Universitat Autònoma de Barcelona

Self-supervised learning for image-to-image translation in the small data regime

Aitor Álvarez Gila

Universitat Autònoma
de Barcelona

Self-supervised learning for image-to-image translation in the small data regime

A dissertation submitted by **Aitor Álvarez Gila** to Universitat Autònoma de Barcelona in fulfilment of the degree of **Doctor of Philosophy** in the Dept. Ciències de la Computació.

Bellaterra, June 2, 2022

Directors	**Dr. Joost van de Weijer** Centre de Visió per Computador Universitat Autònoma de Barcelona **Dr. Estíbaliz Garrote** Computer Vision & Visual Interaction Tecnalia
Thesis committee	**Dr. Rafael García** Institut de Recerca en Visió per Computador i Robòtica Universitat de Girona **Dr. María Vanrell** Centre de Visió per Computador Universitat Autònoma de Barcelona **Dr. Marc Masana** Institute of Computer Graphics and Vision Technische Universität Graz

This document was typeset by the author using LaTeX 2_ε.

ISBN: 978-84-124793-3-1

A Nerea (y los bichitos)

Abstract

The mass irruption of Deep Convolutional Neural Networks (CNNs) in computer vision since 2012 led to a dominance of the image understanding paradigm consisting in an end-to-end fully supervised learning workflow over large-scale annotated datasets. This approach proved to be extremely useful at solving a myriad of classic and new computer vision tasks with unprecedented performance —often, surpassing that of humans—, at the expense of vast amounts of human-labeled data, extensive computational resources and the disposal of all of our prior knowledge on the task at hand. Even though simple transfer learning methods, such as fine-tuning, have achieved remarkable impact, their success when the amount of labeled data in the target domain is small is limited. Furthermore, the non-static nature of data generation sources will often derive in data distribution shifts that degrade the performance of deployed models. As a consequence, there is a growing demand for methods that can exploit elements of prior knowledge and sources of information other than the manually generated ground truth annotations of the images during the network training process, so that they can adapt to new domains that constitute, if not a small data regime, at least a small *labeled* data regime.

This thesis targets such few or no labeled data scenario in three distinct image-to-image mapping learning problems. It contributes with various approaches that leverage our previous knowledge of different elements of the image formation process: We first present a data-efficient framework for both defocus and motion blur detection, based on a model able to produce realistic synthetic local degradations. The framework comprises a self-supervised, a weakly-supervised and a semi-supervised instantiation, depending on the absence or availability and the nature of human annotations, and outperforms fully-supervised counterparts in a variety of settings.

Our knowledge on color image formation is then used to gather input and target ground truth image pairs for the RGB to hyperspectral image reconstruction task. We make use of a CNN to tackle this problem, which, for the first time, allows us to exploit spatial context and achieve state-of-the-art results given a limited hyperspectral image set.

In our last contribution to the subfield of data-efficient image-to-image transformation problems, we present the novel semi-supervised task of *zero-pair cross-view semantic segmentation*: we consider the case of relocation of the camera in an end-to-end trained and deployed monocular, fixed-view semantic segmentation system often found in industry. Under the assumption that we are allowed to obtain an additional set of synchronized but unlabeled image pairs of new scenes from both original and new camera poses, we present ZPCVNet, a model and training procedure that enables the production of dense semantic predictions in either source or target views at inference time. The lack of existing suitable public datasets to develop this approach led us to the creation of MVMO, a large-scale Multi-View, Multi-Object path-traced dataset with per-view semantic segmentation annotations. We expect MVMO to propel future research in the exciting under-developed fields of cross-view and multi-view semantic segmentation.

Last, in a piece of applied research of direct application in the context of process monitoring of an Electric Arc Furnace (EAF) in a steelmaking plant, we also consider the problem of simultaneously estimating the

temperature and spectral emissivity of distant hot emissive samples. To that end, we design our own capturing device, which integrates three point spectrometers covering a wide range of the Ultra-Violet, visible, and Infra-Red spectra and is capable of registering the radiance signal incoming from an 8cm diameter spot located up to 20m away. We then define a physically accurate radiative transfer model that comprises the effects of atmospheric absorbance, of the optical system transfer function, and of the sample temperature and spectral emissivity themselves. We solve this inverse problem without the need for annotated data using a probabilistic programming-based Bayesian approach, which yields full posterior distribution estimates of the involved variables that are consistent with laboratory-grade measurements.

Keywords: *computer vision, neural networks, self-supervised learning, image-to-image mapping, probabilistic programming*

Resum

La irrupció a gran escala de Xarxes Neuronals Convolucionals Profundes (*Convolutional Neural Networks*, CNNs) a la visió per computador des de 2012 ha duït a un paradigma predominant d'interpretació de la imatge consistent en un procés d'aprenentage completament supervisat amb conjunts massius de dades etiquetades. Aquesta aproximació ha resultat ser extremadament útil per a solucionar una miríada de tasques de visió per computador, tant noves com clàssiques, amb resultats sense precedents (sovint superiors als obtinguts per humans), a costa d'emprar grans quantitats de dades anotades manualment, recursos computacionals de gran magnitud, i tot el coneixement previ possible sobre la tasca a resoldre. Tot i que tècniques senzilles de transferència de l'aprenentatge, com el reajustament acurat (*fine tuning*), han obtingut un gran impacte, el seu èxit quan la quantitat de dades etiquetades al domini objectiu és petita es manté limitat. A més a més, el caràcter no estàtic de les fonts de generació de dades habitualment resulta en canvis inesperats en la distribució d'aquestes dades, degradant el rendiment dels models ja desplegats. Com a conseqüència, hi ha una demanda creixent de mètodes que puguin explotar elements de coneixement previ i fonts d'informació més enllà del conjunt d'etiquetes generades per un humà expert en el transcurs del procés d'entrenament de la xarxa, per a que puguin adaptar-se a nous dominis que constitueixen, si no un règim d'escasses dades, sí d'escasses dades *etiquetades*.

Aquesta tesi s'adreça a aquesta classe d'escenaris amb manca de dades etiquetades en tres problemes de transformació imatge a imatge. Contribueix amb una sèrie de metodologies basades en coneixement a priori dels diferents elements del procés de formació de la imatge. Primer introduïm un marc conceptual, eficient en l'ús de dades, per al tractament del desenfocament tant estàtic com degut a moviment, basat en un model capaç de produir degradacions locals sintètiques però realistes. Aquest marc es pot instanciar de tres maneres diferents: en tècnica auto-supervisada, feblement supervisada, o semi-supervisada, en funció de la absència o disponibilitat d'anotacions humanes, així com la seva naturalesa, i resulta ser superior a les corresponent versions completament supervisades en una varietat de situacions.

El coneixement del procés de formació del color en la imatge és aprofitat després per a recopilar parelles entrada/objectiu d'imatges en el context de reconstrucció hiperespectral de la imatge. Emprem una CNN per a resoldre aquest problema, la qual cosa ens permet per primera vegada explotar context espacial i aconseguir resultats que estableixen un nou estat de l'art a partir de un conjunt reduït d'imatges hiperespectrals.

En la nostra darrera contribució a l'àmbit de la transformació d'imatge a imatge en problemes amb poques dades anotades, presentem la nova tasca semi-supervisada de *segmentació semàntica amb zero parells i amb vistes creuades*: considerem el cas de recol·locació de camera en un sistema (comú en escenaris industrials) de segmentació semàntica de vista fixada entrenat de principi a fi i desplegat. Assumint que podem obtenir un conjunt adicional de pars d'imatges, no etiquetades però sí sincronitzades, de noves escenes emprant la posició de càmera original i nova, presentem ZPCVNet, un model i procediment d'entrenament que possibilita la generació de prediccions semàntiques denses tant en vistes de inici com vistes objectiu, en temps d'inferència. La carència de bases de dades públiques per a poder desenvolupar la metodolgia proposta ens condueix a la creació de MVMO, una base de dades pública de gran escala, multi-vista, multi-objecte, renderitzada mitjançant *path tracing*, amb anotacions per-vista de segmentació semàntica. Pensem que MVMO promourà

futura recerca en la molt interessant però poc explorada àrea de la segmentació semàntica multi-vista i amb vistes creuades.

Finalment, en una peça de recerca aplicada amb aplicació directa en un context de monitorització de processos de Forn d'Arc Elèctric (*Electric Arc Furnace*, EAF) en una acereria, considerem el problema d'estimació simultània de la temperatura i la emissivitat espectral de mostres emissives calentes. Dissenyem el nostre propi sistema de captura, integrant tres espectòmetres puntuals cobrint un ampli rang de l'espectre Ultra-Violeta, visible i Infraroig, capaç de registrar senyal radiant entrant per un forat de 8cm de diàmetre localitzat fins a 20m de distància. Llavors definim un model físicament precís de transferència radiant, incloent efectes d'absorbció atmosfèrica, de la funció de transferència òptica del sistema, així com de la temperatura i emissivitat espectral de la mateixa mostra. Resolem aquest problema invers sense la necessitat de dades anotades, mitjançant una aproximació Bayesiana basada en programació probabilística, que proporciona estimacions de la distribució posterior sencera sobre les variables involucrades, consistents amb mesures de nivell de laboratori.

Paraules clau: *visió per computador, xarxes neuronals, aprenentatge auto-supervisat, transformació d'imatge a imatge, programació probabilística*

Resumen

La irrupción masiva de las Redes Neuronales Convolucionales (*Convolutional Neural Networks*, CNN) en visión artificial a partir de 2012 condujo a un dominio, en el ámbito de la interpretación de imágenes, del paradigma consistente en un esquema de aprendizaje extremo-a-extremo totalmente supervisado sobre bases de datos de imágenes de gran escala. Esta aproximación demostró ser extremadamente útil para la resolución de innumerables tareas de visión artificial tanto clásicas como de nueva creación, con un rendimiento predictivo sin precedentes —a menudo mejor que el de los propios humanos—, a costa de requerir grandes cantidades de datos anotados por humanos y de recursos de computación, y de tener que descartar todo nuestro conocimiento previo sobre la tarea en cuestión. Pese a que los métodos sencillos de aprendizaje por transferencia, tales como el ajuste fino (*fine-tuning*), han logrado un impacto notable, su éxito se ve mermado cuando la cantidad de datos anotados en el dominio de destino es reducida. Asimismo, el carácter no estático de las fuentes de generación de datos deriva, con frecuencia, en desplazamientos de la distribución de los datos que dan lugar a una degradación del rendimiento de modelos ya desplegados. En consecuencia, existe una creciente demanda de métodos que puedan explotar tanto aspectos de conocimiento *a priori* como fuentes de información adicionales a las anotaciones manuales de las imágenes durante el proceso de entrenamiento, de manera que sean capaces de adaptarse a nuevos dominios que constituyen, si no un régimen de escasez de datos, sí al menos un régimen de escasez de datos *anotados*.

La presente tesis aborda dicho escenario de ausencia total o parcial de datos anotados en tres problemas de aprendizaje para mapeo imagen a imagen. En ella se hacen contribuciones en forma de varias aproximaciones que se apoyan en nuestro conocimiento previo sobre diferentes elementos del proceso de formación de imágenes: en primer lugar, presentamos un marco de trabajo eficiente (desde el punto de vista del uso de datos) para la detección de borrosidad debida a desenfoque o movimiento, en base a un modelo capaz de producir degradaciones locales sintéticas realistas. La propuesta se compone de tres implementaciones (una auto-supervisada, una de supervisión débil, y una semi-supervisada), en función de la ausencia o disponibilidad y de la naturaleza de las anotaciones humanas disponibles, y supera, en cuanto a resultados, a alternativas totalmente supervisadas en varias configuraciones.

A continuación, empleamos nuestro conocimiento del dominio de la formación de imágenes en color para recopilar así parejas de imagenes de entrada y objetivo para la tarea de reconstrucción de imagen hiperespectral. Acometemos este problema haciendo uso de una CNN que, por primera vez, nos permite explotar el contexto espacial y lograr resultados que suponen una avance en el estado de la técnica, dado un conjunto de imágenes hiperespectrales limitado.

En nuestra última contribución al subcampo de los problemas de transformación de imagen a imagen en condiciones de escasez de datos, presentamos la nueva tarea semi-supervisada de *segmentación semántica de vista cruzada con cero-pares*: consideramos el caso de reubicación de la cámara en un sistema —frecuente en la industria— de segmentación semántica monocular de pose fija, entrenado extremo-a-extremo y ya implantado. Bajo la asunción de que se nos permite obtener un conjunto adicional de pares de imágenes sincronizadas pero no anotadas de nuevas escenas desde ambas ubicaciones de cámara, presentamos ZPCVNet, un modelo y procedimiento de entrenamiento que posibilita, en tiempo de inferencia, la generación de predicciones

semánticas densas, tanto en el marco de referencia de la pose original como en la de destino. La inexistencia de bases de datos públicas adecuadas para poder desarrollar este planteamiento nos condujo a la creación de MVMO, una base de datos de gran escala de imágenes Multi-Vista y Multi-Objeto, renderizadas mediante *path tracing*, y dotada de anotaciones en términos de segmentación semántica para cada vista. Esperamos que MVMO estimule futuras investigaciones en las áreas, ahora subdesarrolladas, de la segmentación semántica multi-vista y de vista cruzada.

Por último, en un ejercicio de investigación aplicada de utilidad directa en el contexto de monitorización del proceso en una planta de acería con horno eléctrico de arco (*Electric Arc Furnace*, EAF), consideramos también el problema de estimación conjunta de la temperatura y la emisividad espectral para muestras emisivas calientes distantes. Para ello, diseñamos e integramos nuestro propio dispositivo, el cual incorpora tres espectrómetros puntuales que cubren, en conjunto, un amplio rango espectral en las zonas ultravioleta, visible e infrarroja. El equipo es capaz de registrar la señal de radiancia procedente de un punto de 8 cm de diámetro ubicado a 20 m de distancia. Asimismo, formulamos un modelo de transporte radiativo riguroso desde el punto de vista de la física, que contempla los efectos de la absorbancia atmosférica, de la función de transferencia óptica del sistema, y de la propia temperatura y emisividad espectral de la muestra. Resolvemos este problema inverso sin requerir para ello dato anotado alguno, haciendo uso de una aproximación bayesiana apoyada en un modelo de programación probabilística que ofrece estimaciones de la distribución posterior de las variables aleatorias definidas que son consistentes con las mediciones realizadas en un entorno controlado y con equipamiento de laboratorio.

Palabras clave: *visión por computador, redes neuronales, aprendizaje auto-supervisado, mapeo imagen a imagen, programación probabílistica*

Laburpena

Sare Neuronal Konboluzionalak (*Convolutional Neural Networks*, CNN) 2012tik aurrera ikusmen artifizialean modu masiboan erabiltzen hasi izanak, eskala handiko irudi datu-baseen bidezko muturretik muturrerako guztiz gainbegiratutako ikasketa paradigma erakarri zuen irudien interpretazioaren esparrura. Hurbilketa hori oso baliagarria izan zen ikusmen artifizialeko arazo ugari —bai klasikoak zein sortu berriak— ebazteko aurrekaririk gabeko errendimendu prediktiboarekin, sarritan gizakiena bera baino hobeagoa. Horren truke, konputazio-baliabide anitzak eta gizakiok etiketatutako irudi kopuru handiak bereganatzea beharrezkoa bihurtu zen, bai eta arazoari buruz aldez aurretik genuen ezagutza guztia baztertzea ere. Transferentzia bidezko ikaskuntza-metodo sinpleek, hala nola doikuntza finak (*fine-tuning*), inpaktu nabarmena lortu duten arren, haien arrakasta murriztu egiten da helmuga-eremuan etiketatutako datuen kopurua txikia denean. Era berean, datu-iturburuen izaera ez-estatikoak maiz datuen banaketaren desplazamendua dakar ondoriotzat, eta honek, era berean, dagoeneko lanean dabiltzan ereduen errendimendua degradatzen du. Ondorioz, entrenamendu-prozesuan nahiz gure *a priori*-zko ezagutza zein irudien eskuzko etiketen informazio iturburu osagarriak ustia ditzaketen metodoen eskaerak gora egin du azken garaiotan; era honetan, datu-eskasiaren erregimena —edota, behintzat, *etiketaturiko* datu-eskasiaren erregimena— kontsidera daitezkeen eremu berrietara egokitzeko gai izango direlakotan.

Tesi honek etiketaturiko datuen eskasiaren edota erabateko faltaren eszenatokia jorratzen du iruditik irudira mapatzeko hiru ikasketa arazotan. Bertan, irudiak eratzeko prozesuaren hainbat elementuri buruz aurretik dugun ezagutzaz baliatzen gara zenbait kontribuzio aurkezteko: Lehenik eta behin, metodo efiziente bat aurkezten dugu (datuen erabileraren ikuspegitik) optika eta mugimenduaren ondoriozko desfokuratze efektuak detektatzeko, degradazio lokal sintetiko errealistak sortzeko gai den eredu batean oinarrituta. Hiru inplementaziok osatzen dute aipatutako metodoa (bat auto-gainbegiratua, bat gainbegiratze ahulekoa, eta beste bat erdi-gainbegiratua), eskuragai dauden etiketen eta haien izaeraren arabera, eta hainbat konfiguraziotan hobetzen ditu guztiz gainbegiratutako anternatibek lorturiko emaitzak.

Jarraian, koloretako irudien eraketaren inguruko gure ezagutza erabiltzen dugu sarrerako eta aurreikusi beharreko irteerako irudi bikoteak gauzatu, eta horri esker irudi hiperespektralak berreraikitzeko problemari heltzeko. Horretarako, CNN bat erabiltzen dugu eredu gisa, zeinek, aurreneko aldiz, testuinguru espaziala ustiatzeko aukera emango digun, eta, honen bitartez, arte-egoeraren aurrerapena dakarten emaitzak lortu, irudi hiperespektral multzo mugatu bat besterik ez izanik eskuragai.

Datuen urritasuneko baldintzetan iruditik irudirako bihurketa arazoen azpieremuari egiten diogun azken ekarpenean, ataza erdi-gainbegiratu berri bat aurkezten dugu, *bikote gabeko ikuspegi gurutzatuko segmentazio semantikoa* izenez: jada ezarrita dagoen muturretik muturrera entrenatutako segmentazio semantiko sistema monokular batean (industrian maiz aurki dezakeguna) kamera birkokatzeko beharra azaltzen zaiguneko kasua hartzen dugu kontutan. Kontestu honetan, demagun bi kamera-kokapenetatik sinkronizatutako baina etiketatu gabeko eszena multzo berri baten irudi pare gehigarri bat lortzea baimentzen zaigula. Hau suposatuta, ZPCVNet aurkezten dugu, eredu eta entrenamendu-prozedura bat, inferentzia-garaian pixel bakoitzeko aurreikuspen semantikoak sortzea ahalbidetzen duena, bai jatorrizko ikuspuntuko erreferentzian, zein helmugakoan. Planteamendu hori garatu ahal izateko datu-base publiko egokirik ez zegoenez, MVMO

sortu genuen, ikuspuntu anitzeko eta objektu anitzeko eskala handiko irudi datu-base bat, alegia, *path tracing* teknikaren bidez errenderizatua, eta ikuspegi bakoitzerako segmentazio semantikoko etiketekin batera argitara ematen dena. Espero dugu MVMO-k etorkizunean ikerketa berriak sustatuko dituela gaur egun azpigaratuta dauden ikuspegi anitzeko eta ikuspegi gurutzatuko segmentazio semantikoaren arloetan.

Azkenik, arku elektrikozko labea (*Electric Arc Furnace*, EAF) duen altzairutegi bateko prozesua monitorizatzeko testuinguruan zuzenean ezarri genezaken ikerketa aplikatu batean, urrutiko lagin igorle beroen tenperatura eta emisibitate espektralaren aldi bereko estimazioaren arazoari heltzen diogu. Horretarako, gure gailu propioa diseinatu eta garatzen dugu, hiru espektrometro puntualez osatua, guztien artean espektrotarte zabala hartzen dutelarik barne, tarte ultramore, ikusgai eta infragorrian. Aipaturiko gailua gai da 8 cm-ko diametroko zirkulu batetik datorren erradiantzia-seinalea erregistratzeko, 20 m-ko distantziara. Era berean, fisika aldetik zehatza den transferentzia erradiatiborako eredu bat formulatzen dugu, absorbantzia atmosferikoaren, sistemaren transferentzia optikoaren funtzioaren eta laginaren tenperatura eta emisibitate espektralaren beraren efektuak jasotzen dituena. Alderantzizko problema hau ebazten dugu, horretarako inolako etiketaturiko daturik behar gabe. Horretarako, programazio probabilistikoko eredu batean oinarritutako hurbilketa bayestarra erabiltzen dugu, definitutako ausazko aldagaien ondorengo banaketaren estimazioak eskaintzen dituena, ingurune kontrolatu batean laborategiko ekipamenduaren bidez jasotako emaitzekin bat datozenak.

Gako-hitzak: *ikusmen artifiziala, neurona-sare artifizialak, ikasketa auto-gainbegiratua, iruditik irudirako transformazioa, programazio probabilistikoa*

Acknowledgements

Working in the pursuit of a PhD is often depicted as a long, individual journey. Though as winding and solitary as it can get, the road through here was paved with the aid and support from many distinct individuals that helped enrich the experience and make it a rewarding one. The following lines are my modest attempt to transmit my most sincere gratitude to those who, one way or another, assisted me in fulfilling this personal project.

Let me first acknowledge the many ways in which my supervisors, Joost and Estibaliz, supported me to achieve this goal. Joost, I will always be grateful for your confidence and for having given me the opportunity to join the team. The remote, part-time nature of my position barely allowed me to scratch the surface of what it means to work with you, and yet, I have enjoyed every single bit of work done together. I genuinely admire what you do and how you do it, and, technical insights apart, I hope I have picked up at least a small part of your sense of purpose and that amazing ability to distill the substance from the noise. Esti, you have been a fundamental piece of this journey as well. I am deeply thankful for believing in me since the very beginning, for backing my decisions and for protecting me through this adventure. You fought for this to happen while getting all the bullets, and I would have never reached this point without your support and guidance.

I must of course extend my gratitude to my coauthors and those who contributed scientifically to the presented works: Yaxing, with his hints and immense knowledge of the state of the art. A very especial shoutout to Adrián, with whom it all started (watching Karpathy's CS231n videos in semi-clandestinity) and whose Itzi-supervised text-to-text translation helped get the last bits of this dissertation done. The long research that led to Chapter 4 was only possible with the joint effort of many colleagues from different collaborating institutions: thanks to Asier for providing the necessary media and domain knowledge, to Jan for his confidence in us and the freedom I always felt working with him, which ultimately allowed us to get these results. Thanks should also go to Gabriel, for putting some solid scientific ground under our shaky feet, and of course, to Artzai, whose legendary technical intuition adds to an equal dose of energy, chaos, excellence and questionable sense of humor to yield the perfect mix that constitutes the very heart of Tecnalia's computer vision team.

I would also like to acknowledge all the managers at Tecnalia (Patxi, Mikel, Ana, Jone) that, over the years, understood the long-term value of this endeavour and allowed me to pursue this PhD. Same goes for those other current and past team members and colleagues who contributed to make my time at the office more enjoyable: Jone, for our heated discussions on the most improbable topics, Arantza, for your unswerving optimism, Miguel, Cristina, Usue, Belén, Laura, Alberto, Gorka and all the others.

Thanks should also go for all the other LAMP members that made me feel welcome and part of the team even in a *scarce visit regime*: Xialei, Lu, Bogdan, Bartek, Mikel, Laura, Luis and the rest of the team. Special thanks to Marc, for being the best host one could expect to find, and a real glue for the group. I should also mention Javi for his advice in the pre-PhD era in my search for an advisor, and all the CVC staff members that silently make things just work, particularly Montse and Claire for their enthusiastic help when needed.

Not all the support came from the academic and work environments, of course. A huge thank you to my friends, for all the laughs along these years. Your humor has been an invaluable source of energy to deal with the tough moments.

I could definitely not have undertaken this journey without the aid of my family: to Borja, Ekaitz and Leire, their company cheered me up uncountable times. A very affectionate hug goes to Begoña, in memory of Santos, whose conversations and laughs I once enjoyed and now miss. And yet another one to my sister, Saioa, who has been consistently sweetening our way here, first through her sweets, and with June from now on. I feel immensely grateful to my parents, Jose and Rosa, for their care, unconditional love, and for a lifetime of sacrifices to make sure we had the education that they could never get access to for themselves. I feel very proud of you.

Finally, this thesis is dedicated to my closest ones: to our kids, Aritz and Malen, who have only known me as a PhD student, and will soon gain a full-time, dedicated father. My most loving gratitude for helping me find joy and passion in the most unanticipated places. And last, my deepest thanks-love to Nerea, who stood by me at every single step and got my back countless times without ever expecting anything in exchange. She has been my single most reliable piece of continuous support along this journey, sharing the lows and celebrating the highs, and helping me become a better version of myself when she is around. None of this would have been possible without you.

Contents

List of Figures

List of Tables

1 Introduction

As you are reading this sentence, each human on earth is generating 8 MB of data [51]. As a result of the social media, the ongoing digital transformation of the industrial processes and the increasingly digital nature of consumer products, it is expected that, by 2025, more than 200 zettabytes of data will be in cloud storage around the globe [173] and that 463 exabytes of them will be generated on a daily basis [230]. We are therefore witnessing a constant redefinition of what we consider to be *Big Data*, amid claims of data being the new oil [103].

In this context, only a small fraction of this generated data is currently being processed, analyzed or even stored [214]. Artificial Intelligence (AI) is called to be the missing piece that will allow us to actually exploit such vast amount of information, leveraging the use of Deep Neural Networks (*i.e*. Deep Learning [18, 86]), and transform the raw data —petroleum— into actual oil. In the field of computer vision, it was the aforementioned ubiquitous availability of digitized *and labeled* data, together with an unprecedented availability of parallel computing resources (materialized as mass market oriented, relatively cheap GPUs-Graphics Processing Units) and the slow but incremental progress in the development of robust training algorithms that enabled the ongoing deep learning revolution.

The mass irruption of Deep Convolutional Neural Networks (CNNs) in computer vision can be exemplified by the ILSVRC2012 breakthrough [130], in which, for the first time, a CNN-based model beat all other contenders in the *Imagenet Large Scale Visual Recognition Challenge* [229] by a considerable margin, although its theoretical roots can be traced back at the very least to the late 90s. The use of CNNs to solve image understanding tasks (such as *e.g*. image classification), represented a paradigm shift with respect to the era of *classic* computer vision, which was characterized by pipelines comprising hand-crafted feature extractors (*i.e*. descriptors) transforming the input image in a compact representation of it, and a subsequent machine learning classifier. Under this schema, such descriptors were designed so that they yielded discriminative representations, distinct for input samples belonging to different classes and similar when describing samples within the same category, while remaining invariant to nuance factors of variation within any single class (e.g. illumination, camera pose, etc.), thus posing a trade-off between both desirable aspects. Remarkably, these handcrafted descriptors were designed so that, when addressing such trade-off, they would leverage what we knew about the domain of application, about the physical processes of the world being represented or the image capturing process. In other words, they would encode our *prior knowledge*. For instance, when designing descriptors required to obtain good representations of the colors in an image, these were required to be invariant to the chromaticity of the main illuminant in the scene —so that the overall algorithm would be robust to illuminant changes [77]. In order to do so, we had to leverage our best knowledge about the physics of the interaction of light with object surfaces, formalized, e.g. as a reflection model [235].

Contrary to this approach, deep neural nets [139] operate by replacing both elements —feature extractor and machine learning classifier— by one single model, which is trained directly on pairs of images (raw pixels) and ground truth labels. During the training process, network coefficients are automatically adjusted —through a process known as *stochastic gradient descent* [120, 219], and based on the *backpropagation* algorithm [140]— so that the errors of the predictions made by the model are minimized. While doing so, the neural net learns to extract hierarchical representations (i.e. *features*) of every input image, so that higher

level features (which are semantically more meaningful and compact) are composed of lower level ones.

This end-to-end fully supervised learning approach over large-scale annotated datasets yielded excellent results and proved to be extremely useful at solving a myriad of classic and new computer vision tasks with unprecedented performance, at the expense of (i) high computational requirements (ii) the need for vast amounts of human-labeled data and (iii) the disposal of all of our previously acquired prior knowledge on the task being addressed. In fact, such solutions have consistently been beating traditional workflow-based shallow approaches in almost every computer vision task. Likewise, they allowed addressing new tasks that were not even viable up until the appearance of these techniques, even surpassing human performance at some of them.

However, in the last few years, an emerging notion has arisen, that the manual annotation of such large amounts of data might not be the best or even a desirable path [141]. Even though simple transfer learning methods (such as fine-tuning a large network pretrained on a large, labeled dataset onto a smaller dataset from a different domain [285]) have achieved remarkable impact, their success when the amount of labeled data in the target domain is small is limited. The data generation sources are not static: cameras evolve, industrial environments in which images are captured are subject to quick modifications, models trained from footage from a certain city are deployed in another country, and the resulting shifts in the underlying pixel distribution statistics render the original annotations —which may have taken a big effort to be created— obsolete quickly. In this context, it is becoming increasingly clear that there is a demand for methods that can exploit elements of prior knowledge and sources of information other than the manually generated ground truth annotations of the images during the network training process, so that they can adapt to new domains that constitute, if not a small data regime, at least a small *labeled* data regime.

1.1 Background

In this next section, we elaborate on some of the areas of research that have recently emerged around the need to train neural networks for domains with only little labeled data.

Exploiting synthetic data. Right at the intersection between the fields of computer vision and computer graphics, the use of synthetically generated data to train deep neural networks has shown to be one very promising approach to circumvent the lack of extensive labeled ground truth. The ongoing rise of the *data-centric AI* movement [175], which advocates for a dramatic increase in the efforts devoted to obtaining high-quality datasets, as opposed to the current algorithm-centric approach, is also aligned with this view. In fact, the use of 3D modeling software such as Blender [44], or 3D game engines (*e.g.* Unity [115], Unreal [204], or GTA [217]) allow for generating a virtually infinite set of training image and corresponding pixel-perfect, error-free ground truth pairs while reducing the risk of dataset bias issues [178] present in real-world datasets. Furthermore, synthetic data generation enables an extensive coverage of long tail events or corner cases that would be hard to capture through conventional real-world image gathering. Given an important upfront cost of 3D-modeling a scene —or that of automating its construction itself [88]—, the synthetic approach can produce multi-modal ground truth labels at minimal cost.

The main drawback, on the other hand, of this synthetic pathway (even in the case of photo-realistic, path-tracing based workflows) lays in the distribution shift between the synthetic and the real world. As a result, this transition of the synthetically-trained model to the real world production environment, typically leads to a significant performance drop. Several approaches have been proposed to try to narrow this gap: different variants of domain adaptation techniques in both unsupervised and semi-supervised flavors [268] have shown to help, while recent research hints at the potential prominent role of diversity of scene appearances (e.g. illumination, camera pose, surfaces, etc.) at bridging the gap [255], while downplaying the effect of

photo-realism for some image understanding tasks [257].

On the other hand, the recent rise of high resolution deep generative models [85, 206] paved the way towards their potential use for synthetic dataset construction [110] as an alternative to the aforementioned 3D rendering workflows. The inherent difficulties of adversarial training and issues related to diversity, such as mode collapse [9, 165], are still obstacles to the full development of this approach.

All in all, the availability of synthetic data and labels for training or pretraining becomes of paramount importance when dealing with dense prediction tasks, such as semantic or instance segmentation, or depth estimation. In this context, the time required to manually annotate each image renders the process hardly scalable. Furthermore, 3D annotations (such as 3D bounding boxes or 6D poses [231] are even harder to obtain over real world input.

In this thesis, we make use of synthetically generated data of diverse forms in Chapters 2, 3, 5 and 6.

Deep image-to-image translation. The task of image-to-image translation can be better thought of as an abstract super-task encompassing a number of different computer vision problems that can be modeled under this scheme. In essence, we want, given an image-shaped input, to obtain an output with the same spatial shape as the input and one or more channels, representing some dense prediction defined with pixel-wise resolution, so that we can observe its spatial variations. The meaning and formatting of such outcome is dependent on the specific task being tackled. In this thesis we will consider several image-to-image settings: (i) blur detection (Chapter 2) [238] aims at spatially localizing the blurred parts of an input image and producing real-valued output blur maps, which can be then exploited by downstream algorithms for further processing (*e.g*. deblurring); (ii) in RGB to HSI (HyperSpectral Image) reconstruction (Chapter 3) [6], the output comprises as many real-valued dense prediction channels as those given by the spectral sampling rate inherent to the available capture devices (typically a few tenths), *i.e*. each pixel represents a spatially localized spectrum (normally representing the spectral reflectance or radiance of the depicted object at those coordinates); (iii) semantic segmentation (Chapters 5/6) [155] predictions can be represented by a color-indexed image with a single channel of integers (each representing a class).

Many of these tasks are often approached through a common base of encoder-decoder based CNN architectures [12]. As a consequence, improvements over the state of the art in one of the fields can be, to some extent, easily transferred to other image-to-image translation problems. The massive adoption of the U-Net architecture [223] across domains and tasks as baseline architecture illustrates this. In the mentioned encoder-decoder scheme, the encoder is typically conceived similarly to an image classification network, except for its last few layers in charge of converting the predicted scores into a hard class decision. This module takes an image-like input (*i.e*. high spatial resolution with just three channels in the RGB case) and produces as output a very compact, semantically meaningful representation of it (*i.e*. a code) which we refer to as *bottleneck features*, *latent space* or *latent representation*. Meanwhile, the decoder takes such code and expands it back to the original resolution while containing the desired prediction. Although having access to the semantically rich latent features enables multiple forms of operation with them, thus enabling *e.g*. plausible semantic image editing operations [237], open challenges remain to this respect, such as the generation of a features space with metric properties [174] or one in which different meaningful factors of variations are disentangled and identified [1], facilitating precise modifications on the produced output.

Incorporating prior knowledge. The last few years have witnessed an increasing effort from the computer vision —and, in broader terms, from general machine learning— communities to overcome the large scale, fully supervised end-to-end paradigm in deep learning by incorporating back into the models part of the prior knowledge we have about the specific domains under study [53]. The rationale behind this so-called *theory-guided data science* [116] is two-fold:

First and foremost, including constraints that reflect on our understanding of natural physical processes —normally formalized as mathematical models that have been extensively proven— can help our black box

deep learning models trained through gradient descent reduce the search space of solutions [227]. This is of particular interest whenever they are forced to deal with small (annotated) data regimes.

Moreover, there is no universal approach for the inclusion of such inductive biases into the training process, but in certain cases this can be materialized in the form of specific backpropagation-ready modules of the deep architectures reproducing the behavior of such physical processes [177]. While doing so, these modules shed light into the so-called black box and act as interpretable elements that can even serve as means to gain new insights of the domain [34].

In all fairness, CNNs themselves are all but free of inductive biases: convolutional layers encode our strong beliefs on the stationary nature of natural images across the spatial dimensions, and our shared wisdom that the coefficients learned in the form of compact filters are valid as we apply them throughout the different locations within the image (*i.e*. translation invariance of the learned features). This is, indeed, the single major assumption and design constraint that has enabled the revolution-scale progress of the field of image understanding and computer vision during the last decade. Deep Convolutional Neural Networks represent a sweet spot in the trade-off between the size of the dataset available for training (in the order of a few million image-label pairs [229] for image classification) and the generality of the employed model (or, conversely, the magnitude of the restrictiveness derived from the inductive biases hard-coded in it). And yet, the most recent advances in the field, namely *Transformers* [59], directly imported from the NLP (Natural Language Processing) community [262], hint at the disposal of the convolutional operations, as the availability of even larger-scale datasets (*e.g*. JFT-300M [247]) allows for more general models, solely constructed on top of self-attention mechanisms.

All in all (and whatever the underlying deep architecture), when operating in the small data regime, solutions that, by somehow encoding inductive biases, are able to reduce the solution search space can greatly benefit the vast majority of computer vision tasks that cannot leverage such extra-large datasets. A number of distinct approaches have been attempted so far. Some of them focus on modifying the formulation of the loss function with terms that favor solutions that are plausible from the perspective of the domain being modeled (e.g. fluid dynamics [122]). Knowledge distillation through the use of teacher-student models is another popular mechanism to transfer external knowledge to a deep student network [98]. Recent works propose a hybrid approximation to the problem, with the support of Graph Neural Nets (GNN) complementing the shortfalls of an incomplete and biased probabilistic graphical model acting as generative model of an underlying physical process [73]. As mentioned before, we can find a wide range of methods encoding our previous understanding of the world as specific modules within the architecture itself from a number of subfields, *e.g*. multiple view geometry for the construction of multi-view consistent features [95], or the image capture process in a device from the spectral point of view [177].

While the previous methods focus their efforts on the model itself, another family of techniques put their emphasis on the training data, so that it is the weights in the model that encode this knowledge after training. Many of them are based on the production of simulations of the physical processes under study that can serve to pre-train our model, then to be fine-tuned over real observations [168] or subject to a domain adaptation process [194]. The whole field of synthetic image generation through 3D rendering for dataset construction may well fall within this category: our prior knowledge on plausible scene layouts will be thus contained in the logic for the construction of the 3D scenes, while photometrically consistent physics-based ray-tracing renderers take care of the interaction of light and matter. The 3D rendering workflow provides very complete image formation models that can be exploited for procedural dataset construction. However, often it is not necessary to go that far, and having physically consistent models for specific parts of the image formation process —particularly so for the steps that can introduce degradations in the image *e.g*. noise, blur, etc.— allows for the construction of datasets from existing images, that can be used in the context of *inverse problems* modeling, *i.e*. those where, given an observation (*e.g*. captured image, or magnitude

measured by a sensor), we would like to obtain the real value of the magnitude being captured, the original form of a captured scene (*i.e*. without the degradations inherent to the capture process (*e.g*. deblurring [100], denoising [252], superresolution [24], or hyperspectral reconstruction [8]), or an estimate of one or more of the parameters of the capture process or of the observed event [198]. Closely related to this, but aside the sphere of neural networks, the emerging field of *probabilistic programming* is a powerful tool to address inverse problems provided that we can accurately model the data capture process as mathematical equations. Notably, probabilistic programming provides an unsupervised framework that leverages Bayesian theory and Markov Chain Monte Carlo (MCMC) sampling to tackle these tasks without the need to learn a model, and has been successfully employed in a variety of topics, *e.g*. astronomical parameter estimation [266], stock market modeling [273] or scene understanding [132].

In this thesis, we leverage different ways of incorporating prior knowledge to the model training workflow, either for synthetic data generation (Chapters 2, 3, 5) or to construct a physically consistent model of the signal capturing process (Chapter 4).

1.2 Challenges

Based on our analysis of the state-of-the-art, we identified several challenges related to the topics highlighted in section 1.1 which we will outline in this section.

1.2.1 Blur detection from few labeled images

The first domain in which we investigate data-efficient approaches to the image-to-image mapping problem is that of blur detection under few labeled data. Image blur is an effect that alters the definition of a picture, as a result of which there is a loss of detail in the affected regions. Unless when being used as conscious component of a creative photographic process, it is regarded as an undesired outcome, as it negatively impacts the perceptual experience. There are two main causes that lead to partially or fully blurred images: defocus blur is caused by a non-punctual aperture of the camera-lens system, which projects real world points onto non-punctual circles of confusion in the sensor. Motion blur, on the other hand, is the result of either the camera or the subjects of the scene moving during the duration of the exposure. The task of blur detection aims at accurately segmenting the blurred areas of a given image, independently of their nature. The attained outcome may then be leveraged to tackle additional downstream tasks in diverse settings, such as computational photography (*e.g*. for defocus blur magnification), as a proxy task for other dense predictions that correlate well with it, such as semantic object segmentation, depth estimation or saliency prediction.

As is the case in many other subfields of computer vision, recent deep learning-based methods have been approaching the blur detection problem by learning an end-to-end mapping between the blurred input and a binary or grayscale mask representing the localization of its blurred areas. Nevertheless, as opposed to many of those other tasks, in which the irruption of fully supervised CNN-based methods trained on large scale data elicited an enormous disruption in the state of the art in terms of predictive performance, blur detection experienced a relatively modest performance gain from this approach [121, 159, 187, 287, 291, 294, 295]. A number of factors exist that directly impact the effectiveness of this approach. Most notably, the main bottleneck hindering the exploitation of its full power of CNNs is the absence of large enough and varied datasets with pixel-wise annotations of presence of blur degradation. Several recent efforts have tried to partially address this gap [238,291,294,295] by contributing with novel datasets; however, their scale is orders of magnitude below that of other areas in computer vision [46, 298]. Some of them are also limited in scope and gather only samples from one of the blur variants. At the root of this situation lies the inherent difficulty of producing accurate manual segmentations of blurred regions, which often results in labor-intensive and yet

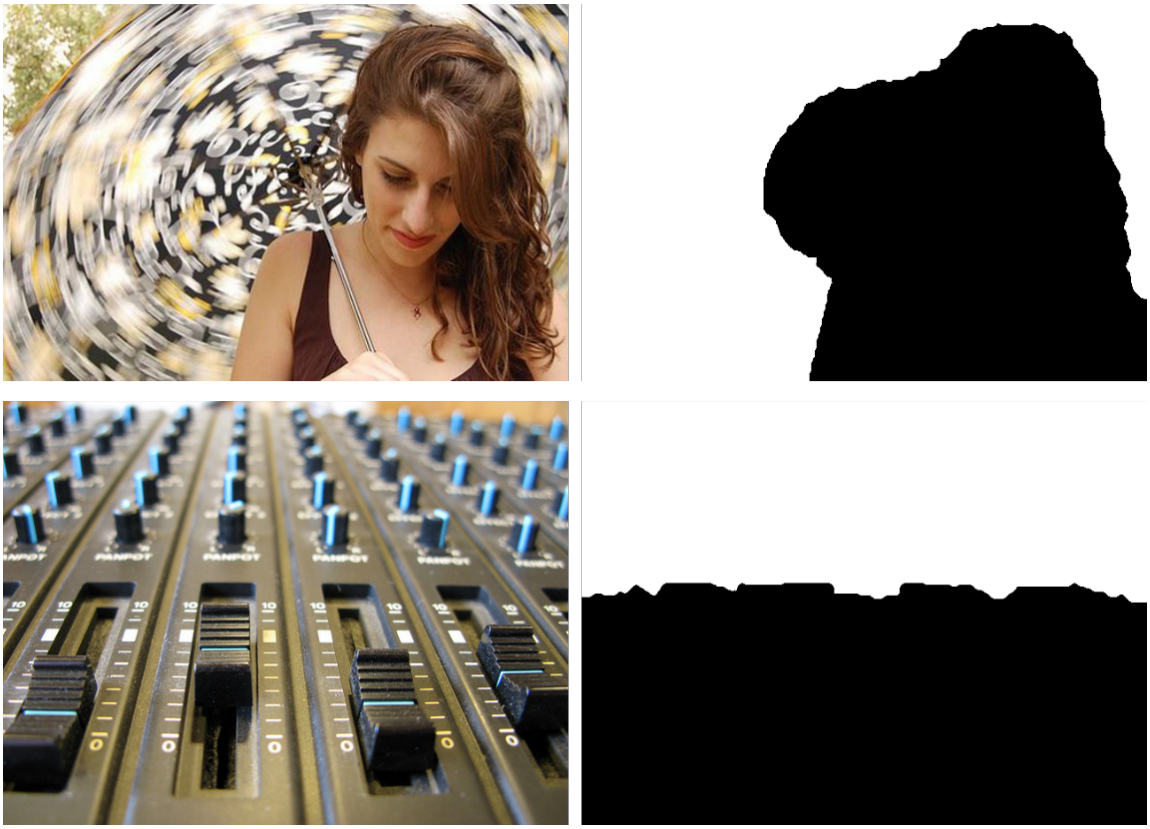


Figure 1.1: Two samples of images partially affected by motion blur (top) and defocus blur (bottom), and their corresponding human-annotated ground truth blur localization masks from Shi's dataset [238]. As can be seen, mask boundaries are subject to the arbitrary criterion of annotators.

poor quality annotations. Except for images exhibiting a clear main subject vs. blurred background separation (which represents one of the abovementioned creative tools for photographic composition), blur intensity is a continuous magnitude in nature. Therefore, producing ground truth annotations through binary masks is an oversimplified model for blur presence most of the time, with hand-drawn boundaries placed at arbitrary locations and subject to the annotator's criterion (see Fig. 1.1).

From the purely algorithmic point of view, additional factors can also explain the moderate advances in the field. At its lowest level, fundamental ambiguities exist between out-of-focus pixels in an image and regions of the image that depict flat surfaces or smooth edges. Scale-ambiguity refers to the inherent difficulty of inferring the level of blur at one single scale [238]. One final challenge relates to the dependence of our perception of sharpness of an image on the image size and resolution [258, 263]. These conditioning factors are often ignored in the design of end-to-end CNN-based solutions in the field, which, to some extent, explain their limited success.

In this thesis, we investigate several alternatives to the end-to-end fully supervised deep black box approach to the problem of blur localization in natural images, in order to circumvent the scarcity of reliably

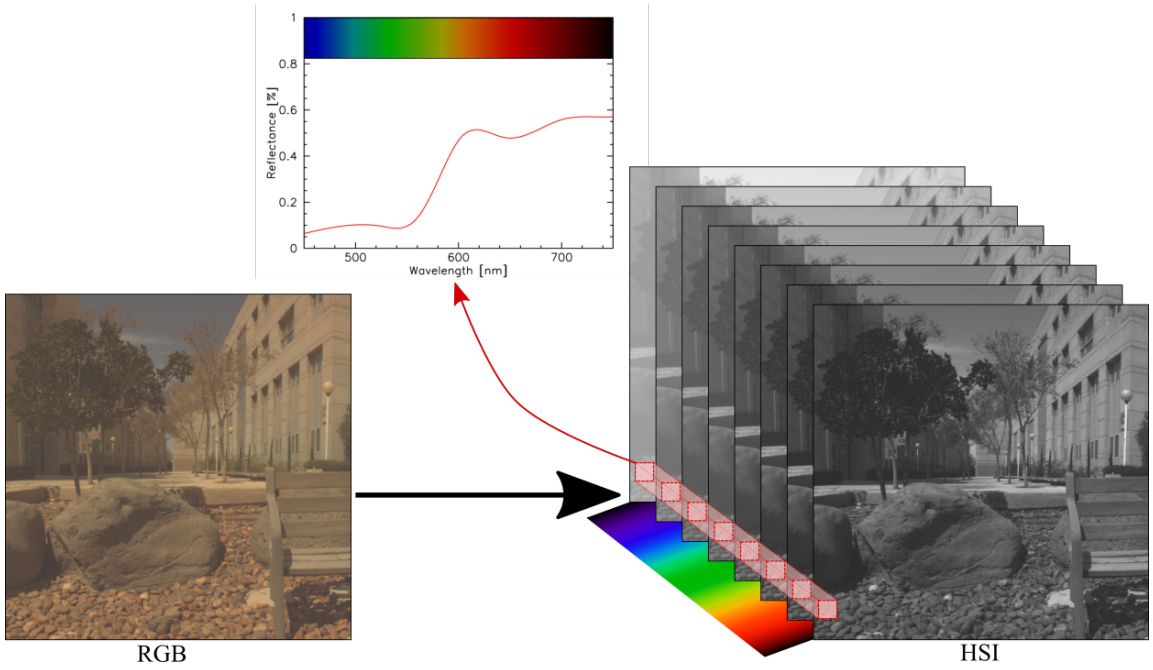


Figure 1.2: Illustration of the hyperspectral image reconstruction task: given an input RGB image, we must estimate a full spectral radiance for each pixel.

annotated large datasets in the field.

1.2.2 Exploiting context for hyperspectral image reconstruction from RGB

The second field in which we address training under few or absent ground truth data for image-to-image mapping is hypespectral image reconstruction. Hyperspectral signal reconstruction (also referred to as spectral super-resolution) aims at recovering the original spectral input that produced a certain trichromatic (RGB) response from a capturing device or observer (Fig. 1.2). Without any deeper spectro-photometric modeling of the problem, one can think of it as the task of recovering a correlate for the spectral radiance associated to each input sample. Given the heavily underconstrained, non-linear nature of the problem, traditional techniques leverage different statistical properties of the spectral signal in order to build informative priors [6] from real world object reflectances for constructing such RGB to spectral signal mapping.

However, a number of challenges still remain open when dealing with extending the task from traditional punctual sample reconstruction to full scenes, even if we limit our scope to the visible range of the spectrum (*i.e.* around $380-780nm$): First, the acquisition of hyperspectral images represents a highly complex procedure on its own. Most HyperSpectral Imaging systems are based on either spectral or spatial scanning of the scene [69], through narrowband variable filters or push-broom methods that capture one hyperspectral line of pixels at a time. This limits their usefulness for capturing natural images of non perfectly static subjects. Even though a few versions of snapshot (*i.e.* matricial) multi-spectral capture devices —relying on repeated mosaic-like local filter-array patterns— are readily available as commercial systems, these are

Figure 1.3: Left: Illustration of an Electric Arc Furnace (EAF) during tapping of molten steel into a vessel, and cross-section of a furnace showing the electrodes. Right: a worker manually checks the temperature using a thercocouple attached to a long probe while slag is poured. Image: Deutsche Fotothek (CC BY-SA 3.0 DE).

still restricted to a very reduced number of bands (up to 25 [27, 75]). This has traditionally dragged the creation of large enough and diverse hyperspectral image datasets [6, 8, 70, 176], that could be employed for HSI-based tasks through data-driven approaches, such as end-to-end deep learning. In particular, HSI recovery from RGB data has been further limited by the additional difficulty of capturing simultaneous co-registered RGB and HSI pairs. This explains the absence of methods in the literature combining spatial and spectral information while tackling RGB to HSI reconstruction in a supervised framework. Ideally, given a good enough reconstruction model comprising a wide set of priors about the spectral features of natural objects we could use standard RGB cameras to produce high spectral and spatial resolution hyperspectral cubes as a cheap alternative to current expensive multi/hyperspectral capturing devices and feed downstream applications with the super-resolved spectra.

In this thesis, we aim to leverage synthetic, physics-based generation of (RGB, HSI) *pairs to perform a CNN-based end-to-end learning of a hyperspectral image reconstruction mapping over a medium-sized dataset that, for the first time, leverages spatial context as well as spectral information.*

1.2.3 Temperature-spectral emissivity separation of hot samples

A third area in which we make contributions under the small labeled data regime assumption is that of remotely estimating temperature and spectral emissivity properties of hot targets in severe industrial settings. Steelmaking plants relying on Electrical Arc Furnaces (EAF) are some of the harshest industrial environments out there. The first and most important step of the steel manufacturing procedure consists, in essence, in melting large volumes of ferrous scrap by forming an electric arc on it (Fig. 1.3, left). Each plant can produce a range of different steel types, and the outcome of the process is also dependent upon the composition of the raw scrap. Therefore, a continuous adjustment must be performed in the furnace through the addition of different chemical compounds. Consequently, there is a need to monitor a number of process variables (*e.g.* temperature) and the composition of both the steel and, especially, the slag (a by-product of the process that

comprises the undesired elements from the scrap and, when in molten state, serves as a mean for stabilizing the electric arc and improving its efficiency).

To this respect, the Holy Grail of EAF-based steelmaking would be achieving a continuous, online, real-time monitoring of both slag temperature and its chemical composition [192]. However, this is far from today's state: in current practice, such chemical composition analysis is run offline through the manual acquisition of a molten sample of slag, its preparation and subsequent analysis in a laboratory X-ray fluorescence spectrometer [197]. Likewise, temperature is manually sampled through the introduction of a thermocouple once or twice during each casting [20] (Fig. 1.3, right).

One giant leap towards that final goal would be that of obtaining remote estimations for some of those variables, rather than physically collecting samples and measurements. Nevertheless, an accurate remote measurement of sample temperature at the range of temperatures taken into consideration (around $800-1700^{\circ}C$), even with a commercial pyrometer, requires as additional input a value for the emissivity of the sample, which is also unknown, since it depends on the composition; as a result, we face an ill-posed problem. At the same time, the closest correlate one can aim at for the chemical composition of the liquid slag (the underlying steel is occluded by the slag and can only be accessed physically) in a context of hot remote samples (where the reflected component is largely masked by the dominant emissive component) would be its spectral emissivity, provided that it was obtained on a wide enough range of the electromagnetic spectrum, preferably comprising ultraviolet, visible, and IR bands up to the far IR zone.

And yet, one cannot simply directly measure the spectral emissivity of a hot, remote sample. From a radiometric perspective, we can only capture -through a spectrometric device- a noisy observation, *i.e.* a proxy of the spectral radiance of such sample, which, at the considered temperatures, is fundamentally dominated by the emissive spectrum of the black body and further affected by unknown and highly variable atmospheric attenuation and the noise and degradations inherent to the optical capturing device. In addition, no spectrometer covers on its own the whole aforementioned spectral range of interest. Should we want to derive the spectral emissivity of the source, we would first need to secure a good estimate of the sample temperature, as the spectral emissivity is defined as the ratio of the spectral radiance of a sample to that of a black body at the same temperature as that sample [275]. Furthermore, Planck's law evidences the exponential dependence of such back-body radiance with the value of the temperature, thus any deviation in its estimation would result in large errors in the predicted spectral emissivity. This is, in essence, a blind inverse problem again, largely unmanageable from a pure supervised data-driven perspective, and a heavily under-constrained one, as there are many combinations of temperature, atmospheric attenuation variables and spectral emissivity profiles that could potentially result in any single noisy measurement of the spectral radiance.

In this thesis, we aim to resolve the lack of suitable signal capturing hardware for EAF environments and circumvent the absence of exploitable datasets for end-to-end learning of a temperature-spectral emissivity separation model for distant hot samples. We instead approach the inverse problem in an unsupervised way, leveraging a physically sound forward model of the signal capture process and a Bayesian approach on the acquired data.

1.2.4 Cross-view semantic segmentation for unlabeled views

Finally, the last image-to-image mapping domain that we focus on in this thesis under constrained labeled data scenarios is the particular case of semantic segmentation in multiple-view setups. Semantic segmentation is one of the computer vision tasks that has most prominently benefitted during the last years from the end-to-end supervised learning approach enabled by CNN models [236], growing from being hardly within reach before up to current highly reliable models [37, 223, 297]. However, certain scenarios still pose relevant challenges:

Figure 1.4: Left: object-cluttered 3D scene captured from a reference camera pose. Center: a complementary camera pose featuring a close baseline with respect to the reference (*e.g.* those in commercial stereo rigs and typical autonomous driving datasets) results in small disparities and little additional insight over the occluded scene parts. Right: a wide-baseline complementary camera location could potentially provide a substantial information gain over the imaged scene, but the fusion of information from both views and the production of cross-view predictions are far from trivial.

- **Multi-view semantic segmentation:** single view semantic segmentation of complex, densely populated scenes is hindered by self and inter-object occlusions that prevent accurate classification of the involved objects, and it could greatly take advantage from additional viewpoints of the same scene. Nonetheless, no trivial solution exists for this setup when baselines between cameras are wide.

- **Cross-view semantic segmentation:** although significant progress has been achieved lately on novel view synthesis [162, 166, 210], the transfer of learned semantic features for inference across views on wide baseline multi-object scenes remains an open research question.

Both fields represent exceptional opportunities to develop solutions that are able to combine our prior knowledge from classic multiple view geometry [93] with the expressive power of contemporary deep neural models. Still, the main inconvenience in this regard so far has been the non-existence of an adequate, large scale image dataset of scenes comprising multiple objects captured from diverse, wide-baseline camera locations (see Fig. 1.4) and with pixelwise semantic annotations. This lack can be largely explained by the disproportionate cost of producing dense labels if compared to creating image-level annotations. Consequently, we assess that data-efficient (*e.g.* semi/weakly/self/un-supervised) approaches to multi and cross-view dense prediction tasks are called to provide significant advancements in these fields.

In this thesis, we attempt to draw attention to the overlooked fields of multi and cross-view semantic segmentation by first providing the community with a useful dataset suitable for large scale data-driven learning and embedding of multi-view geometry-related inductive biases. In addition, we define a new multi-view dense prediction task and introduce a first data-efficient approach to handle it.

1.3 Objectives and approach

Based on the analysis of the challenges identified in section 1.2 we define a selection of precise objectives to be addressed in this thesis for each of the topics discussed above:

1.3.1 Blur detection from few labeled images

Following our assessment that the absence of large labeled blur detection datasets is the main impediment for the development of the field, we define the following research objective:

> **Self-supervised blur detection through synthetic blurring of scenes**: propose a self-supervised framework to bypass the need for large-scale blur localization datasets. Study the combined effectiveness of unsupervised object proposals and synthetic modeling of blur degradations to produce augmented training data.

In Chapter 2 we quantitatively show the limitations of end-to-end fully-supervised approaches and propose a self-supervised framework based on synthetically producing partial defocus and motion blur degradations on images from an unrelated dataset without blur labels. We identify and resolve the issues derived from this approach (e.g. halo artifacts) and we unfold the framework on three distinct instantiations with different levels of supervision: (i) a **weakly-supervised** approach, in which semantic-segmentation ground truth object masks are used as proxy for the image regions to be blurred, (ii) a **self-supervised** approach, in which we leverage unsupervised object proposal techniques to produce multiple semantically plausible partitions of image regions, acting as an implicit form of augmentation, and (iii) a **semi-supervised** approach, in which we combine the latter with actual annotations from one of the few datasets devoted to blur detection. We show that our framework improves state-of-the-art results for two distinct test datasets even without ever observing any real blurred image, and that the addition of a small number of annotated images consistently outperforms the fully supervised approach to the problem.

1.3.2 Exploiting context for hyperspectral image reconstruction from RGB

As we have previously highlighted, most of the existing works treat each sample independently, and there is a void of approaches leveraging spatial information from the local vicinity of the RGB pixels when trying to obtain high spectral resolution reconstructions of RGB images. Concretely, CNNs have never been proposed for the task. Therefore, we propose the following research objective:

> **Spatio-spectral feature fusion for hyperspectral reconstruction**: current approaches for RGB to hyperspectral image recovery disregard the potentially useful information provided by nearby pixels. Hence, we propose the use of a Deep Convolutional Generative Adversarial Network to exploit the spatio-textural information from the local neighborhood while producing plausible reconstructions.

Chapter 3 of this thesis focuses on the development of this approach. We pose hyperspectral natural image reconstruction as an image to image mapping learning problem, and apply a conditional generative adversarial framework to help capture spatial semantics. In order to overcome the lack of independently captured RGB-HSI pairs, we synthetically produce the RGB versions of each of the original hyperspectral scenes taken from the first medium-sized publicly available HSI image dataset (*i.e.* ICVL [6]). By applying well-known, psycho-physically consistent models for color formation, we get sRGB versions of the scenes. The proposed adversarial network system's convolutional generator is in charge of performing the high resolution spectral

recovery, while the discriminator assesses the plausibility of the reconstruction. In chapter 3, we quantitatively show that we can actually benefit from the contextual information that the spatial dimensions can provide for this task by training a deep adversarial dense regression network.

1.3.3 Temperature-spectral emissivity separation of hot samples

Given the enormous complexity of the stated final goal of achieving accurate online remote chemical composition estimations for the hot slag in a steelmaking plant, we choose to focus on securing faithful estimations for temperature and spectral emissivity, which remains an open research question. Consequently, we target two of the identified open challenges, and address them in chapter 4:

> **Robust multi-spectrometer hardware system for remote spectral radiance acquisition**: there is a lack of adequate rugged commercial hardware able to obtain wide range remote spectral radiance raw data from hot remote sources. Thus, we tackle the design and development of our own capture device, which is required to provide fast, reliable and calibrated wide range radiance readings over a reduced size spot from a safe distance from the EAF.

Section 4.3 describes the built portable device, which was conceived integrating the signal received from three independent punctual spectrometers spanning the [0.2,12 µm] range. The device allows for accurate radiance captures over a sample size of 12 mm at 20 m, which is achieved through a collimator and a laser as aid for remote pointing. Critically, a thorough two-step calibration procedure is defined together with the hardware design: a blackbody-based calibration of the spectrometer non-linearities, and an in-field calibration with a calibration lamp to account for potential fiber misalignments and other unaccounted effects.

> **Self-supervised model for simultaneous temperature and spectral emissivity estimation**: we aim at defining and validating a method that can leverage a well-established radiative transfer model within a blind inverse problem modeling framework without extensive available ground truth, so that we are able to estimate both sample temperature and spectral emissivity over captures obtained with the described hardware.

Section 4.4 shows the envisioned approach, which leverages the emerging paradigm of Bayesian probabilistic programming to define a radiative transfer model describing the full emission and noisy capture process in terms of probability density functions over the expected values of the key magnitudes involved. The model comprises the emissive source emitting at an unknown temperature and with unknown spectral emissivity and the spectral characterization of the capturing device, and accounts for the atmospheric attenuation induced by unknown concentrations of the most influential gases to yield full distribution estimates for each of these variables in quasi-real time.

The whole system and probabilistic model were validated under controlled conditions against laboratory-grade equipment for spectral emissivity measurement, using two solid samples with well characterized, stable spectral emissivities (namely alumina and boron nitride). The results and achieved specifications show the suitability of both system and method, which were developed at Tecnalia for ArcelorMittal, for inline in-field use for remote monitoring of the steelmaking process.

1.3.4 Cross-view semantic segmentation for unlabeled views

The conducted analysis of the state-of-the-art raised the problem of an absence of suitable, challenging datasets that could be exploited for cross-view or multi-view dense prediction tasks. Hence, as our first objective we choose to fill in this void:

> **Multi-view, multi-object dataset for semantic segmentation**: create a new synthetic path-tracing based dataset, comprising scenes with multiple, varied objects per-scene and multiple, wide-baseline views where each of them is annotated in terms of semantic segmentation. Release the code and dataset to the community for public access.

In chapter 5, we justify, present and characterize MVMO (Multi-View, Multi-Object dataset): a synthetic dataset of 116,000 scenes containing randomly placed objects of 10 distinct classes and captured from 25 camera locations in the upper hemisphere. MVMO comprises photorealistic, path-traced image renders, together with semantic segmentation ground truth for every view. Unlike existing multi-view datasets, MVMO features wide baselines between cameras and high density of objects, which lead to large disparities, heavy occlusions and view-dependent object appearance. Therefore, we expect that MVMO will propel research in multi-view semantic segmentation and cross-view semantic transfer. In section 5.4, we also provide baselines that show that new research is needed in such fields to exploit the complementary information of multi-view setups.

As a consequence of the nature of this thesis, which was developed in a part-time appointment while executing industry-oriented projects in Tecnalia (see section1.4), we are particularly sensible to problems arising in industrial environments. One of such scenarios would consist on the need for a relocation of the camera in a trained and fully deployed monocular semantic segmentation system, which would most likely result in a severe performance drop. Consequently, we pose the following additional research objective:

> **Semi-supervised cross-view semantic segmentation**: define and validate a data-efficient approach for semantic transfer of dense predictions across wide-baseline views upon the event of a forced camera-relocation.

In chapter 6 we address this by introducing the task of *zero-pair cross-view semantic segmentation*, in which we assume that we can capture an additional set of unlabeled scene image pairs from original and new camera poses. Under this setting, we present ZPCVNet, a model based on the joint training of a set of encoder-decoders that are designed to have a common latent bottleneck representation. At inference time, we can combine aligned encoder and decoders to obtain semantic segmentation results from the desired viewpoints. In our experiments on the MVMO dataset we show that classic geometry-based baselines are ineffective and that ZPCVNet outperforms these and other learning-based baselines by a large margin.

1.3.5 Relation between chapters

This thesis touches upon several applications of different machine learning techniques. To help better understand the points in common among the addressed problems, we provide Table 1.1. All chapters in this thesis address the problem of the scarcity of annotated data as an alternative, more realistic scenario than the end-to-end large scale dataset paradigm. Other than that, Table 1.1 also covers additional characteristics of the proposed methods, which play a prominent role within each of the topics, and are discussed herein.

Three of the chapters pose their respective problems as learning of different image-to-image mapping functions. Another recurring theme is the learning strategies applied in the thesis. The small-data scenario is faced by the use of self-supervised or semi-supervised learning approaches. Each of them makes use of diverse pieces of acquired prior knowledge. Finally, also in most of the chapters, this allows for synthetically generating the input data and/or the corresponding ground truth and learning from them.

Chapter	Small data	Technique	I2I	SfS/SmS	PK/PB	Synthetic
Ch.2	Few/Auto labels	CNN	Yes: 1CR	Yes: SfS	Yes: Blur	GT
Ch.3	Auto labels	CNN	Yes: MCR	Yes: SfS	Yes: Color	I
Ch.4	No labels	PP	No: PM	Yes: SfS	Yes: RT	-
Ch.5&6	Few labels	CNN	Yes: SS	Yes: SmS	Yes: PT	I, GT

Table 1.1: Summary of the characteristics of the contributions covered in this thesis. **Chapter**: Ch.2: Blur detection from few labeled images. Ch.3: Exploiting context for hyperspectral image reconstruction from RGB. Ch.4: Temperature-spectral emissivity separation of hot samples. Ch.5: Multi-view, Multi-object dataset. Ch.6: Cross-view semantic segmentation for unlabeled views. **Technique**: CNN: Convolutional Neural Networks. PP: Probabilistic Programming. **I2I**: Image-to-Image mapping. 1CR: 1-Channel Regression. MCR: Multi-Channel Regression. PM: Punctual Measurements. SS: Semantic Segmentation. **SfS/SmS**: Self-Supervised/Semi-Supervised learning. **PK/PB**: elements in the workflow that leverage Prior Knowledge or Physics-Based modeling. Blur: model for blurred image formation. Color: model for color image formation from spectral radiance. RT: Radiative Transfer model. PT: Path Tracing and 3D scene modeling. **Synthetic**: data components generated synthetically through the mechanism shown in the PK/PB column. I: Input data. GT: Ground Truth.

1.4 Industrial PhD at TECNALIA

In this section, we provide some context on the development of this thesis as an industrial PhD, with a remote, part-time dedication at the CVC and a full-time position at Tecnalia, which is the industrial partner of this thesis.

Employing a workforce of more than 1,400 people (44% women – 56% men) of 31 nationalities, Tecnalia is the largest private center of applied research and technological development in Spain, a benchmark in Europe and a member of the Basque Research and Technology Alliance. Through R&D and Innovation, Tecnalia aims at creating solutions that generate economic impact on companies, prosperity for the country and value for society, in terms of quality of life and progress. Hence, it defines its mission as to "transform technological research into prosperity".

Born in 2011 from the fusion of eight sectorial centers located in the Basque Country, Tecnalia puts forward a multi-sectoral and multi-technological proposal, following the assertion that today's companies need cross-sectorial and multi-technology responses to their global challenges, which are impossible to satisfy without the hybridization of technologies. Tecnalia's main scopes of action are: Smart Manufacturing, Digital Transformation, Energy Transition, Sustainable Mobility, Personalized Health, Circular Economy and Urban Ecosystem. From the perspective of its technological strategy, however, Tecnalia structures them into vertical technologies (with an impact on a single market) and transversal technologies (which may have an impact on more than one market).

Among the latter, the *Computer Vision and Visual Interaction* technology group's *Computer Vision* team focuses on developing advanced, primarily deep learning-powered image understanding systems for applications in production-line quality control, bio-medicicine, agriculture, steelmaking or recycling industries, among others. Remarkably, its activity is based on the analysis of not only conventional color images, but of a wide range of image or punctual signal acquisition technologies, such as HSI, LIBS (Laser-Induced Breakdown Spectroscopy), Raman and FTIR (Fourier Transform InfraRed) spectroscopy, thermography, OCT (Optical Coherence Tomography), MPT (Multi-Photon Tomography), bright-field microscopy, THz, etc. This

enables extending its scope to deal with material and process characterization objectives as well.

Tecnalia's place within the R&D value chain, which is a reflection of its mix of private and both non-competitive and competitive public funding, covers a long range from basic research to fully deployed turnkey solutions in some cases. Consequently, there is a constant tension between the need for being at the edge of the state-of-the-art and being able to yield fully functional implemented systems to the industry, as its success is evaluated via both academic and pure market-driven performance indicators. Therefore, Tecnalia encourages its employees to obtain a PhD on a relevant research topic. Currently, 18.1% of its employees are doctors.

In this context, most of the present thesis was developed with funding from the SPRI-Basque Government's ELKARTEK program, but without an explicit link to any ongoing privately funded project. However, the knowledge obtained from pursuing this thesis has proven to be useful for several projects within Tecnalia, which also has seen a rising interest in deep learning —particularly so in data-efficient deep learning— from the industry in recent years. Chapter 4 was a notable exception to this, as it directly describes the solution conceived and implemented for one of the most prominent real-world problems currently faced by steelmaking industry. This work also led to a European patent application being filed [196] covering its content.

In conclusion, while the part-time dedication and double affiliation regime posed additional challenges on the development of this thesis, it also paved the way for a certain cross-fertilization of the purely academic/industrial spheres, therefore yielding an enriched version of both activities.

2 Self-supervised Blur Detection from Synthetically Blurred Scenes*

2.1 Introduction

Image blur is a phenomenon that degrades the definition of an image, producing a loss of detail in the affected regions. The two main causes leading to a fully or partially blurred image are (i) defocus blur, which is inherent to a wide aperture optical image capturing device that projects scene points that are away from the focus plane onto a non-punctual circle of confusion on the sensor and (ii) motion blur, which is caused by the movement of either the camera (*i.e*. camera shake) or the imaged objects during camera exposure time. Even if both defocus and object-motion blurs are sometimes sought-after as part of a creative photographic process (*e.g*. to pop-out the subject or to evoking a sense of motion, respectively), most of the time blur is considered as an undesired effect [291] or image artifact.

In any case, trying to localize blurred regions within an image (or, equivalently, achieving a segmentation in terms of blurry/unblurred parts) is a useful task with a wide range of applications in computational photography, e.g. defocus blur magnification [13, 81], image deblurring [74, 184, 239, 292], or camera focus point or depth of focus estimation [81]. In addition, due the underlying correlation between blurred and non-blurred areas within the same image, blur detection has been also applied in general computer vision tasks, *e.g*. depth estimation [91], saliency prediction [57] or semantic object segmentation [195]. However, blurred region segmentation is a challenging task, due to the fundamental ambiguities existing between the out-of-focus pixels and the originally flat regions or smooth edges. Scale-ambiguity (*i.e*. the difficulty of inferring the level of blur over one single scale [238]) and the dependence of the perception of sharpness on the image size are additional challenges that affect the performance of current approaches.

A number of previous works have investigated blur localization directly or implicitly from the feature engineering and physical modeling approaches, either taking one single [81, 238, 248] or multiple [66, 299, 300] images as input, and aiming at detecting only one [33, 186, 239, 248, 284, 302, 303] or both kinds of blur [32, 151, 246]. Most of these try to leverage information extracted directly from the intensities [151], from the gradients [185, 246, 302, 305], or from transformed domains [32, 238, 248, 249, 303].

More recently, supervised learning-based approaches [185], and particularly those based on the use of Convolutional Neural Networks (CNN), have shown enormous potential for tackling tasks that require a dense, per-pixel prediction, such as semantic segmentation [36, 236], instance segmentation [94] or crowd counting via density map estimation [153]. Blur segmentation can also be viewed as one of such dense prediction tasks, and several works have already explored this approach, either for predicting both types [121, 159, 291] or defocus only blur [187, 287, 294, 295].

Nevertheless, the performance gain obtained by these fully-supervised, CNN-based approaches trained end-to-end is relatively modest when compared to gains in other fields. The main bottleneck which hinders the full power of CNNs in their application to the field of blur localization is the absence of large enough datasets with pixel-wise annotations. Several recent efforts have partially addressed this void [238, 291, 294, 295], but these manual annotations are scarce, labour intensive and costly to acquire.

In the absence of large public sets of annotated data, the generation of synthetic image degradations has been successfully applied for training deep CNNs in recent years. This approach has been adopted for

*This chapter is based on a publication in Image and Vision Computing journal, 2019 [3].

computer vision tasks as diverse as image super-resolution [142], inpainting [190], image quality assessment [152] or perceptual similarity estimation [290]. It is also a core component of many self-supervised learning methods, which, facing the lack of large sets of annotated data for a certain final task, define a pretext task (*e.g.* inpainting [190], warped image matching [180], artifact spotting [111]) for which ground truth labels can be automatically derived without manual intervention and be used for feature learning. A particular case of such scenario would be that of pretext and final tasks being equal.

Synthetic blur generation has also been explored as part of deblurring workflows [133, 184], as the blur generation process has been extensively studied, especially within deconvolution approaches for image deblurring. In the context of blur localization, however, synthetic blur operations have only been applied either over very simple partial blur masks (*e.g.* consisting of a hard rectilinear partition of the image in halves [295]) or globally over full patches of reduced size extracted from the whole image [202].

Our main contribution in this chapter is the introduction of a deep self-supervised partial blur detection framework, which successfully localizes both defocus and object-motion blur types for a single image without making use of any blur segmentation annotation for training the model. Instead, we circumvent the lack of large annotated blur detection image sets by selectively applying procedural synthetic blurring operations to varying regions within images taken from an unrelated non-annotated dataset of natural images. By controlling the definition of the regions being synthetically blurred, we can automatically generate the associated ground-truth blur masks on the fly. Fig. 2.1 illustrates the process.

The framework comprises three different instantiations, defined by i) how those regions are extracted and ii) whether the training is purely based on synthetically blurred images or not:

1. A **self-supervised** approach, in which the regions to be blurred are defined by an object proposal [129] model, which can generate class-agnostic plausible object masks and thus generalize well across a variety of datasets. This method can extract multiple different blur masks per image, so that the model learns from different image and blur-mask pairs. This allows us to obtain a highly variable training image stream to feed our CNN-based semantic segmentation model at train time.

2. A **weakly-supervised** approach, in which such regions are instead selected from the ground truth labels of any given semantic segmentation image set.

3. A **semi-supervised** approach, in which the synthetically generated blurred image and labels from the object proposal method are used in conjunction with real partially blurred images and their respective manual blur-mask annotations in order to augment the given fully supervised blur segmentation dataset.

By extensively evaluating the learned model on the two largest publicly available pixel-wise annotated real blur localization datasets [238, 294], we show that our approach generalizes adequately for both blur types, improving the performance of all the considered classic approaches and that of recent fully supervised deep CNN-based ones that require large human-labeled training sets. Experimental results show that the proposed solution can be successfully employed to train a deep blur segmentation model, either without the need for any specific blur localization dataset or by making use of a very reduced set of images exhibiting real blur degradations. This is especially important for domains other than that of natural RGB images, for which no labeled blur masks exist, such as infrared or multispectral [38] imagery, medical imaging [156] or text documents [161].

2.2 Synthesizing realistic blur

In this chapter, we are interested in learning to localize blur on partially blurred images without using any labeled example for this specific task (in the self-supervised and weakly supervised approaches). We therefore

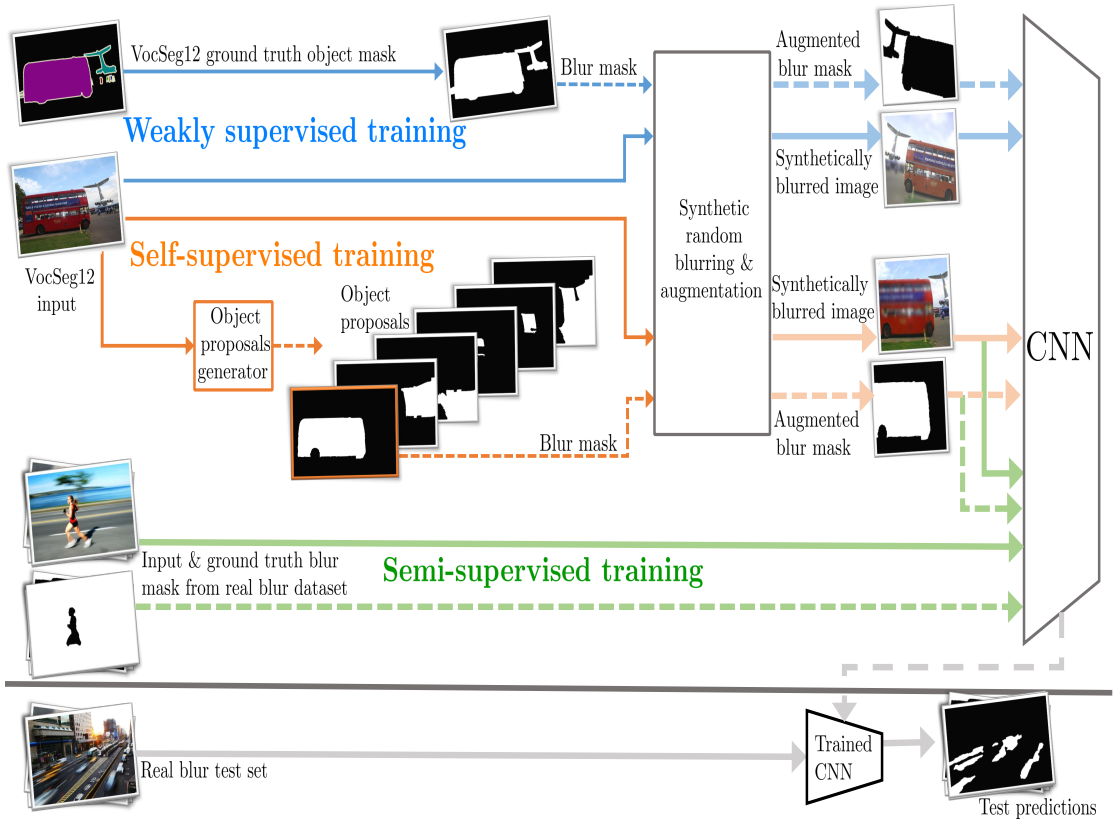


Figure 2.1: General overview of our framework train and testing processes, with each path color representing one of its three possible instantiations *i.e.* self-supervised, weakly-supervised and semi-supervised approaches.

switch to the problem of generating plausible blurred scenes from non-blurred images.

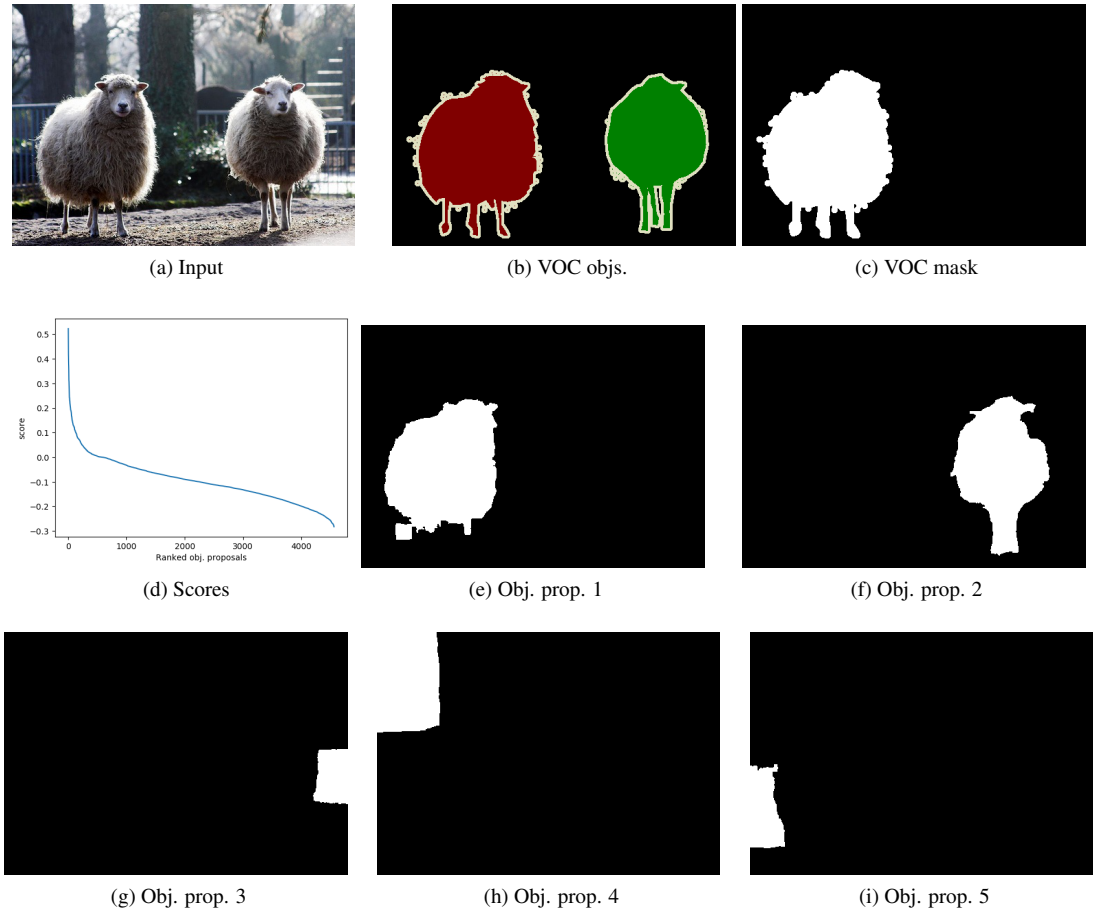
2.2.1 Blur mask extraction

The first step of the process is the determination of the parts of the image that will be subject to the synthetic blurring operation, to which we refer as the *blur mask*.

Semantic object masks as blur masks

One possible approach to achieve this is that of using a dataset of images on which several objects have been manually segmented, without any explicit relation to its blur content. This effectively implies making use of a distant supervision (one grounded on easier to obtain semantic object annotations) for the final blur segmentation task. We refer to this as our weakly supervised approach.

This kind of labeled data is readily available from datasets generated for semantic segmentation tasks, such as the Pascal VOC Segmentation challenge 2012 (SegVOC12) [65]. This dataset includes ground truth annotations of the most prominent objects present in the image (see Fig. 2.2b) corresponding to 20 different



(a) Input (b) VOC objs. (c) VOC mask

(d) Scores (e) Obj. prop. 1 (f) Obj. prop. 2

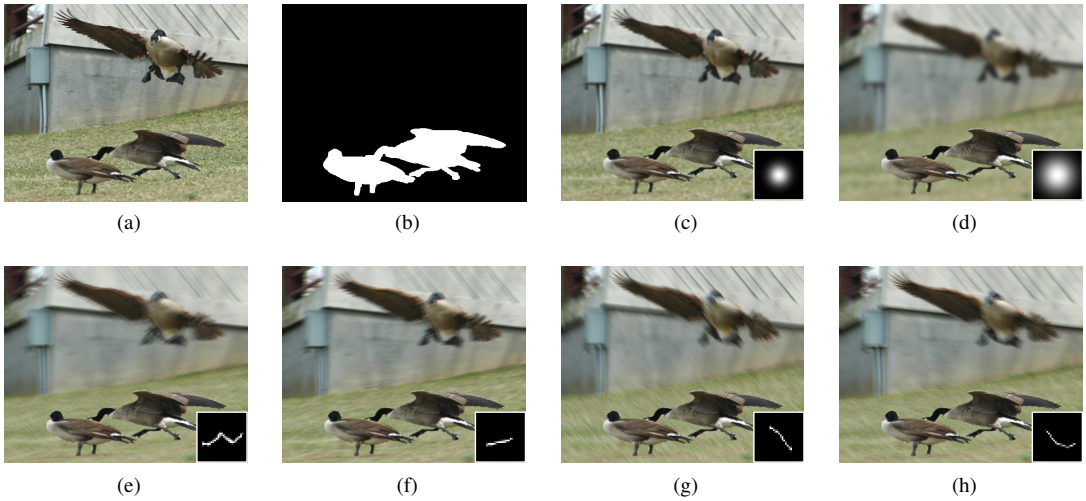
(g) Obj. prop. 3 (h) Obj. prop. 4 (i) Obj. prop. 5

Figure 2.2: Blur mask extraction from an input image (a) of the Pascal VOC 2012 dataset [65]. (b) Ground truth object masks from the segmentation challenge. (c) Blur mask given by the largest connected component. (d) Sorted scores of objectness given by MCG for this input. (e-i) First five object proposals generated by MCG [199].

classes. Under this setting, at train time, we build the blur mask for each input image by computing the connected components of the largest object present in the corresponding ground truth (Fig. 2.2c).

Object proposals as blur masks

Our purely self-supervised method takes one step ahead by removing the need for manual object segmentation, and replacing it with the inclusion of a class-agnostic object mask proposal generation step. The goal of object proposal generation methods is, given an input image, to yield a set of either bounding boxes or segmentation masks that correspond to different object location hypotheses. Its primary application is serving as a first candidate location filtering stage of two-phase object detection methods, so that more resources can be allocated for the representation and analysis of the resulting subset of regions.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.3: (a) Input image from the VOC dataset (b) Segmented Foreground (c-d) Blurring of input image with Gaussian blur, $\sigma = 1.5, 3$ pixels (e-h) Blurring of input image with randomly-generated non-linear motion blur. Blur kernels displayed at the right-bottom on each image. All images have been blurred with the halo-artifact removal described in section 2.2.3. Blur differences are better appreciated when focusing in the roof on the top of the image.

The Multiscale Combinatorial Grouping (MCG) object proposal generation algorithm [199] represents one of the most accurate approaches of its kind. Moreover, although some learning-based steps are involved in its model creation process, its ability to generalize across different datasets renders this method virtually parameter-free. It yields a ranked set of segmented object proposals after a process that comprises: (i) a multi-scale input image segmentation step (based on low level features), (ii) a rescaling and alignment of the segmentation results, (iii) the combination of such results onto a merged multi-scale hierarchy of binary spatial image partitions (*i.e.* regions) and (iv) a final combinatorial grouping stage, which explores the region tree looking for sets of regions that, merged together, are likely to represent complete objects. The resulting set of -hundreds of- proposals is then ranked in accordance to a score representing such likelihood. During training, we apply MCG to every input image and, at each epoch, we randomly sample a blur mask from the probability distribution given by applying a softmax operation over the mentioned scores.

Fig. 2.2 shows an example of blur mask extraction for both ground truth semantic object mask (b-c) and MCG-based object proposal (e-h) methods. It is important to note that, should we directly use the blur mask extracted as described in either case, we would be introducing a strong bias in the blur detection training, favoring the prediction of blurred elements in the foreground or background of the images. This is due to a comparatively significant amount of the extracted blur masks corresponding to foreground objects. In order to mitigate this and promote the invariance of the model to this respect, we invert the blur mask with a probability p_{inv}.

(a) (b) (c) (d) (e) (f)

Figure 2.4: (a) Original image from the VOC dataset (b) Inpainted foreground (c,d) Naively blurred background (e,f) Result of blurring after inpainting background.

2.2.2 Synthetic blur

Once the input blur mask has been created, we can now randomly apply different kinds of synthetic blur operations over it. Given an image $I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ for which the image domain Ω has been already partitioned into background Ω^B and foreground Ω^F, represented as the blur mask, we generate a partially blurred version of I by first defining a blur kernel K. Then, the resulting artificially blurred image I_b can be easily obtained by computing:

$$I_b(x,y) = \begin{cases} K * I(x,y), & \text{for } (x,y) \in \Omega^B, \\ I(x,y), & \text{for } (x,y) \in \Omega^F. \end{cases} \tag{2.1}$$

The adequate definition of the kernel K is critical in order to accomplish the goal of generating realistic blur. While blur coming from situations on which an object is in-focus and the background is defocused can be easily simulated by Gaussian kernels K_σ of varying standard deviations σ, blurs of different natures, *e.g.* motion blur, are harder to emulate. Since our aim is to produce fast blurred versions of a given background on the fly, we design a simple pipeline for generating non-linear motion blur:

1. Build a horizontal line of length $1 \times m$ in a discrete domain $\Omega_k \subset \Omega$, obtaining a kernel K_m that defines a linear horizontal motion blur.

2. Rotate it by α degrees, obtaining a kernel $K_{(m,\alpha)}$ that defines a linear diagonal motion blur.

3. Apply an elastic deformation $\mathscr{E} : \Omega_k \rightarrow \Omega_k$ of the underlying image grid Ω_k that turns $K_{(m,\alpha)}$ into a non-linear kernel $K_{(m,\alpha,\mathscr{E})}$.

In training time, a random decision of whether to apply defocus blur or motion blur is made for each training image. The definitions of σ or $(m, \alpha, \mathscr{E})$ are also randomized by drawing parameters from a suitable range of values. In this way, the same image can be transformed in infinitely many ways.

In Fig. (2.3a) we show one image from our Pascal VOC training set. In Figs. (2.3c) and (2.3d) the background, as given by Fig. (2.3b), is blurred with two Gaussians of varying standard deviations, whereas in Figs. (2.3e–2.3h) different non-linear motion blurs are applied. It should be noted that there exist more sophisticated mechanisms to simulate non-linear image blurring based on physical considerations [133]. However, for the purposes of this chapter we prioritize a simple and efficient strategy like the one described above.

Note that, although ideally we would want to use perfectly sharp images to perform the selective blurring, the employed SegVOC12 dataset does indeed contain some pictures exhibiting preexisting partial blur. This means that the artificially generated ground truth will contain a certain amount of noisy pixel-level labels. In spite of this, as shown in section 2.4, our framework is able to successfully learn to segment real blur.

2.2.3 Removing halo artifacts by inpainting

The naive approach of blurring the image, and then placing back the unblurred foreground on its original location has some disadvantages. Namely, this procedure leads to the appearance of halo artifacts around the borders of the foreground, as shown in Figs. (2.4c) and (2.4d). The reason of this problem is that sharp intensity jumps at borders of foreground objects artificially distort image statistics when averaging with a blur kernel around those pixels.

If a model is trained with images containing these halos, it will likely learn to localize blur by simply finding the position of such artifacts. In order to avoid this situation, we propose a different approach to obtain a blurred-background image. After extracting the sharp foreground from a given image, we proceed to inpaint the foreground pixels applying the method from [250], as shown in Fig. (2.4b), before blurring the background. This process allows to remove halo artifacts from the artificially blurred scenes before supplying them to our model, see Figs. (2.4e) and (2.4f).

2.3 Convolutional Neural Networks for Blur Segmentation from Synthetic Data

2.3.1 Architecture

Section 2.2 provides us with all the necessary tools to create a procedural -and thus, highly variable- training image flow. By posing blur localization as a dense, per-pixel classification task, we can make use of one of the well-established set of deep architectures devoted to semantic segmentation and feed it with such synthetically distorted image stream for training. We consciously avoid the *ad hoc* design of an architecture specifically suited for our target task, and instead rely on an off-the-shelf deep network, which allows us to isolate the contribution derived from the proposed training procedure.

We select the DeepLabv3 [36] Network as our reference CNN architecture, since it has recently shown state of the art performance on various dense labeling tasks. This is partly due to the fact that it features an effective receptive field significantly larger than those of other standard architectures. These are enabled by the use of *atrous* convolutions (first introduced in this context in [35, 286]), which also contribute to the attainment of output maps with a large spatial resolution. It has been previously shown that high semantic level features can greatly contribute to solve the task of blur localization [159]. Hence, we can expect that the use of wide receptive fields, covering significant parts of the scene, will bring important benefits to our technique. The DeepLabv3 architecture also benefits from the fusion of feature maps from multiple scales (which has proven crucial for solving the scale ambiguity problem [121]): the use of a new *Atrous Spatial Pyramid Pooling* (ASPP) module helps capture context at various ranges. In this chapter, we employ the DeepLabv3-ResNet101 variation of the architecture, which is constructed by a Deeplabv3 model with a ResNet-101 backbone. We start from a model initially pre-trained on a subset of the COCO train2017 dataset [148] containing the 20 classes also present in the Pascal VOC 2012 segmentation dataset, and we substitute its final classifier with a $1 \times 1 \times 256 \times 2$ $2D$ convolution, followed by a $logSoftmax$ operation.

2.3.2 Training procedure

Our training set for the purely self-supervised and weakly supervised approaches is generated by randomly applying synthetic defocus and motion blurring operations (generated as described in section 2.2) to image batches (with a batch size of 18) from the training set of the SegVOC12 dataset [65]. Such degradations are selectively applied according to the binary mask created on the fly as mentioned in sections 2.2.1 and 2.2.1

for the ground truth semantic object masks and object proposal-based mask, respectively, and the blur mask itself is used as the ground truth label.

Additionally, the usual standard random data transformations (affine transforms, flips, color jitter, cropping) are applied to the input and target image pairs before resizing them from their original size to 224×224 pixels. A randomly applied JPEG compression-based augmentation is also performed at different stages of the input image preprocessing workflow, with the aim of gaining invariance to low level, dataset-specific regularities.

We employ a negative log-likelihood loss that we minimize using the Adam optimizer, and let the model train until the validation loss stagnates for 20 epochs. A reduced number of hyperparameter tuning configurations was tried for each of the experimental setting, and the setup yielding the lowest validation loss was kept for evaluation. In most of the configurations, this corresponded to a learning rate value of 10^{-5} and a weight decay of $5 \cdot 10^{-4}$.

2.4 Experimental Results

We evaluate the trained model on the largest publicly available[†] dataset of images annotated in terms of defocus and object-motion blur localization, i.e. that of Shi *et al*. [238][‡]. The dataset consists of 1000 partially blurred natural images with human-made binary (blur/no-blur) pixel-wise annotations, of which 704 correspond to defocus blur and 296 to motion blur-affected images. Following [159], we partition the dataset in an odd and an even subset, both of them containing an approximately equal amount of images affected by both types of blurs), and keep the latter held out for testing purposes. Unless otherwise stated, all the results shown in this section were thus evaluated on the 500 images from Shi *et al*.'s even subset. Meanwhile, a 20% of the odd subset (100 images) is used for validation, and the remaining 400 images are used for training in those experimental setups that require a supervised training component (i.e. our semi-supervised approach and the fully supervised model from Table 2.1 and Fig 2.6).

At inference, we run a simple test-time augmentation (TTA) process, consisting of averaging the predictions yielded by the original input image, together with its horizontally flipped version, and upscale the resulting 224×224 pixels-sized predictions to the original, varying image sizes before performing the evaluation. Even though the raw performance values could, to some extent, benefit from performing a re-scaling of the resulting blur map in the $[0, 1]$ range, we purposefully avoid such kind of post-processing, in order to enable the prediction of completely sharp or blurred scenes. Consequently, we can consider the values of the predicted mask pixels as a measure of absolute blurriness, as opposed to a measure of blur level of each region with respect to the sharpest area of the image.

2.4.1 Self-supervised setup

We first consider the purely self-supervised instantiation of our framework, in which we directly test the model trained on synthetically blurred images over the 500 even samples of the test dataset. We compare our method against most of the best performing hand-crafted feature-based approaches which do not require any dedicated dataset for training: Liu *et al*. [151], Chakrabarti *et al*. [32], Su *et al*. [246], Shi *et al*. [238], LBP [284] and HiFST [81]. In addition, we include the performance reported for the same even subset by one of the most recent deep CNN-based defocus and motion blur detection methods in the literature, *i.e*., the *Deep Blur Mapping* approach by Ma *et al*. [159].

[†]Zhang *et al*.'s dataset [291] and Zhao *et al*.'s supplementary train set [295] are not open.

[‡]Pretrained models are available on https://github.com/aitorshuffle/synthblur.

Method	AUC			AP		
	Defocus	Motion	All	Defocus	Motion	All
Liu *et al*. [151]	0.722	0.714	0.720	0.792	0.683	0.760
Chakrabarti [32]	0.745	0.640	0.714	0.837	0.675	0.789
Su *et al*. [246]	0.807	0.750	0.790	0.859	0.707	0.814
Shi *et al*. [238]	0.836	0.735	0.806	0.876	0.699	0.823
LBP [284]	0.855	0.678	0.802	0.876	0.683	0.819
HiFST [81]	0.901	0.804	0.873	0.928	0.744	0.874
Ma *et al*. [159]	0.947	0.861	0.922	0.966	0.784	0.912
Ours self-supervised	0.945	0.905	0.933	0.960	0.838	0.924
Ours weakly supervised (segmentation masks)	0.941	0.897	0.928	0.959	0.849	0.926
Ours semi-supervised (joint with 400 odd img.)	0.956	0.904	0.941	0.974	0.840	0.934
Fully supervised (finetuned to 400 odd img.)	0.943	0.875	0.923	0.965	0.819	0.922

Table 2.1: Quantitative evaluation over Shi *et al*.'s dataset's [238] even partition. Best, 2nd best and 3rd best results are highlighted for each metric and blur type.

Table 2.1 shows the obtained results in terms of Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) and Average Precision (AP), as computed over the different values of recall given by the Precision-Recall (P-R) curve. Both the AUC and AP values were computed individually for each output map and then averaged over the whole even test subset. The overall performance value (*All*) is also shown disaggregated for the *Defocus* and *Motion* subsets of the database. Note that, as is the case with most of the other methods, in absolute terms our approach performs better over the defocused images than over the motion-blurred ones. This will hold true for the results of all the three variants of our framework.

From Table 2.1 we can see that our self-supervised approach performs significantly better than all the other non-deep methods for every metric and blur-type subset. Furthermore, the proposed self-supervised learning method, when used to train the off-the-shelf DeepLabv3_resnet101 network and without ever observing a single image with real blur, yields better overall AUC and AP values than Ma *et al*.'s CNN architecture [159], whose design was tuned *ad hoc* for this task and trained end-to-end in a fully supervised setup over the 500 odd samples of the dataset[§]. In particular, our method performs close to Ma *et al*.'s on the defocus blur subset, but significantly better for the motion blur samples. The last row of Table 2.1 (*i.e*. *Fully supervised*) shows the results obtained by fine-tuning the DeepLabv3_resnet101 network over Shi's odd subset. The mixed results of this fully supervised model when compared to Ma *et al*. suggest that the performance gain of our self-supervised method is largely due to the benefits of our training scheme, rather than being just a product of the use of a better architecture.

Other relevant fully supervised CNN based results were left out of this comparison for various reasons: Zhao *et al*. [294] target exclusively images degraded with defocus blur, and their training procedure involves (i) pre-training the model employing very simple artificial defocus blurring operations over half of the image on samples from additional datasets and (ii) further fine-tuning it over 604 out of the 704 images of Shi *et al*.'s defocus blur partition and testing it over the remaining 100. Finally, Zhang *et al*. [291] train their *ad hoc*

[§]The performance evaluation protocol in [159] is slightly different, as they compute a single AP value for a one-dimensional vector containing the predictions of all the pixels of every image in the even subset. Although we believe that the AUC/AP values should be computed on a per-image basis, under their protocol our self-supervised overall AP is 0.952, vs. their reported value of 0.880

	AUC			AP		
	DF	MT	All	DF	MT	All
DF	**0.949**	0.826	0.913	**0.967**	0.773	0.910
MT	0.934	0.894	0.922	0.953	0.831	0.917
All	0.945	**0.905**	**0.933**	0.960	**0.838**	**0.924**

Table 2.2: Blur type based ablation test over Shi *et al*.'s dataset's [238] even partition, in our self-supervised setup. Rows represent the synthetic blur type being applied on training (DF=Defocus, MT=Motion, All=Defocus and Motion). Columns represent the test (sub)set. Bold is best.

designed *ABC-FuseNet* architecture end-to-end on their unreleased *SmartBlur* blur segmentation dataset of 10.000 images before evaluating on Shi *et al*. Even with such amount of training samples, their reported AP is 0.869 for the whole dataset, sensibly below our results on the even partition.

Fig.2.5 contains visual results for a small random subset of images affected by both types of blur (defocus blur in the top seven rows, motion blur in the bottom seven), as predicted for most of the considered methods. We can observe that, even without the utilization of any single ground truth blur segmentation annotation from the target dataset for direct supervision, our self-supervised approach obtains accurate masks, comparable in visual quality to those produced by fully-supervised deep CNN-based methods, such as [159].

Blur type ablation Finally, Table 2.2 shows the results obtained over the same sets when only defocus blur or only motion blur synthetic degradations where applied during training in our self-supervised setup. The results suggest that devoting the full capacity of the network to learning to detect only one specific type of degradation does help in the case of defocus blur training, but not so when the training is constrained to observing samples affected solely by object motion. For the latter, the increase of variability introduced by feeding the net with both kinds of degradations seems beneficial, probably due to a regularization effect. Cross-blur type evaluation is asymmetric: while training on motion-only blur achieves decent results on the defocus-only test subset, a model trained uniquely on defocus blur synthetic degradations suffers a significant performance degradation if applied to object motion-affected images.

In addition, this experiment reveals one of the advantages inherent to our blur detection framework: by selectively tunning the ratio of images being synthetically blurred with defocus or motion kernels during training, we can prioritize the performance over either kind of blurs, or operate anywhere in between.

2.4.2 Weakly supervised setup

The *Ours weakly supervised*-labeled row from Table 2.1 shows the results achieved with our weakly supervised approach, based on the use of manually annotated semantic object segments as blur masks instead of the unsupervised proposals yielded by the MCG algorithm. While one could intuitively think of this as an upper bound for our self-supervised method's performance, we observe that, in fact, both methods perform almost on pair, with the self-supervised approach yielding slightly better results for the images affected by defocus blur and with no clear strongest method in terms of motion-blur and overall results, depending on the metric of interest. This experiment shows that applying blur degradations to object proposals computed with an object proposal method does not negatively impact our results, and that similar outcomes are obtained when compared to a method trained with ground truth segmentation masks.

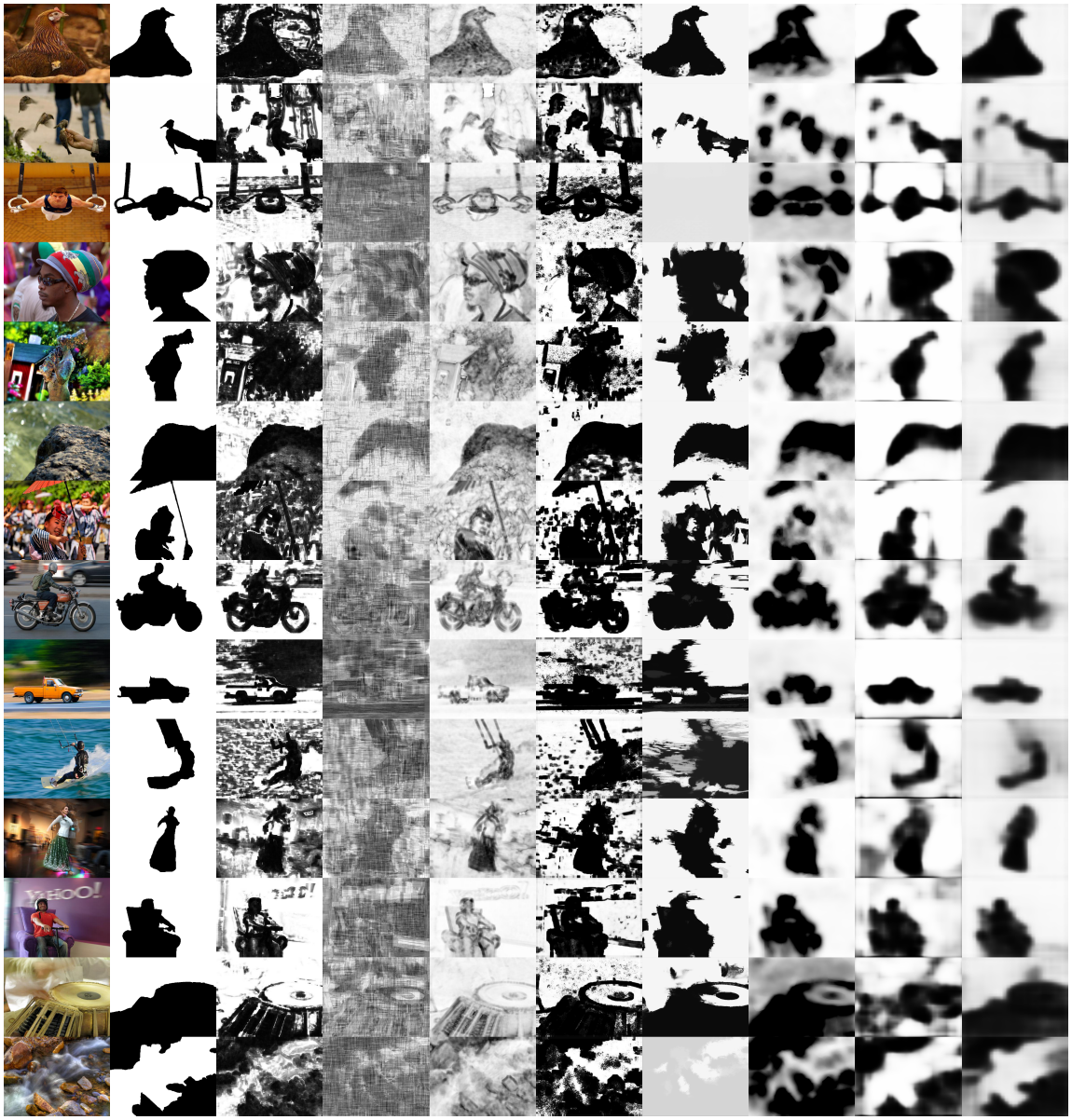
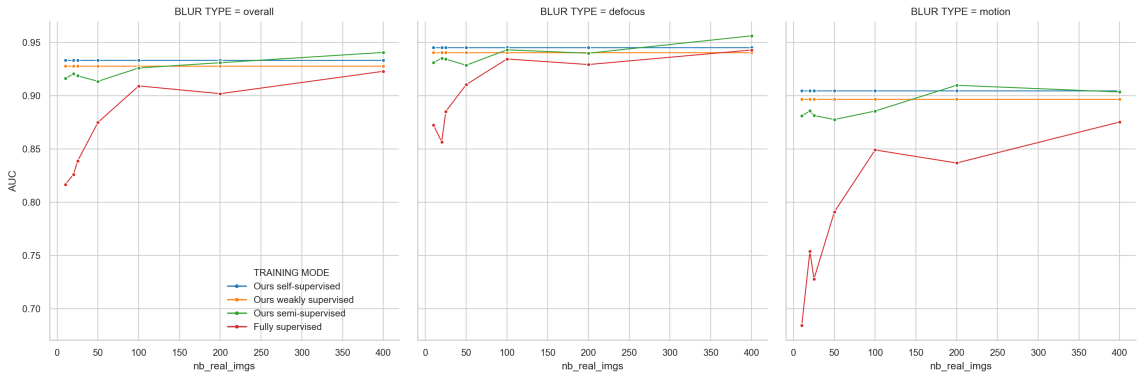


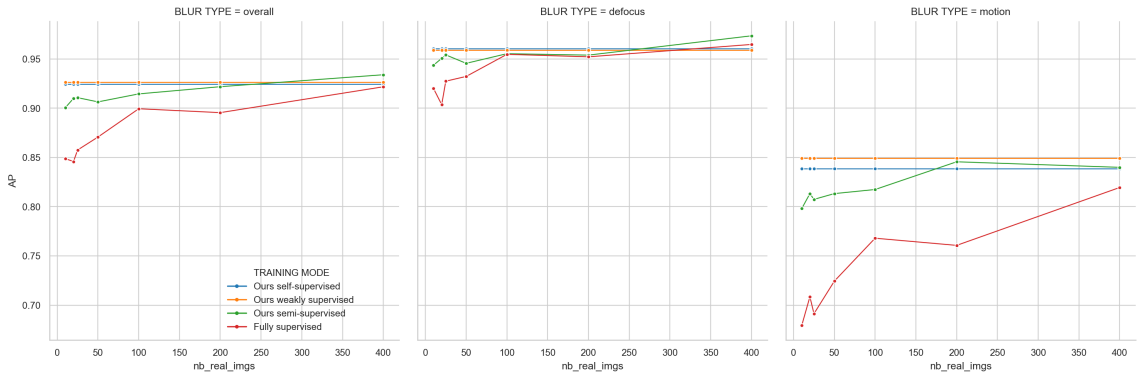
Figure 2.5: Qualitative results for a sample of images from Shi's dataset [238] affected by defocus (top 7) and motion (bottom 7) blur processed by the different evaluated algorithms.

2.4.3 Semi-supervised setup

We now introduce a modification in the training process in order to test the usefulness of our proposed synthetically blurred image-based training when a limited number of blur segmentation annotations are



(a)



(b)

Figure 2.6: (a) AUC and (b) AP as a function of the number of images with real blur from the Shi *et al.*'s dataset [238] used in the training process, in the following setups: (i) Joint training on images with synthetic and real blur (*Ours semi-supervised.*) (ii) Direct fully-supervised fine-tuning on images from Shi *et al.* (*Fully supervised*). The following setups are shown for comparative purposes, but do not use any image from Shi *et al.* : (iii) MCG object proposals-based self-supervised training (*Ours self-supervised*), (iv) SegVOC12 semantic segmentation masks-based weakly supervised training (*Ours weakly supervised*). The set of images with real blur ar part of Shi *et al.*'s odd subset, containing both defocus and motion blur. All the experiments were done using the DeepLabv3 architecture [36] with a Resnet101 backbone.

available for the target dataset. This corresponds to our semi-supervised experimental setting, in which a joint training is performed: mini-batches are now composed of equal amount of image and ground truth blur mask pairs produced (i) in a MCG-based synthetic blurring operation, and (ii) sampled from the train subset of the target dataset [238].

In order to assess the usefulness of this semi-supervised setting, we consider the 400 images from Shi *et al.*'s training set that were previously separated and employ a varying fraction of them for joint training with the synthetically blurred images. Specifically, we conduct the experiment with a 2%, 4%, 5%, 10%, 20%, 40% and 80% of the odd part being used as training aid, which, in absolute terms, correspond to 10, 20, 25, 50, 100, 200, and 400 images, respectively. The evaluation protocol is not affected, and the trained model is then

tested on the 500 even pairs of the dataset.

We compare our semi-supervised results with those presented in the preceding sections (i.e. *Ours self-supervised* and *Ours weakly supervised* approaches from Table 2.1), whose performance metrics do not vary with the number of real images being used for joint training. Finally, for comparison we extend the fully supervised approach presented in section 2.4.1 by fine-tuning the same deeplabv3_resnet101 model over the same fractions of the odd train partition.

Fig. 2.6 shows the AUC and AP values obtained for each of the aforementioned methods, both as overall metrics and with disaggregated values for defocus and motion blur affected images. We observe that:

- For all the considered amounts of annotated images, our joint, semi-supervised approach outperforms the fully supervised one by a significant margin for both blur types, especially so as we operate with few real samples. This means that the use of synthetically generated blurred image-ground truth pairs has proven useful as additional source of training data for improving the performance of fully supervised blur detection approaches trained end-to-end.

- The concrete AUC and AP values obtained by these variations for 400 real blur images were also added, for comparison, to Table 2.1. As shown there, the semi-supervised training scheme achieves the best overall results and, for every subset and metric, there is always one of the three instantiations of our framework outperforming every other deep fully supervised alternative. In particular, our best method (semi-supervised setting) reaches an overall AUC of 0.941 and an AP of 0.934, 0.019 and 0.022 points better than Ma *et al*.'s, respectively.

- Even the self-supervised or the technically simpler but more annotation-dependent weakly supervised variants of our framework can clearly outperform both their semi-supervised counterpart and, most notably, the fully supervised training in the lower part of the range.

- This gap is closed by the fully supervised training scheme only for the defocus blur subset, as we keep adding images with real blur, but not so in the case of the motion blur subset.

2.4.4 Cross-dataset generalization

One of the most frequent limitations of current end-to-end trained CNN-based solutions to many computer vision tasks is the inability of models trained on a certain dataset to generalize to other datasets with some underlying distribution shift, either revealed as some readily apparent visual difference (e.g. illumination, object appearance) or due to some hard-to-perceive low level statistical regularities within the dataset. Domain adaptation solutions [259] aim at mitigating the harmful effect of such domain shifts, but they are frequently based on domain-adversarial training strategies that can be cumbersome to implement, and mostly ineffective in providing a significant performance gain when both domains are easily told apart.

The following experiment seeks to evaluate this cross-dataset generalization ability for our self-supervised approach as compared to that of other methods. To that end, we introduce the dataset of defocus blurred images provided by Zhao *et al*. in [294], which will serve for testing purposes. Table 2.3 shows the results obtained by Zhao *et al*. themselves, along with Ma *et al*.'s and the aforementioned fully supervised DeepLabv3 model. The reported values are all result of the direct application of the models learned over Shi *et al*.'s dataset. These are further compared with our weakly-supervised and self-supervised training methods, showing that the latter exhibits a significantly better generalization ability (0.950 vs. 0.923 of AUC for Ma *et al*.'s approach).

As a general conclusion, we show that synthetically blurring parts of images with a certain semantic coherence is, on its own, a useful technique to perform self-supervised or weakly supervised blur localization,

Method	AUC	AP
Zhao *et al*. [294]	0.913	0.946
Ma *et al*. [159]	0.923	0.956
Ours self-supervised	**0.950**	**0.976**
Ours weakly supervised	0.915	0.953
Fully supervised	0.904	0.952

Table 2.3: Direct testing of models from Table 2.1 on Zhao *et al*.'s defocus blur dataset [294], together with Zhao *et al*.'s [294] own results. None of the models in this table have seen Zhao *et al*.'s dataset during training. Bold is best.

and that its use as aid when performing supervised end-to-end training (even in extreme few-shot cases) can help boost detection accuracy. Those domains where annotated data is scarce can particularly benefit from such approaches when applied in conjunction with other downstream computer vision tasks. Histological imaging (to focus a potential disease classification model on sharp parts of a digitized slide), multispectral imaging (where a significant, hardly avoidable blur effect is often found due to the chromatic aberrations derived from the large bandwidth of the captured spectra, and it is difficult to tell it apart from motion or defocus blur effects) or document scanning are some examples of such situations where a self-supervised approach could have a large impact.

2.5 Conclusions

This chapter presents a framework for deep defocus and object motion blur segmentation built upon the procedural application of both types of synthetic blurring distortions over regions of images. The self-supervised and weakly supervised versions of the framework exploit different ways of obtaining the candidate blur masks for automatic ground truth generation, and can be applied without any blur localization annotation. In the semi-supervised case, this source of data augmentation can leverage the availability of a few labeled images to further improve the obtained segmentation accuracies. Extensive quantitative and qualitative experiments show that a segmentation CNN trained on this kind of synthetic data under any of the three mentioned framework configurations is able to accurately localize the blurred regions of a hold-out set, showing performances well above other recent CNN-based approaches.

3 Adversarial networks for spatial context-aware spectral image reconstruction from RGB*

3.1 Introduction

Hyperspectral (HS) imaging has gained relevance over the last couple of years in the applied vision community. Remote sensing, UAV-based imaging, precision agriculture or autonomous driving are only some of the fields that are already benefiting from the use of imaging devices that provide a response that spans the spectral dimension with narrow-band channels to produce an image with higher spectral resolution than the standard RGB trichromatic one.

While the evolution of HS imaging devices has undergone major breakthroughs, it is also true that there is still a trade-off inherent to the fact that we are ultimately capturing three dimensional information with a two dimensional sensor, which limits the quality or resolution of the acquired signal in either of those dimensions: spatial, spectral or temporal. On top of that, the cost of such devices is orders of magnitude above that of conventional RGB cameras.

In this context, HS signal reconstruction from broadband or limited acquisition channels (typically, from RGB sensors) arises as a natural computational alternative, either to compete against native HS systems or to be included as part of their signal post-processing backends. The spectral reconstruction problem is a severely underconstrained, highly non-linear one, and the algorithms trying to solve this mapping should exploit the low dimensionality of the natural HS images [31] and learn informative priors of diverse forms from real world object reflectances, to be leveraged in the reconstruction phase. Note, however, that most of the existing solutions handle each pixel individually. By doing so, they are not taking advantage of the latent contextual information available in the spatially local neighborhood [31].

Generative adversarial Networks (GAN) are a class of neural networks which have shown to be able to successfully generate samples from the complex manifold of real images. In this chapter, we use this class of algorithms to learn a generative model of the joint spectro-spatial distribution of the data manifold of natural HS images and use it to optimally exploit spatial context information. To our knowledge, this is the first time Convolutional Neural Networks (CNN) are used in the task of spectral reconstruction of natural images. We quantitatively evaluate our approach on the largest HS natural image dataset available to date, i.e. ICVL, by comparing against [6], and show error drops of 33.2% (RMSE) and 54.0% (relative RMSE) over their state of the art results.

3.1.1 Related work

A number of works are relevant to the proposed approach. This task was first addressed by isolating its spatial component and focusing on the reconstruction of homogeneous, well-established reflectances of real world surfaces such as Munsell chips, either from multispectral, RGB components [97] or from the tristimulus values [2, 11].

Initial attempts on the spectral reconstruction of natural images from full size RGB input required additional constrains or multiple input forms to help in their task: [117] and [26] use the aid of a low

*This chapter is based on a publication in the IEEE International Conference in Computer Vision Workshops (ICCVW), 2017 [4]. The method was also used in our submission to the NTIRE 2018 Challenge on Spectral Reconstruction from RGB Images [7]

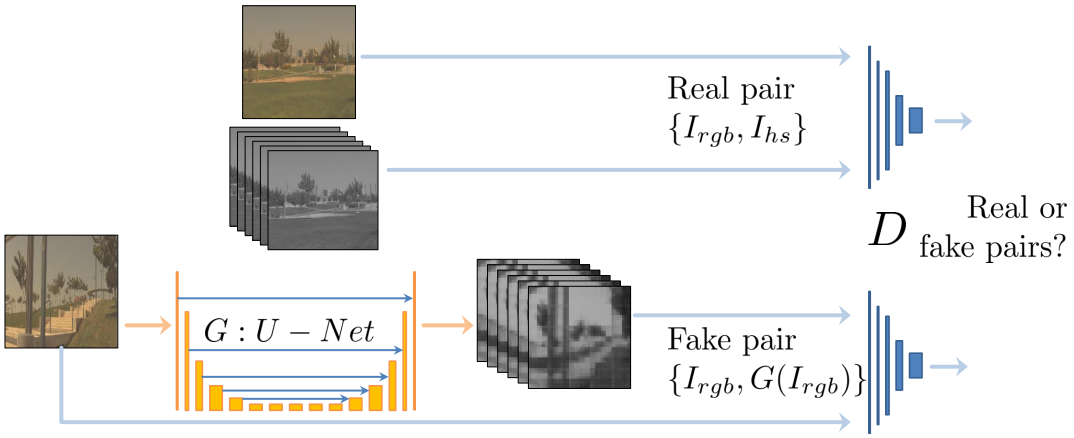


Figure 3.1: Adversarial spatial context-aware spectral image reconstruction model.

resolution HS measurement in addition to the RGB input, [157] restricts to the skylight samples domain, and [80, 188, 189], among others, rely on the aid of computational photography-like multiplexed narrow band lighting. The latter does, however, use spatial information for learning, as does [31], which focuses on the statistics for this class of images and defines a representation basis and computation method for the associated coefficients, but does not tackle reconstruction.

Solutions relying on a single RGB image input at test time are scarce, and almost none of them leverage the spatial context: [176] uses a Radial Basis Function network and produces an estimate of scene reflectance and global illuminant, but assumes a known camera color matching function, and directly depends on the performance of a white balancing stage as part of the workflow. [296] presents the *matrix R method* for spectral reflectance reconstruction, which additionally requires a calibration target to build a camera model. [6] learns a sparse dictionary of HS signatures as bases for the reconstruction. By treating each pixel independently, the ability to use the surround information is lost e.g. for producing distinct spectral outputs for metameric RGB pairs dependent on the context.

Remarkably, [220] exploits spatial material properties of the imaged objects by extracting not only spectral, but also convolutional features resulting from the application of the filter banks from [261], and adopting a constrained sparse coding-based reconstruction approach. In parallel to our development, we found a similar approach [72] which makes use of a CNN-based encoder-decoder to address this task.

Finally, there exists a certain relation between the HS reconstruction and the image colorization [42] tasks, which has been previously addressed in a similar fashion [108, 289], but under different evaluation requirements. We can think of the former being a generalization of the latter for an arbitrary number of input/output channels.

None of these methods would have been possible without the existence of publicly available HS natural image datasets. Until recently, the amount of images per set was the limiting factor for the development of HS reconstruction algorithms that learn on the basis of images or image patches [31, 62, 70, 71, 176, 283]. [6] changed this releasing a set of 201 high resolution images that we show is enough for the successful training of deep neural networks.

3.2 Adversarial spectral image reconstruction from RGB

This section describes the core functioning of our method, along with some of the mathematical developments that derived into the proposed models.

3.2.1 Adversarial learning

Generative Adversarial Networks (GANs). GAN-s [85] are generative statistical models that learn to produce realistic samples y that lay in the data manifold by relying on a setup consisting on two competing agents: the generator G takes noise z as input as a source of randomness, and creates *fake* data samples $G(z)$. It is trained to make the generated samples as realistic as possible. On the other end, the aim of the discriminator, D, which randomly takes as input both samples from the training data set and those generated by G, is to learn to tell if the received input samples are real or fake. Typically, both G and D are neural nets, and they are trained iteratively to progressively become better in their respective tasks. The objective function associated to such a setting is:

$$\mathscr{L}_{GAN}(G,D) = \mathbb{E}_{y\sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{z\sim p_{noise}(z)}[\log(1-D(G(z)))] \tag{3.1}$$

where G tries to minimize this loss and D attempts to maximize it, yielding the objective function:

$$G^* = \arg\min_G \max_D \mathscr{L}_{GAN} \tag{3.2}$$

This adversarial framework has successfully been applied to the unsupervised generation of data of different modalities, including natural images [54], and empirical architecture guidelines for G and D have been derived [205] for such cases, along with common tricks to stabilize the training process [233].

Conditional Generative Adversarial Networks (cGANs). cGANs [167] extend this framework by feeding both G and D with additional information x to be used to condition on the output of the generator. Such conditioning input could adopt different modalities, and range from simple categorical labels [167] to more sophisticated content, such as text [208] or images [145], either alone or as a combination of multiple input modalities [209,301]. This has been proved useful for a number of tasks and output types [163,269]. Eq. (3.3) shows the updated loss function for conditional GANs. In this case, G attempts to generate images that look realistic given the additional provided input x (be it the class of y, a descriptive text, or an additional image), and D tries to determine whether the given (x, y) pair makes sense or not as a mapping.

$$\mathscr{L}_{cGAN}(G,D) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)] + \mathbb{E}_{x\sim p_{data}(x),z\sim p_{noise}(z)}[\log(1-D(x,G(x,z)))] \tag{3.3}$$

As a result, cGANs open the door to using generative statistical modeling for our HS reconstruction problem by conditioning the generation of an HS outcome on a given input RGB image.

Adversarial image to image mapping. Many modern computer vision tasks can better be regarded under the common reference framework of image to image mapping learning, in which a generator model G is learned that translates an input image x into the most probable representation y of such image in the output domain. This is the case *e.g.* for semantic segmentation [236], instance segmentation [50], or depth and surface normal estimation from single image [14], among others. Most of these tasks have been recently addressed making use of Convolutional Neural Networks that yield deterministic results as generators, and which are specifically tailored, in terms of architecture design, objective function or other specific training details, for their respective tasks.

There are, in addition, some tasks for which this mapping is not unique, and one same input image could have multiple equally correct representations in the output domain. Realistic image rendering from semantically labeled images (inverse of the semantic segmentation problem) or from hand-drawn sketches, or image colorization [42], are just a few examples of this. The choice of the objective functions to use in each of these cases is a particularly challenging design aspect; applying an otherwise useful ℓ_2 loss to x, y image pairs is known to be problematic and yield blurry results [136], as the generator tends to average over the space of valid image representations.

For all of the above, [108] proposes a common image to image mapping learning framework based on the cGAN adversarial setting, which, provided that one can feed it with co-registered image pairs of input and output domains, is able to learn the most suitable loss function for each of the tackled tasks in a data-driven approach. This is done implicitly using the adversarial objective from eq. (3.3), enforced by the discriminator trying to identify the fake images and, this way, encouraging the generator to become better at trying to deceive it.

By doing this, [108] manages to get rid of the blur inherent to ℓ_2 distance-based models and produce sharp results. Nevertheless, it has been previously shown [191, 241] that combining one of the traditional loss functions with the adversarial objective $\mathscr{L}_{cGAN}$ can help produce more spatially consistent results and make the generator less prone to artifacts inherent to the adversarial scheme. They thus place an additional ℓ_1 term (eq.(3.4)) on the generator, which is known to yield less btlur:

$$\mathscr{L}_{\ell_1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_{noise}(z)}[\|y - G(x,z)\|_1] \tag{3.4}$$

and produce the following combined objective function:

$$G^* = \arg\min_G \max_D \mathscr{L}_{cGAN}(G,D) + \lambda \mathscr{L}_{\ell_1}(G) \tag{3.5}$$

where λ is a weighting factor for the ℓ_1 term, which is set to 100 in [108]. In essence, $\mathscr{L}_{cGAN}(G,D)$ would be in charge of producing sharp, realistic looking results, while ℓ_1 takes care of the global image structure.

Interestingly, the stochastic output pursued by the noise input to cGAN-like models does not manifest itself under this design (see details in section 3.2.2), and the resulting mapping is a fundamentally deterministic one. A probable interpretation is G learning to ignore the effect of the noise. As a result, [108] gets rid of the noise input and leaves test-time dropout as unique source of randomness.

Adversarial spectral reconstruction networks. The forward correspondence learning between the RGB and hyperspectral signals is a heavily under-constrained one, which could benefit from an approach that aims at exploiting the underlying priors present in both the spectral and spatial dimensions and learn a model that specifically produces realistic outcomes as a target. It not only requires mapping a 3-dimensional image to a much higher dimensional one (typically 31 spectral channels and the two spatial dimensions), but such mapping can be context-dependent as well, as is in the case of metameric colors. The inverse mapping, however, i.e. the rendition of RGB images from their spectral counterparts, is well defined, and deterministic under the only assumption of the color matching functions defining the observer, or the spectral sensitivity functions that characterize specific sensors. This makes it immediate to generate perfectly aligned (RGB, hyperspectral) image pairs (see section 3.3) to be used under the described solution.

Hyperspectral image reconstruction from RGB can then be posed as one of the aforementioned image to image mapping learning problems and thus be solved under the conditional adversarial network-based image to image translation framework proposed by [108].

Figure 3.2: Random RGB samples from the ICVL dataset [6].

The resulting adversarial and combined objectives would then become:

$$\mathscr{L}_{adv} = \mathbb{E}_{I_{rgb},I_{hs}\sim p_{data}(I_{rgb},I_{hs})}[\log D(I_{rgb},I_{hs})] + \mathbb{E}_{I_{rgb}\sim p_{data}(I_{rgb})}[\log(1 - D(I_{rgb},G(I_{rgb})))] \tag{3.6}$$

$$\mathscr{L}_{rgb2hs}(G,D) = \mathscr{L}_{adv} + \lambda\mathscr{L}_{\ell_1} = \mathscr{L}_{adv} + \lambda\,\mathbb{E}_{I_{rgb},I_{hs}\sim p_{data}(I_{rgb},I_{hs})}[\|I_{hs} - G(I_{rgb})\|_1] \tag{3.7}$$

where I_{hs} represents the original hyperspectral image, I_{rgb} is the corresponding input image in the RGB domain and λ is scalar weight used to balance both loss terms (and is set to 100 in all our experiments, unless otherwise stated). Note that we have explicitly removed any reference to the input noise, and the RGB image remains as the only input to G.

Figure 3.1 shows an overview of the whole adversarial spatial context-aware spectral image reconstruction process. We depart from a database of perfectly aligned RGB and hyperspectral image pairs, which are extracted one pair at a time. In a first iteration, a first pair of real images of size $H \times W$ is taken: $\{I_{RGB}, I_{HS}\}$. The generator G takes I_{RGB} as input, and yields the corresponding hyperspectral reconstruction of size $H \times W$, $\hat{I}_{HS}$. The discriminator D is now fed with two pairs of images, $\{I_{RGB}, I_{HS}\}$ and $\{I_{RGB}, \hat{I}_{HS}\}$ and uses the associated labels indicating if they are real or fake $\{1,0\}$ to compute the adversarial loss and update its gradients. G's weights are also updated, and both D and G continue to become better at their respective tasks iteratively.

3.2.2 Architecture design and training

As for the specific implementation of the models, since G needs to yield full-size detailed images, a *U-Net*-like architecture [223] is used. Regular autoencoder networks [124] exhibit a progressively reduced representation size until a bottleneck layer and there is no way for the last layers of accessing the original data, which negatively affects the results when we aim at detailed outcomes. Unlike these, the *U-Net* incorporates skip connections between layers of equal representation size, and concatenates local activations from the upscaling phase with those coming from the downscaling stages, which has shown to achieve superior performance on tasks were the details are relevant. It was first proposed with semantic segmentation tasks in mind, but

original spectral signal reconstruction falls within the kind of tasks that can clearly benefit from accessing the original input levels at each sample (i.e. pixel).

The discriminator D, defined as *PatchGAN*, is simpler in terms of convolutional layer count, and is focused solely on modeling high-frequency structure. Each of the $M \times M$ output neurons is restricted to see only a limited $N \times N$ receptive field from the input image, which can be significantly smaller than the input image size. Consequently, only the adversarial loss term is placed over D (eq (3.5)).

The use of this design solution for D is consistent with our initial hypothesis that local spatial context can help better reconstruct the spectral signal. Specifically, we hypothesize that the proposed approach could help disentangling the illuminant and object body reflectance components of a pixel's trichromatic response, as defined by the dichromatic reflection model [235]. The design of D, with its attached ℓ_1 objective, helps capture the high frequencies that characterize the textures in the image. These are, together with the body color component, one of the main features characteristic of the different materials which, ultimately, produce distinct spectral responses. Therefore, convolutionally integrating the trichromatic response of adjacent pixels should yield a better estimate of the central spectral response. To this respect, the *PatchGAN* design isolates D's response associated to pixels separated by more than one input patch. For small enough patch sizes, this effectively implies that the discriminator is learning a loss function tailored for texture or material recognition, making sure that the reconstructed spectra falling within the patch are not only plausible in the spectral domain, but also spatially consistent in the close proximities.

The illuminant-specific component of [235], on the other hand, is typically largely constant or slowly varying across big portions of the image (especially in terms of chromaticity and conversely, spectral shape), and the ℓ_1 norm does a good job taking care of its global image-wide consistency, along with that of the low-mid frequency spatial structures.

Avoiding Batch Normalization. Given the intrinsically exact nature of our task (some of the described design choices help leverage spatial structure consistency for our task, but we ultimately want the reconstructed spectra to be accurate), we choose to remove all the Batch Normalization [106] layers present in the generator architectures proposed in [108]. While this technique has shown to be useful to help accelerate and regularize the training process for a wide variety of tasks by reducing the internal covariate shift, the fact that it makes the signal lose track of its original value, along with the deterministic nature of the desired output, makes it non-advisable for reconstruction tasks. We experimentally found that including Batch Normalization produced inferior results.

3.2.3 Implementation details

We now provide some details on the configurations used for our implementation. We use Keras with Theano backend and take the implementation of [108] made by [47] as starting point, modifying it for our purposes. We use Adam optimizer [123] for both G and D, with a learning rate of $2 \cdot 10^{-4}$ and $\beta_1 = 0.5$. We use a minibatch size of 1 in order to benefit from the regularization provided by the gradient estimation noise [119], and following common practice [108]. The training is performed iteratively and alternates between the two models: at each step, the discriminator is first trained for 50 iterations and then the generator gets trained for 25 more minibatches.

We crop the original 1392×1300 images during the training phase by extracting one random crop of size 256×256 (the H, W values from section 3.2) per image and epoch. The models are fed with these crops during training, while, for the testing phase, each full size RGB image is divided in tiles of 256×256 with no overlap, which effectively yields image sizes of 1280×1280 pixels. Each tile gets processed by the generator independently and we reconstruct the full image back before evaluating it.

The generator G accepts input images of size 256×256. Its encoding stage is composed by eight successive

3×3 convolutions with stride 2 and a leaky ReLU after each of them, thus yielding a 1×1 activation in the most narrow point of the main branch. The initial number of filters is 64, which gets doubled at each convolutional layer up to 512, keeping it constant after that. On the decoding part, eight transposed convolution blocks successively double the activation size up until the original 256×256 size, while progressively reducing the number of filters in a symmetric way with respect to the encoding stage. Each block comprises the transposed convolution itself, followed by a train-time-only Dropout layer (with a drop rate of 10%) and a leaky ReLU activation. After each Dropout, the correspondent activations from the encoding stage are concatenated, thus producing eight skip connections between levels of equivalent activation size. Finally, two 1×1 convolutions are added at the end before the output *tanh* activation, with a leaky ReLU in between, in order to get the direct input images adequately combined with the upstream features.

The discriminator D is a simple single-branch net composed of four 3×3 convolutional layers with stride 2, each of them followed by a leaky ReLU, with filter numbers doubling at each step. A fifth 3×3 convolution with a sigmoid yields the output 8×8 prediction.

3.3 Experimental evaluation

This section contains an overview of the experiments performed to quantitatively assess our algorithm's performance as compared to previous methods.

3.3.1 Dataset

Given the amount of images, diversity and resolution, we evaluate our approach on the dataset presented in [6]. At the time of writing, it comprised 201 hyperspectral images (see Figure 3.2 for RGB renditions of a few random samples) of 1392×1300 spatial resolution and 519 spectral bands in the $400nm - 1000nm$ range, with a spectral resolution of 1.25nm. As for the acquisition, a *Specim PS Kappa DX4* hyperspectral camera was used, together with a rotary stage for spatial scanning. This aspect is noticeable in some of the samples, in which common objects such as cars exhibit aspect ratios that do not match those we find in real life. There is also a spectrally downsampled version of 31 bands in the $400nm - 700nm$ range. Following practice from [6], we use the latter for our reconstruction experiments. There is no illuminant information available for each of the images, which would allow for object reflectance recovery; therefore, our task consists on the estimation of the radiance correlate represented by the captured hyperspectral images.

3.3.2 Preparation

In order to get the aligned image pairs dataset required by our method, and given the deterministic correspondence between spectral and RGB samples once the observer (or sensor sensitivity functions) and the output color space are specified, we render wide band trichromatic RGB versions of the spectral images in the sRGB color space as follows: we first obtain the CIE XYZ tristimulus values for each spectral image pixel location x, making use of the color matching functions corresponding to the CIE 1964 10° standard observer:

$$\mathbf{X}(x) = K(x) \sum_{\lambda=400nm}^{700nm} S(\lambda, x)\bar{\mathbf{x}}(\lambda)\Delta\lambda \tag{3.8}$$

where $S(\lambda, x)$ is the relative spectral power distribution of pixel x, $\mathbf{X} = \{X, Y, Z\}$, $\bar{\mathbf{x}}(\lambda) = \{\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)\}$ are the color matching functions, $\Delta\lambda = 10nm$ and $K(x)$ is the normalization factor, defined, for illuminant $L(\lambda, x)$,

Method	RMSE	RMSERel	GFC	ΔE_{00}
Galliani *et al*. [72] (reported)	1.980	0.0587	–	–
Arad *et al*. [6] (reported)	2.633	0.0756	-	-
Arad *et al*. [6] (optimized parameters)	2.184 ± 0.064	0.0872 ± 0.004	–	–
Ours (weighted avg.)	$\mathbf{1.457 \pm 0.040}$	$\mathbf{0.0401 \pm 0.0024}$	$\mathbf{0.99921 \pm 0.00012}$	$\mathbf{2.044 \pm 0.341}$
Ours (fold 0)	1.452 ± 0.101	0.0383 ± 0.0024	0.99906 ± 0.00001	1.861 ± 0.324
Ours (fold 1)	1.463 ± 0.022	0.0420 ± 0.0024	0.99936 ± 0.00023	2.228 ± 0.358

Table 3.1: Summary results of the conducted experiments over ICVL dataset. Black pixels contained in the original hyperspectral images (derived from the variable image width) are not taken into account for evaluation purposes in any of the experiments, and folds are weighted accordingly. RMSE values are in the $[0-255]$ range. Two train-test cycles were run and the results averaged.

as:

$$K(x) = \frac{100}{\sum_{\lambda=400nm}^{700nm} L(\lambda, x)\bar{y}(\lambda)\Delta\lambda} \tag{3.9}$$

Note that, before going through this computation, the original spectral power distribution captured by the camera for each image $S'(\lambda, x)$ is preprocessed with min value subtraction and max value scaling. The final step is producing the sRGB renders. We do so by applying the associated 3×3 transformation matrix and unlinearizing (i.e. *gamma-correcting*) the result with a $1/2.4$ power law gamma with a linear segment in low luminance values.

While not suffering from the same lack of an adequate performance evaluation method that affects typical generative modeling tasks [251], spectral signal reconstruction algorithms assessment is an active research field that lacks consensus on what is the most adequate metric to measure spectral match of signals [105]. When the signals comprise the visual spectrum, the task can be tackled from a variety of perspectives, ranging from the pure signal processing point of view of spectral curve difference metrics, to a full spectrum of metric families that place different levels of perceptual load on their computation: metameric indexes, CIE ΔE color difference equations, or weighted spectral metrics.

If we widen the scope onto full reference image difference metrics, little work has been done on the spectral extension of these families [138]. We here focus on four of the most widely used metrics, namely RMSE (Root Mean Squared Error, computed across the spectral dimension for each pixel and then averaging for whatever number of pixels present in the image or the dataset), RMSERel (i.e. RMSE relative to the value of the real signal), GFC (Goodness of Fit Coefficient [222]) and ΔE_{00} (CIEDE2000) perceptual color difference formula [43] computed over the reconstructed tristimulus values.

3.3.3 Experiments and discussion

We first compare our method with [6]. In their general experiment over the whole set, which was back then composed of 100 images, they perform a leave-one-out procedure, and learn from pixels sampled along the whole set except for the unique image being tested at a time. Their reported results for this setting are shown in Table 3.1 as *Arad* et al. *(reported)*, together with those reported in [72] on their own evaluation setting. We choose to instead split the full dataset in two equal partitions of 100 images each [†], training on

[†]Train-test splits available at https://aitorshuffle.github.io/publication/2017-10-10-alvarez-gila_adversarial_2017

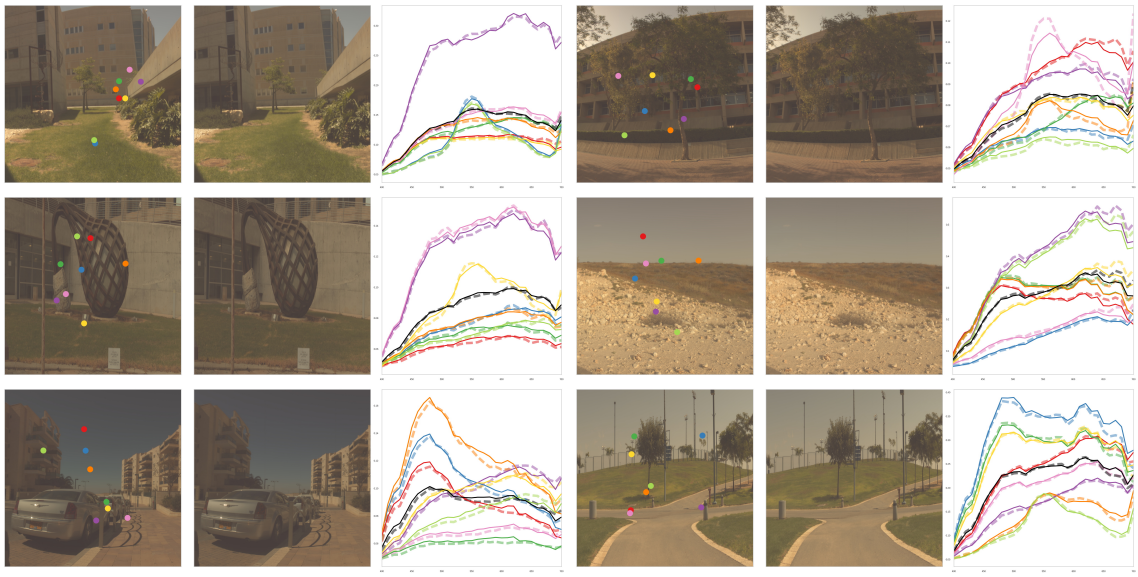


Figure 3.3: Sample results for our method. For each triplet, left, center: sRGB rendition of original and reconstructed hyperspectral signals, respectively. Right: Original (dashed) and reconstructed (solid) spectra of eight random pixels identified by the colored dots.

one and reporting on the other, running two full train-test cycles and averaging the results across folds and runs. Table 3.1 compares the obtained values for the aforementioned metrics over each of the testing sets, showing an average per-pixel error drop of 33.2% in terms of RMSE and 54.0% in terms of RMSERel with respect to [6] evaluated over the same splits and with their hyperparameters (i.e. dictionary size, number of samples per image, iterations and sparsity target) optimized over the test sets. While [6] does not provide any further evaluation metric, note that our average GFC values are all above the GFC threshold which [222] considers a *very good reconstruction*, and one which implies missing only 0.2% of the signal energy in the process. Also, the average per-pixel color difference (which does not account for spatial perceptual effects) is constrained around as low as $2\Delta E_{00}$ units.

Figure 3.3 shows the sRGB rendition of original and reconstructed hyperspectral images for some randomly chosen test image samples. In addition, for each image, we show the original and estimated spectra for eight randomly selected pixels from the image.

Does the spatial information actually help?

In an attempt to empirically validate our main hypothesis of contextual spatial information on a local neighborhood being relevant for the correct spectral reconstruction of any given central pixel, we conduct a branch pruning experiment. We depart from a minimal version of our net, in which both the main branch and all the skip connections have been removed, except for the one connecting the 256×256 input with the last pair of 1×1 convolutions (such model predicts each output pixel independently by design, without incorporating any spatial contribution), and keep adding skip connections at successively deeper levels (extending the receptive field of the model and thus increasing the spatial contribution at each step) until we end up with the

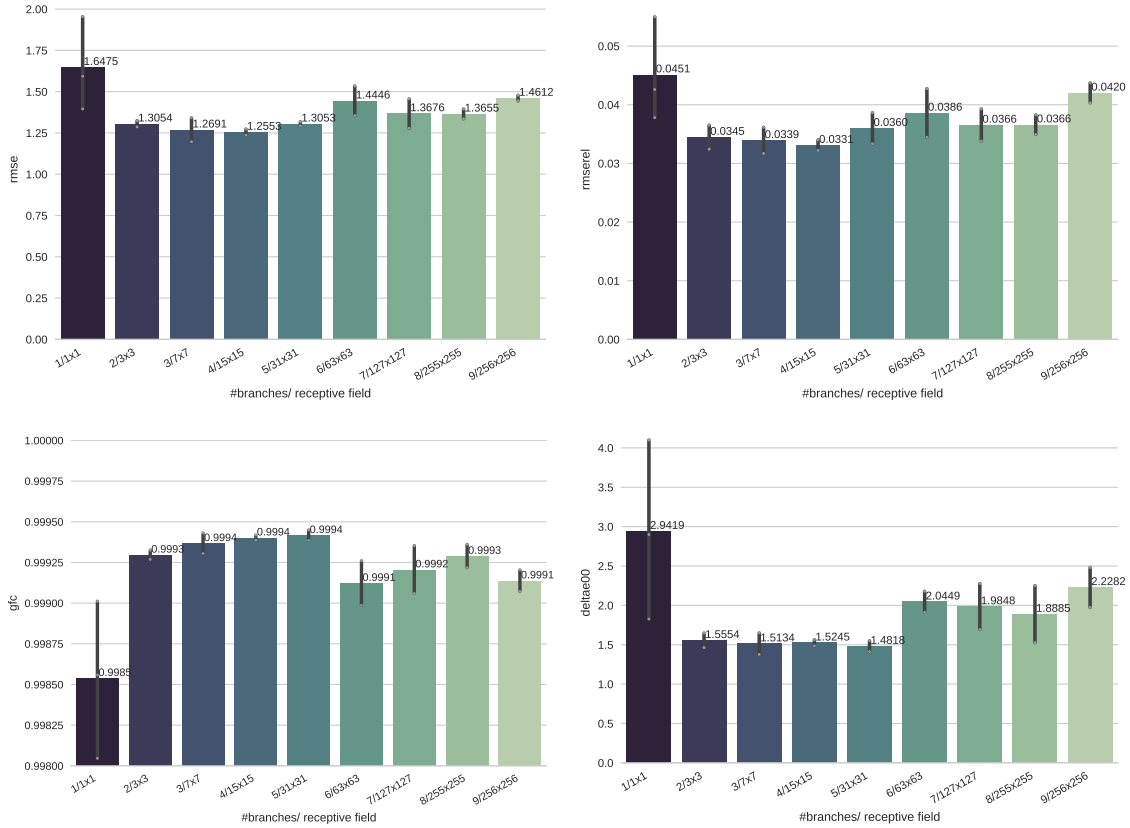


Figure 3.4: Branch pruning experiment results. Top-left: RMSE. Top-right: RMSERel. Bottom-left: GFC. Bottom-right: ΔE_{00}. Leftmost bar is the model with a single skip connection at 256×256 activation size level and 1×1 receptive field (RF). Each additional bar adds one skip connection at increasingly deeper levels of the U-Net. The rightmost bar is the full net, resulting from the addition of the main branch, and its RF (which would be 512×512 in an unconstrained scenario), is here limited by the 256×256 patch size. The addition of this last layer is justified by the notion of effective RF presented in [158], which may be significantly smaller than its theoretical counterpart.

full net, after the addition of the main 1×1 stream branch. Figure 3.4 shows the results of running at least two train-test cycles on each of these nets, and testing over the 1280×1280 versions of the images in fold 1. All the four metrics show a closely correlated outcome, with a very significant average performance improvement (-20.8% RMSE, -23.5% RMSERel, $-47.1\%\Delta E_{00}$) when transitioning from the model with a single skip connection and a 1×1 receptive field to that with 2 skip connections and a 3×3 receptive field. Further increases of the model's theoretical receptive field (by adding new branches) yield only marginally better results (models labeled as $3/7 \times 7$, $4/15 \times 15$) and, from there on, additional deeper skip connections produce increasing test error rates. We hypothesize that this is due to the influence of overfitting for experiments $5/31 \times 31$ and onwards. Given this, a straightforward way of improving the reported results could be that of

increasing the regularization associated to the deepest branches by, e.g., increasing their dropout rate.

It is also noticeable the substantially higher variance that the results seem to suggest for the model with a 1×1 receptive field. This would mean that the addition of local spatial information does not only improve the overall prediction accuracy, but it does so in a more robust manner as well.

3.4 Conclusion

We propose a convolutional neural network architecture that successfully learns an end-to-end mapping between pairs of input RGB images and their hyperspectral counterparts. We adopt an adversarial framework-based generative model that shows itself effective in capturing the structure of the data manifold, and takes into account the spatial contextual information present in RGB images for the spectral reconstruction process. State of the art results in the ICVL dataset suggest that individual pixel-based approaches suffer from the fundamental limitation of not being able to effectively exploit the local context when applied to spectral image data in their attempt to build informative priors. The observed performance in terms of both reconstruction error and speed open the door to a full range of potential higher level applications in sectors of increasing demand for spectral footage at a lower cost.

4 A Probabilistic Model and Capturing Device for Remote Simultaneous Estimation of Spectral Emissivity and Temperature of Hot Emissive Materials*

4.1 Introduction

The steelmaking process consists of obtaining steel products through a series of industrial steps of ferrous raw materials (scrap or iron ore) melting, liquid steel solidification and thermo-mechanical transformation. Those processes must be properly adjusted based on the features of the raw materials and the required properties of the steel products to be obtained. Currently, the two main routes for producing steel are the integrated route and electrical route [10]. However, no matter the process route used, obtaining inline detailed information, not only for steel, but also for slag in the different reactors is of vital importance for the optimization of the processes. In the digital era, in which industrial facilities tend to be automatized through process smartization, the development of online analysis tools is becoming an important topic of concern and heavy processes like EAF (Electric Arc Furnace) represent a giant challenge. In this sense, several solutions have been reported for liquid steel control in the EAF, including continuous temperature measurement by using optical fiber or optical sensors [134], dynamic metallurgical models that continuously solve mass and thermal balances for the whole system [20] or foamy slag detection by noise and arc harmonics analysis [264].

As for the EAF operation, the process begins with a mixture of different ferrous raw materials being introduced in the reactor. These materials are melted down in the initial phases of the procedure and then, during the steel refining phase, the liquid steel is heated up till about 1650°C. For succeeding in the steel transformation process, most of the undesired elements contained in the scrap are oxidized and moved to the slag, whose function is also to improve the efficiency of the huge amounts of energy (electrical and chemical) continuously being introduced during the EAF operation. At the end of the process, the slag is removed from the furnace and the steel is poured into a refractory ladle to be refined in the ladle furnace process.

The analysis of the slag temperature and chemical composition is of vital importance during the steel manufacturing process. The slag composition is measured manually by the use of a manually acquired and prepared sample on a spectrometer that allows calculating the steel-slag thermodynamic state. Although some automated methods have been implemented to avoid the slag sample preparation and to directly measure the slag composition based on its spectral reflectance [197], they still require the manual extraction of the slag sample, still representing a slow and manual process.

Novel methods are therefore required to measure the temperature and composition of the liquid slag without interfering with the steel manufacturing process. Currently, there exists no viable solution for accurately measuring the temperature of liquid slag, and few technologies have been posed for online understanding of slag evolution (i.e. solid/liquid fraction distribution, chemical composition or temperature) [192]. Proposing a method for remote extraction of continuous slag characteristics would allow for the optimization the EAF process by continuously adjusting the process parameters according to the evolution of the slag.

In this work, we present a novel portable spectrometer device and Bayesian probabilistic algorithm that are capable of direct remote estimation of temperature and spectral emissivity from remote radiant samples. The system captures the radiance signal incoming from an 8 cm diameter spot located up to 20 m away. A fully Bayesian model integrates all the signal pipeline, simultaneously estimating the sample temperature,

*This chapter is based on a publication in the IEEE Access journal, 2021 [198]. First authorship is shared with Artzai Picon with equal contribution. It also led to the associated European patent application [196].

spectral emissivities, the absorption caused by the presence of water vapour and CO_2 along the optical path that explains the observed radiance with a maximum likelihood. The proposed method was validated with alumina ($\alpha - Al_2O_3$) and hexagonal boron nitride ($h - BN$) samples and compared with standard laboratory analysis obtaining good correlation. The proposed system and methods can be used in steel factory settings for in-situ electric arc furnace monitoring without the need for active thermocouple or calibration blackbody.

4.2 Related work

On the EAF process, temperature is a key factor that regulates thermo-chemical processes in order to yield appropriate properties of the resulting steel [20, 264]. On the other hand, slag composition determines the thermodynamic system status between the steel and slag [197]. In this sense, coming to the idea of developing methods for characterizing materials involved in the steel manufacturing process, a compact and portable device that remotely provides reliable composition and temperature values under actual industrial conditions is desirable.

The thermal infrarred range has been widely analysed by different authors in order to establish relationships between the spectral emissivity and the different materials. Different works [22, 143, 149, 169, 182, 260] analyse the relationship between the temperature and the emissivity of different materials, evidencing the dependency of the spectral emissivity on the chemical composition for slag materials such as SiO_2, Al_2O_3, FeO, Fe_2O_3, CaO and MgO.

Temperature measurement devices are commonly divided in two different categories: those that require contact (thermometers, thermocouples, thermistors, etc.) and those that do not (e.g. pyrometers, thermographic cameras). The latter, which sense and measure the incoming infrared radiation in order to estimate the temperature, are frequently used for monitoring high-temperature furnaces. To make an accurate temperature measurement, these methods require precise knowledge on how the emissivity behaves in the spectral range that the radiation is detected, as $T = T(\varepsilon(\lambda))$. The emissivity $\varepsilon(\lambda)$ is a parameter that indicates how a material emits radiation when compared to a black body, the perfect emitter.

A number of works have proposed novel methodologies to measure emissivity and temperature in field with the use of non-contact and portable devices: Rego-Barcena et al. [212, 213] described a technique to obtain emissivities in situ at different locations, with a portable, rugged and inexpensive device. Their method is capable of estimating a single spectral emissivity value and the average temperature by means of least-squared optimization. However, this method considers the boiler emitter as a gray body. This is not the case on the EAF furnace, as the emissivity depends on the composition of the slag-steel fraction [20, 22, 216]. As opposed to theirs, our method makes no assumption on the shape of the underlying spectral emissivity of the sample.

Other methods focus on estimating spectral emissivities under controlled conditions, where the temperature is not estimated but used as a model input [143]. More recently, different Bayesian methods have been employed to model the complex interactions between the radiative physical processes: [101] used Montecarlo techniques to simulate the radiative heat transfer between surfaces, while a few works [56, 63, 96] apply Bayesian approaches to address the composition estimation problem.

Temperature Emissivity Separation (TES) methods aim to simultaneously estimate sample temperature and emissivity. Given the inherently under-constrained nature of the problem, most of them tackle the task by imposing certain strong priors on the proposed models [19, 52, 109, 170–172, 225]. A few others claim to avoid making such heavy assumptions by instead imposing a maximization of the entropy [15, 16, 150], although still require to explictly select the temperature and emissivity ranges within which the solution is expected to be. Besides, in practice, these solutions are often tied in some way to the specific use case for which they were conceived. Many of the aforementioned approaches were indeed proposed and tested in

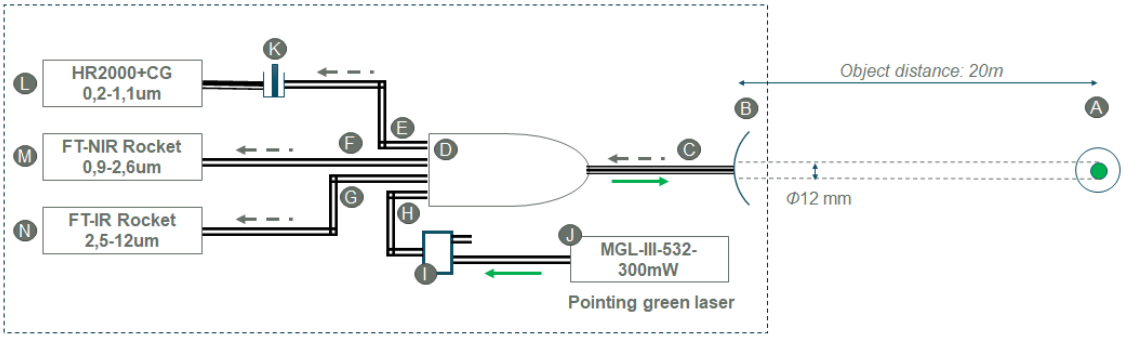


Figure 4.1: Design diagram of the acquisition system. Targeted point is illuminated by the green laser (J). A collimator (B) captures the signal, which is transmitted to the three spectrometers (L), (M) and (N) through the fiber bundle (D). Filter (K) eliminates the signal from the green laser (J).

the context of remote sensing applications, and thus operate under a very constrained range of plausible temperature values [16, 225].

Some of these approaches have also been applied in steel manufacturing processes. Sonoda et al. [244] e.g. propose a bootstrap filter-based method to estimate the probability distribution of liquid steel temperature. Meanwhile, only machine learning methods have been proposed for the estimation of the slag parameters [272].

In this work we introduce a compact and portable device that remotely produces reliable temperature and visible-IR spectral emissivity estimates under actual industrial conditions in quasi real time. The device comprises three punctual spectrometers with partially overlapping spectral ranges, covering a combined range of [0.2, 12.0μm] for an improved predictive performance. Along with it, we present and validate a fully Bayesian radiative transfer model that seamlessly estimates and accounts for the spectral emissivity, the sample temperature and the different unknown variables that may affect the received radiance on real industrial situations such as presence of gases on the optical path or fiber misalignments on the spectrometers. It does so by leveraging the signal received by the three available spectrometers, and could easily generalize to an arbitrary number of them. As opposed to machine learning-based approaches, the model can be considered unsupervised once the calibration procedure has been completed, as it requires no further training data. In addition, it yields full density estimates of the considered random variables, thus considering the uncertainty associated to each prediction. This is possible due to its Bayesian nature, which also implies the need to impose a prior distributions over each of the target random variables. The distributions proposed in section 4.4.2 for such priors provide a good parametrization for the very wide range of temperatures considered in this work. However, these are soft constraints that may be further softened by substituting them by other less informative distributions (or hardened, should the user wish to apply it in more constrained setups); the described framework would still be valid in both cases, and the influence of the prior distributions is in any case overridden as the evidence provided by the captured observations becomes more prevalent.

4.3 Design of the device

As mentioned in section 4.2, several systems have been proposed to simultaneously calculate spectral emissivity and temperature of hot emissive samples. However, these systems have been designed to work

under strict controlled laboratory conditions and are not capable of estimating the temperature and spectral emissivity under real industrial conditions. The targeted goal for our device is to allow operating on the proximity of the electric arc furnace from the steel factory. These conditions imply high temperature, dirtiness, tolerance to mechanical impacts and vibrations and uncontrolled vapour conditions.

Based on preliminary laboratory testing and analysis, the following system requirements were defined and set as design guidelines. (i) As for the optical requirements, a spectral range of $200-12\,000$ nm was defined with a minimum spectral resolution of 50 nm/pixel on the far infrared region. (ii) The device should allow for the remote acquisition of a spot with a diameter of 12 mm at a distance of 20 m. (iii) The system should be able to perform simultaneous spatial and temporal acquisition, (iv) with a minimum acquisition rate of 0.5 samples per second. (v) Regarding usability, the device should provide in-field pointing capabilities and on-field calibration mechanism without the need for sending the system to laboratory for calibration. (vi) Finally, as for the working conditions, the system should be tolerant to heat and vibrations and (vii) agnostic to the presence of water vapour, CO and CO_2 on the optical path. The present work describes how a prototype fulfilling these requirements has been designed, built, calibrated ant tested both in laboratory and in real industrial conditions at the ArcelorMittal research casting factory of Sestao.

4.3.1 System description

Fig. 4.1 shows the acquisition device diagram. In order to capture the radiance from a 12 mm diameter area from region *(A)* located up to a distance of 20 m we employ a UV-Enhanced Aluminum Reflective collimator *(B)* with an SMA connector that assures a 12 mm beam and a good reflectivity throughout the required range. The captured radiance passes through a special fiber *(C)* comprising a fiber bundle that ends in 4 SMA-905 outputs *(D)*. This fiber bundle is composed by one input and four outputs, allowing spectral coverage in the $200-12\,000$ nm range. This fiber bundle is composed by seven fibers: 1×200 µm UV-VIS fiber, 3×240 µm Polycristaline infrared (PIR) fibers and 4x VIS-NIR 200 µm fibers. Two VIS-NIR fibers *(H)* are connected to an inline splice bushing connector (Thorlabs 20-02) *(I)* which is also connected to a green laser source *(J)* (MGL-III-532-300 mW) that serves as system pointer. The UV-VIs fiber *(E)* connects into a UV-VIS spectrometer *(L)*, an Ocean Optics HR2000 with a composite grating covering $200-1100$ nm. A notch filter *(K)* (Thorlabs NF533-17) is placed to remove green laser signal from the spectrometer entrance. The other two VIS-NIR fibers *(F)* are connected to a Fourier transform infrared spectrometer *(M)* composed by a CaF beamsplitter, with an $InGaAs$ detector covering the $0.9-2.6$ µm range (ArcOptics). The two PIR fibers are connected to a second Fourier Transform Infrared spectrometer *(N)* with a $ZnSe$ beamsplitter covering the $2.5-12$ µm range (ArcOptics).

The system is packed into an acquisition case (see Fig. 4.2) that allows taking measurements on industrial conditions. The case provides mechanical protection to the optical components as well as data connectivity based on a single USB output connection for all the different acquisition devices.

In order to obtain a continuous spectral signature in each capture, a dedicated software has been developed. The software performs real time acquisition of the data from the three spectrometers and it applies the latest optical system's calibration function on the fly. It allows plotting, recording and saving the captured information, and provides access to the specific low-level settings of each independent spectrometer.

4.3.2 System calibration

The calibration of the system have been divided into two different processes. A first process models the spectrometers' non linear response and generates the optical system transfer function. However, the intended working conditions of the system may significantly affect the system calibration. Fiber re-installation, vibrations and in-place temperature variations affect the amount of radiance acquired by each spectrometer.

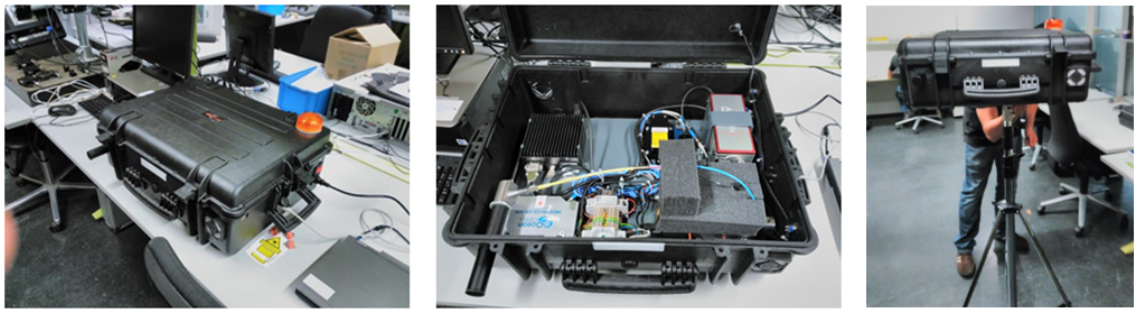


Figure 4.2: Packed prototype. (left) Acquisition case, (middle) Packed opto-mechanical system, (right) Tripod mounting.

This makes necessary an additional in-situ calibration of the system. To cope with this, a second calibration stage based on a stabilized calibration lamp is added to the system operation.

Calibration of the non-linearities of the spectrometers

Each spectrometer wavelength was calibrated by fitting a 2^{nd} degree polynomial function that maps the collected counts by the spectrometer at each λ_i into the theoretical blackbody radiance at the blackbody temperature (T_j). The polynomial coefficients W_{λ_i} are estimated at each wavelength λ_i by minimizing the root mean square error between the theoretical blackbody emissivity at T_j: $L_{bb_{\lambda_i}}(T_j)$ and the mapping of the counts $C_{\lambda_i}(T_j)$ acquired by the spectrometer when looking into a calibrated blackbody at temperature T_i at λ_i as described in (4.1):

$$\operatorname{argmin} W_{\lambda_i} \sqrt{\sum_{T_j=T_0}^{T_n} [L_{bb_{\lambda_i}}(T_j) - f_{calib}(C_{\lambda_i}(T_j), W_{\lambda_i})]^2} \tag{4.1}$$

where f_{calib} is a 2^{nd} degree polynomial function that takes W_{λ_i} as its polynomial coefficients to map spectrometer acquired counts into theoretical radiance.

To calibrate the system, acquisitions over a blackbody furnace between 500°C and 1500°C were performed at the closest possible distance of 600 mm from the tip of our system collimator and the end of the blackbody furnace. Fig. 4.3 depicts the calibration setup. It is noteworthy to remark that this optical path contained gases composition that corresponds to the part of the atmospheric transmittance corresponding to the optical path that is inside of the calibrated device. However, this effect can be considered negligible in comparison to the full optical path.

The root mean squared error resulting from this process was 2% for spectrometer 1 (200−1100 nm), 1% for spectrometer 2 (970−2600 nm) and 0.5% for spectrometer 3 (2500−12 000 nm), as illustrated in Table 4.1. Correlation among theoretical radiance and calibration-corrected captured samples for three specific wavelengths and six temperatures is depicted in Fig. 4.4.

Field calibration

Working in industrial conditions makes the system be affected by mechanical vibrations, dust, temperature changes, etc. Even the movement between the different fibers occasionally led to observed intensity changes of up to 15% on the collected signal. Fortunately, these changes are related to the coefficient of transmission

Figure 4.3: Acquisition system under calibration set-up. (left) System on the calibration room close to the blackbody furnace, (right) Close-up.

Calibration distance (mm)	Test distance (mm)	Spectrometer	RMSE (%)
1000	1700	1	2%
1000	1700	2	1%
1000	1700	3	0.5%

Table 4.1: Acquisition system calibration error (%).

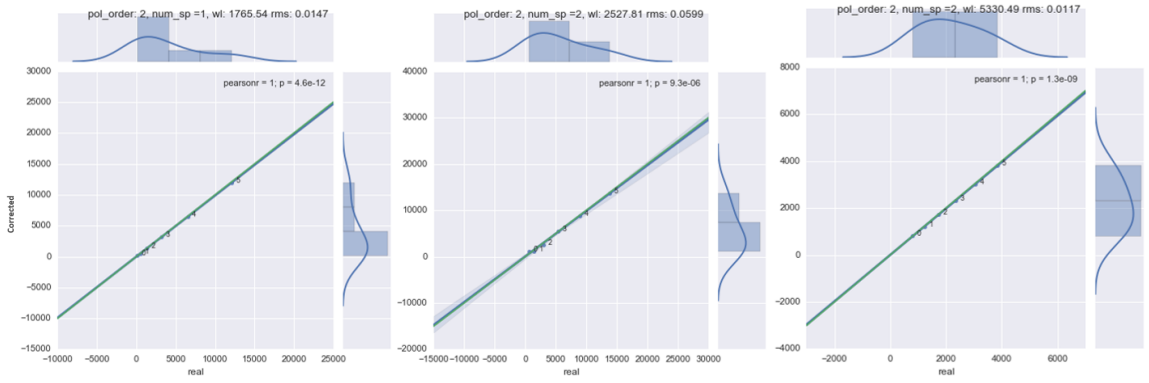


Figure 4.4: Correlation plot of the calibrated system response vs. an ideal blackbody source for three separate wavelengths (1765.54 nm, 2527.81 nm, 5330.49 nm) at six different furnace temperatures. x axis: theoretical blackbody radiance at the specific wavelength. y axis: radiance estimated by the calibration polynomial. Right and top subplots depict respectively the estimated and real distribution of the temperatures.

between the different fibers and devices. We model this transmission coefficient as independent from wavelength. Based on this hypothesis, we assume that the change in the amount of radiance acquired by each

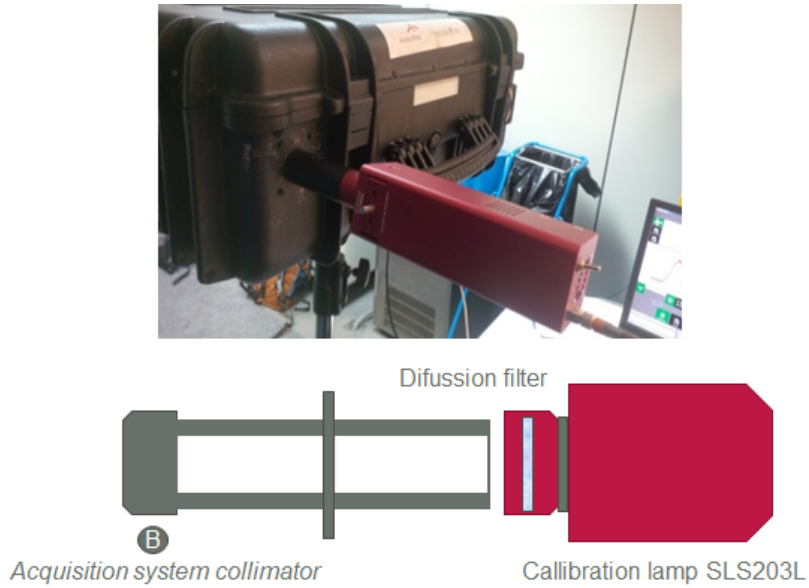


Figure 4.5: Calibration lamp for field calibration. (top) System mounted for calibration, (bottom) Lamp calibration diagram: calibration lamp, diffusion filter and SLS203L camera.

spectrometer S can be modelled by a proportional factor C_s that does not depend on the wavelength.

In order to reduce this effect, a stabilized spectral calibration lamp emulating a blackbody at 1500 K (Thorlabs SLS203L (500−9000 nm)) is proposed for daily calibration at the beginning and at the end of each acquisition campaign. The calibration lamp is used with an installed diffusion filter to allow for diffuse illumination into the collimator. Fig. 4.5 depicts the imaging system with the coupled calibration camera.

The field calibration procedure is defined as follows: (i) During the spectrometers' non-linearity calibration phase (section 4.3.2) the calibration lamp is mounted into the system and switched on, and the radiance captured by the different spectrometers $L_{s0}(\lambda)$ is stored as reference. (ii) Later, during the field measurements, the calibration lamp is set again and the new signal $L_{s1}(\lambda)$ is recorded. (iii) For each spectrometer, the ratio of the average intensity between the two lamp signals along a specific spectral range $[\lambda_l, \lambda_h]$ is used for signal correction, as shown in (4.2). An spectral range of 450−900 nm is used for spectrometer 1, 960−2500 nm for spectrometer 2 and 3500−4000 nm for spectrometer 3. Figure 4.6 shows the resulting three obtained correction factors K_s, one per spectrometer. The RMSE before correction was 11% for the signal induced by the calibration lamp, whereas this error was reduced to 0.4% after applying the correction factors.

$$K_s = \frac{\sum_{\lambda=\lambda_l}^{\lambda_h} L_{s1}(\lambda)}{\sum_{\lambda=\lambda_l}^{\lambda_h} L_{s0}(\lambda)}, \ \forall s \in \{1,2,3\} \tag{4.2}$$

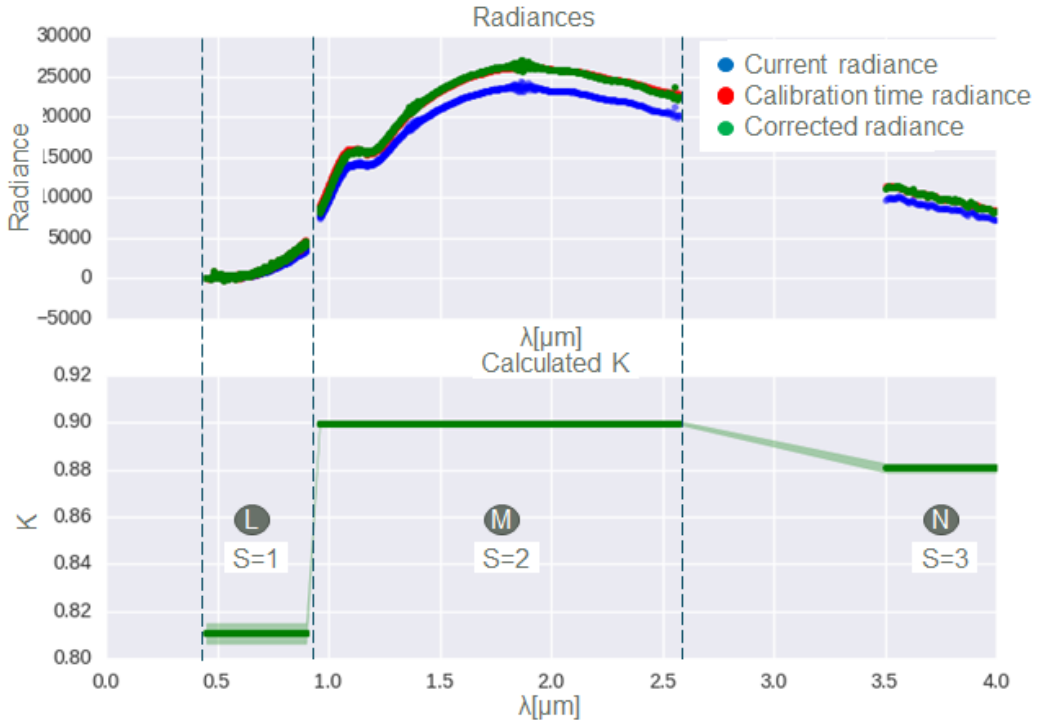


Figure 4.6: Calibration correction by using the calibration lamp. (top) In blue, the current radiance received by the sensor, in red the reference radiance obtained at calibration time and in green the corrected radiance after applying the correction factors K_s. (bottom) The three K_s correction factors calculated, following (4.2), as the ratio of current collected collected lamp-induced radiance $L_{s1}(\lambda)$ and the radiance collected with the lamp right after the nonlinear response calibration $L_{s0}(\lambda)$.

4.4 Radiative Transfer Model

4.4.1 Model formulation

Our final aim is the simultaneous estimation of the temperature T_{bb} and spectral emissivity $\varepsilon(\lambda, T_{bb})$ of the observed hot sample, taking as sole input the number of counts (i.e. digital level magnitude) yielded by the capture software as a function of the wavelength, $C(\lambda)$. We model the scenario described in the preceding sections as a perfect blackbody radiator $L_{bb}(\lambda, T_{bb})$ emitting at our unknown target temperature T_{bb}, whose emission is successively filtered by (i) a selective filter with a transfer function that equals the spectral emissivity that characterizes the sample (both conforming a selective radiator), (ii) the atmospheric spectral transmittance of the optical path between the device and the sample, $\mathcal{T}_{atm}(\lambda)$, and (iii) the transfer function of the full multi-spectrometer optical system, $\mathcal{T}_{OS}(\lambda)$ (comprising the fibers, optical components, detector and the slit function modelling its spectral convolution), which is determined by the calibration process (see section 4.3), and relates the observed physical magnitude -radiance, in $W/m^2\mu m$- to the counts yielded by each of the spectrometers. Equation (4.3) formalizes this model, which is graphically described in Fig.4.7 [275].

$$C(\lambda) = \mathcal{T}_{OS}\left[L_{bb}(\lambda, T_{bb}) \cdot \varepsilon(\lambda, T_{bb}) \cdot \mathcal{T}_{atm}(\lambda)\right] \tag{4.3}$$

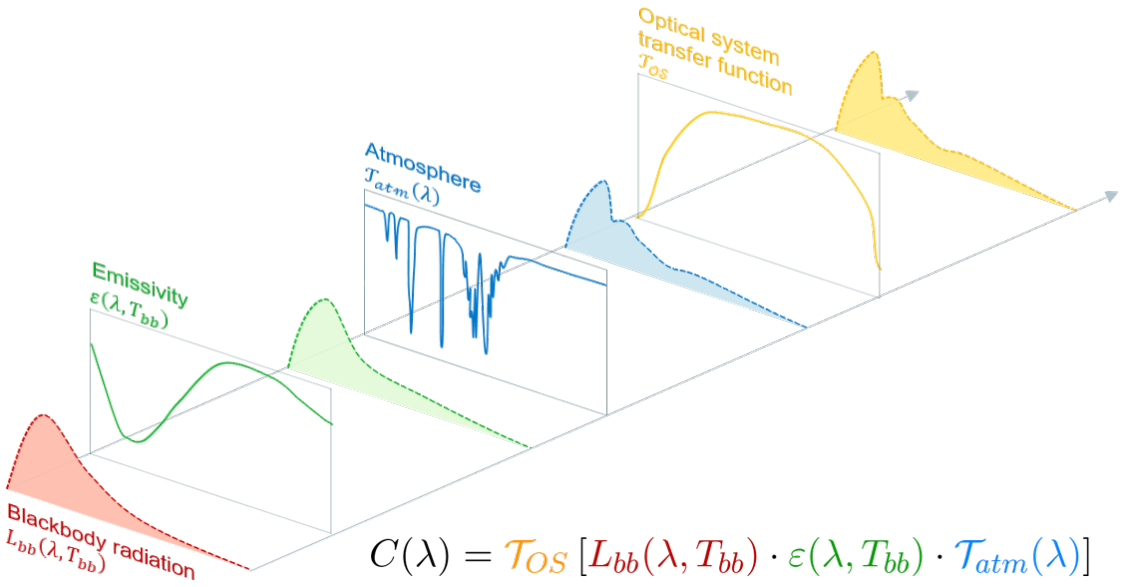


Figure 4.7: Radiative transfer model: The radiance from a blackbody emitter at the unknown sample temperature is successively filtered by the spectral emissivity of the sample, atmospheric transmittance of the optical path, and transfer function of the capturing optical system and sensor. The model is solved through Bayesian probabilistic inference and yields full probability density estimates of sample temperature, spectral emissivity, atmospheric CO_2 and H_2O concentrations, and other auxiliary variables. Dashed lines represent signals, continuous lines represent the different steps modeled by their spectral transmittance.

Thus, the calibrated radiance being observed by our system, $L_{obs}(\lambda)$ can be obtained by inverting the precomputed optical system's transfer function (resulting from the calibration process described in section 4.3.2) and directly applying it to each captured observation:

$$L_{obs}(\lambda) = \mathcal{T}_{OS}^{-1}\left[\mathcal{T}_{OS}\left(L_{bb}(\lambda, T_{bb}) \cdot \varepsilon(\lambda, T_{bb}) \cdot \mathcal{T}_{atm}(\lambda)\right)\right] \tag{4.4}$$

The remaining terms in (4.4) contain certain variables that are either our target magnitudes (i.e. spectral emissivity, $\varepsilon(\lambda, T_{bb})$ and temperature T_{bb} of the equivalent ideal blackbody) or side parameters that need to be estimated from our observations. In order to solve for these all simultaneously we adopt a probabilistic programming framework and build a Markov-Chain Monte Carlo (MCMC) based Bayesian inference model. The modeling details of each of such terms are introduced in the rest of this section:

The **ideal blackbody radiance**, $L_{bb}(\lambda, T_{bb})$, is defined as:

$$L_{bb}(\lambda, T_{bb}) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T_{bb}}} - 1} \tag{4.5}$$

where k_B is the Boltzmann constant, h is the Planck constant, and c is the speed of light. Equation (4.5) shows the heavy dependence of the radiance $L_{bb}(\lambda, T_{bb})$ with the target temperature T_{bb} of the sample. An accurate estimation of this parameter is thus critical. We consider such temperature a stochastic variable to be estimated by the probabilistic model.

The **spectral atmospheric transmittance**, $\mathcal{T}_{atm}(\lambda)$, can be defined as a function of the distance, d, between the capturing device and the observed radiating sample, and the attenuation coefficient of the atmosphere, γ_{atm}. The latter can be further decomposed in terms of the molar concentration x and the unitary absorption coefficient γ of each of the considered absorbents, a. Therefore:

$$\mathcal{T}_{atm}(\lambda) = e^{-d\cdot\gamma_{atm}} = e^{-d\cdot\sum_{a} - x_a\cdot\gamma_a} \tag{4.6}$$

The distance d is set manually according to the acquisition set-up, while the attenuation coefficients are obtained from the HITRAN2016 (High Resolution Transmittance) molecular spectroscopic database [87], which comprises spectroscopic parameters for a number of gaseous molecules with a high spectral resolution, and constitutes the *de facto* standard for the simulation of atmospheric molecular absorption under arbitrary conditions. The HITRAN Application Programming Interface (HAPI) [126] was used during this work in order to obtain accurate and line-by-line information about the absorbents being considered in our measurements. HAPI provides the molecule-specific attenuation coefficient information in the form of cross section (σ_i) with $[^{cm^2}/_{molecule}]$ units, which we then convert appropriately. In our case, the molar concentration of the absorbents (x_a) are unknown probabilistic variables to be estimated by the model. We initially considered a set of potentially significant absorbents comprising H_2O, CO_2, O_3, CO, CH_4, N_2O, and O_2. However, after simulating the spectral atmospheric transmittance due to each of them within a range of typical concentrations, sampling distances and ambient conditions in industrial environments (see Fig. 4.8), only H_2O, CO_2 where kept in the model as the ones with non-negligible contributions to the overall absorbance, in order to reduce potential sources of overfitting. Hence, (4.6) can be further decomposed:

$$\mathcal{T}_{atm}(\lambda) = e^{-d\cdot\gamma_{atm}} = e^{-d\cdot(x_{CO_2}\cdot\gamma_{CO_2} + x_{H_2O}\cdot\gamma_{H_2O})} \tag{4.7}$$

Fig. 4.9 shows the effect of the resulting typical atmospheric transmittance over the emission of an ideal blackbody at 1550 °C at a sampling distance of 1.5 m.

The definition domain of the **spectral emissivity**, $\varepsilon(\lambda, T_{bb})$, of the observed radiative source -our main target variable, which is continuous as a physical magnitude- is given by those of the three spectrometers of the capturing device, which, together, sample the continuous spectrum at $\lambda_i\ \exists\forall\, i \in [1, N]$.

However, in order to prevent overfitting, we regularize the spectral emissivity with a set of M probabilistic variables ε_k, with $k = 1,\dots M$ and $M \ll N$, each paired with one specific anchor wavelength in the considered range, and whose values represent the value of the spectral emissivity of the radiative source at each of such spectral sampling points.

In order to obtain the spectral emissivity defined for every value of λ_i, we take advantage of the smooth variation of the spectral emissivity function [23], and pose M fuzzy sets defined by their triangular membership functions μ_k:

$$\mu_k(\lambda) = \begin{cases} 1 - \left|\frac{\lambda-\lambda_{ck}}{D}\right| & \lambda_{ck} - D < \lambda < \lambda_{ck} + D \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

where λ_{ck} is the central wavelength corresponding to ε_k, $\forall k \in [1,\dots M]$. Using this representation, the spectral emissivity at any arbitrary wavelength i can be estimated as the weighted value of the peak emissivity values, ε_k, defined at the center of the each fuzzy set (Fig. 4.10).

Multi-spectrometer setup. Although useful for standalone-spectrometers, the presented model does not take into account the case where the observed spectral radiance signal is a combination of a set of various spectrometers operating in adjacent, partially overlapping regions of the spectrum. Regions perfomning high noise were removed from the model. For spectrometer 1 wavelengths between $0.2 - 0.4um$ and $0.9 - 1.2um$

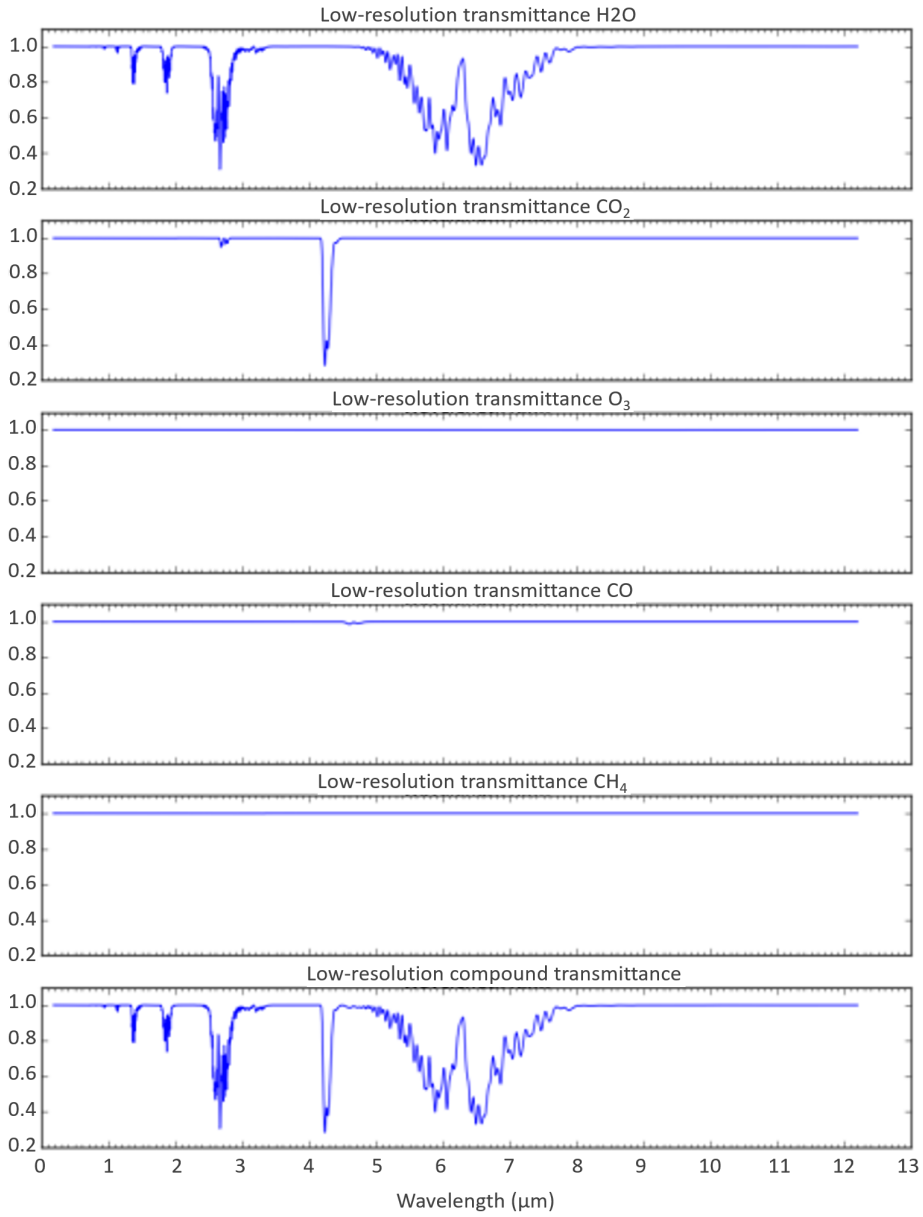


Figure 4.8: Spectral atmospheric transmittance due to each of the considered absorbents (H_2O, CO_2, O_3, CO, CH_4, N_2O, and O_2.) and the combined transmittance for typical concentrations and 27°C, at a distance of 1.5 m.

were removed due to spectrometer lack of sensitivity. For spectrometer 2, wavelengths between $2.5-2.7um$ were removed and for spectrometer 3, wavelengths between $1.6-3.0um$ and were removed. The optical system of our device comprises, though, three different spectrometers, six bunches of optic fiber and a

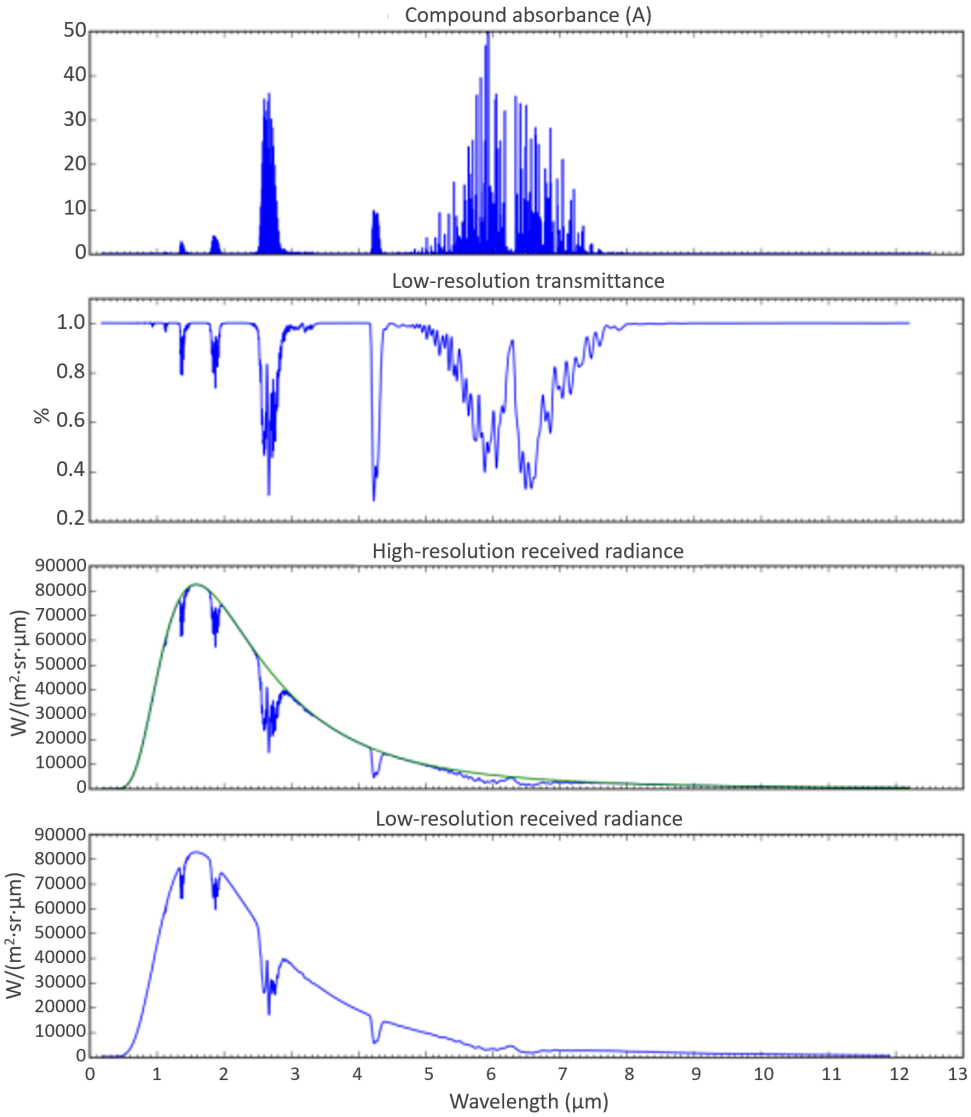


Figure 4.9: Radiance from an ideal blackbody at 1550°C filtered by simulated atmospheric transmittance due to H_2O, CO_2 absorption at typical concentrations and 27°C, sampled at distance of 1.5 m. 1^{st} row) Combined absorbance of the optical path at high spectral resolution, as yielded by the HAPI model (i.e. line-by-line cross-section information). 2^{nd} row) Equivalent transmittance as a result of converting the original, line-by-line magnitude to a low resolution one -matching those of the spectrometers- by convolving the signal with the slit function that characterizes each of the spectrometers' sampling processes (a *sinc* function for the FTIR spectrometers and a triangular function for the non-FTIR one). 3^{rd} row) Blackbody radiance and high resolution radiance once filtered by the atmosphere. 4^{rd} row) Atmosphere-filtered radiance convolved and downsampled to match the spectrometers' resolution.

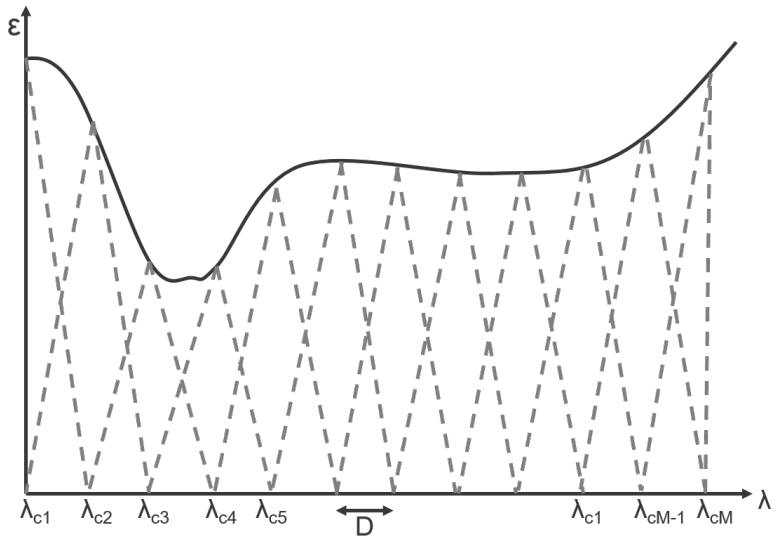


Figure 4.10: Triangular shaped membership functions μ_k, defined over their respective M central wavelengths $\lambda_{ck}, k = 1 \dots M$ and with a distance D between adjacent central wavelengths λ_{ck}.

reflective collimator, whose aperture causes that the three spectrometers do not measure exactly the same area. In order to account for this and for possible fiber misalignments, we introduce a set of three additional proportionality correction probabilistic variables k_s (with $s = 1,2,3$) into the model. These are constant for every value of λ within each spectrometer, and their prior value can be pre-calculated by the use of a calibration lamp as shown in (4.2) and explained in section 4.3.2. Minor variations over the pre-computed K_s can be estimated by the model should they provide a better explanation (i.e. higher likelihood) of the observed data.

The resulting radiative transfer model is thus:

$$L_{obs}(\lambda) = \mathcal{T}_{OS}^{-1}\left[\mathcal{T}_{OS}\left(k_s \cdot e^{-\gamma_{atm}(\lambda)\cdot d} \cdot \varepsilon(\lambda, T_{bb}) \cdot L_{bb}(\lambda, T_{bb})\right)\right] \tag{4.9}$$

which contains M+6 unknown parameters that need to be estimated, i.e. $\theta = \{T_{bb}, \sigma, x_{CO_2}, x_{H_2O}, k_1, k_2, k_3, \varepsilon_k\}\ \forall k \in [1 \dots M]$. T_{bb} was directly one of our ultimate targets, and the other one, i.e. $\varepsilon(\lambda, T_{bb})$, can then be reconstructed for every value of $\lambda_i\ \forall i \in [1 \dots N]$ via the membership functions (μ_k) from $\varepsilon_k\ \forall k \in [1 \dots M]$.

4.4.2 Solving the model through probabilistic inference

The only information available to estimate these unknown variables are the observations acquired by the capturing device, L_{obs}, and the radiative transfer model defined by (4.9). In order to do so, a probabilistic Bayesian inference approach is proposed. The field of Bayesian inference addresses the classical problem of fitting a probabilistic parametric model —our equation (4.9)— to the available noisy observations, i.e. learning the values of the model parameters, θ, that best explain the measured data, assuming that the model describes the data generation process faithfully. As opposed to well-known point estimators for such parameters (e.g. Maximum Likelihood, or Maximum A Posteriori), Bayesian inference constructs a full

probability distribution over the values of each model parameter, in such a way that we gain insight about the goodness of our fit, effectively modeling the inherent uncertainty of any measurement process and accounting for the fact that there is no single optimal set of parameter values which are compatible with the noisy observations.

At the core of this approach lays the classical Bayes' formula:

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)} \tag{4.10}$$

where:

- $P(\theta \mid x)$, known as the *posterior probability distribution*, is the probability of each parameter value given the measured data. This is ultimately our quantity of interest, the one we need to compute.

  In this case we have:

  $$P(\theta \mid x) = P(T_{bb}, \sigma, x_{CO_2}, x_{H_2O}, k_1, k_2, k_3, \varepsilon_1, \ldots \varepsilon_M \mid L_{obs}(\lambda)) \tag{4.11}$$

  and we are ultimately willing to know what are the sample temperature (T_{bb}) and emissivities $\varepsilon_k(T_{bb})$ that best explain the measured radiance data.

- $P(x \mid \theta)$ is the *likelihood*, i.e. how we think the data is distributed given a parameter set θ, or how likely it is that the data was generated by our model with such given parameter set. This is what we use to evaluate how well our model explains the data, and where we describe how our data was generated, guided by (4.9).

- $P(\theta)$ is the *prior*, i.e. the probability distribution over the different parameter values. This is the magnitude that we can use to incorporate any prior knowledge we could have over the parameter values. Table 4.2 summarizes the prior assignments used in our solution.

- $P(x)$ is the *evidence* that the data was generated by this model, which could be computed by integrating over all possible parameter values: $P(x) = \int_\theta P(x,\theta) d\theta$. For non-trivial models, though, we are not able to compute this integral in a closed form.

The intractability of $P(x)$ makes the exact computation of $P(\theta \mid x)$ impossible in most of real-world examples. However, we can try to approximate the posterior making use of Markov Chain Monte Carlo (MCMC) methods, which work by constructing a Markov Chain that generates samples yielding a distribution that matches that of the posterior:

$$P(\theta \mid x) \propto P(x \mid \theta)P(\theta) \tag{4.12}$$

This can be achieved by defining just the priors and likelihood, thus avoiding the need to work with the evidence term. The main of the few caveats to be taken into account is that the samples from the resulting distribution are not independent i.e. there is a certain non-zero serial correlation between them, but the resulting distribution can be monitored to this respect so as to ensure a good-enough *mixture* (i.e. sampling).

The subject of probabilistic programming has recently emerged as a field that helps programmatically describe how the available data have been generated in terms of random variables, probability distributions and deterministic relations that can be used to model real-world processes. Specific programming languages exist that follow and implement this paradigm, but some general-purpose languages, such as Python, also offer library-based approaches to it, such as PyMC3 [234], which was our library of choice.

The first modeling decision was to represent the difference between the expected theoretical spectral radiance (given by the successively filtered Planck's law) and the obtained measurement as a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where each sample corresponds to a given wavelength. This corresponds to our likelihood. We can thus define our observed radiance data, L_{obs}, as:

$$L_{obs} \sim \mathcal{N}(\mu, \sigma^2) \tag{4.13}$$

Where the expected value of the distribution, μ, around which the measured samples are expected to be located, follows a deterministic transformation of the random variables according to (4.9):

$$\mu = L_{expected}(\lambda) \tag{4.14}$$

And the standard deviation of the distribution is also a random variable with a prior following a Half-Cauchy distribution [76] with a fixed $beta$ parameter value:

$$\sigma \sim HalfCauchy(\beta = 10) \tag{4.15}$$

This means that we are assigning high likelihoods to those unknown random variables that generate a theoretical $L_{expected}(\lambda)$ radiance which is numerically close to the $L_{observed}(\lambda)$ radiance acquired by our device. MCMC obtains the posterior probability $P(\theta \mid x)$ by sampling from some probability distributions that represent our prior knowledge over each of the unknown model parameters from the model independently of the observed data. We define such parametrized prior distributions in Table 4.2, by observing their expected occurrence in nature:

- T_{bb} is the expected temperature distribution of the target sample. We use the minimum and maximum temperature values and set a uniform distribution between those values.
- The values of the molar concentrations of CO_2 and H_2O are based on real typical atmospheric ranges.
- K_s are pre-calculated with a calibration lamp as defined in (4.2) (section 4.3.2) and the model assumes that there can be some variation over this pre-calculation. Consequently, they are modeled as normal variables.
- ε_k is the expected value of the spectral emissivity of the radiative source at the k^{th} control wavelength.
- σ models the intensity of the noise existing on the captured signal, independently of the region of the spectrum where it is produced. The extremes of the spectral range of each of the spectrometers are more prone to larger noise levels, while central wavelengths of the respective ranges exhibit less noise. The imposed Half-Cauchy distribution (4.15) aims at modeling such behavior.

Once we have described the observed data generation process (atmosphere-corrected and calibrated radiance as given by the spectrometers), we can apply MCMC to obtain samples from the posterior distribution of each of the unknown variables. Note that, as a result, we will have simultaneous probabilistic estimates for all the random variables that we defined along this description. These will be our model parameters:

$$\theta = \{T_{bb}, \sigma, x_{CO_2}, x_{H_2O}, k_1, k_2, k_3, \varepsilon_1, \ldots \varepsilon_M\} \tag{4.16}$$

In reality, MCMC comprises a full family of distinct algorithms for generating such samples. One of the most efficient ones in the task of going over the full multidimensional posterior space generating samples is the NUTS (No-U-Turn Sampler) sampler [99], used in our solution, which takes advantage of the use of

Parameter	Name	Prior distribution
T_{bb}	Sample temperature	$\sim \mathcal{U}(min = 400°C, max = 1500°C)$
x_{CO_2}	Molar concentration of CO_2	$\sim \mathcal{N}(\mu = 450ppm, \sigma^2 = 50^2)$
x_{H_2O}	Molar concentration of H_2O	$\sim \mathcal{N}(\mu = 36000ppm, \sigma^2 = 500^2)$
k_s	Spectrometer-wise misalignment proportionality constant	$\sim \mathcal{N}(\mu =$ [closed form K_s, eq. (4.2)]$, \sigma^2 = 0.001^2)$
ε_k	Spectral emissivity anchor value	$\sim \mathcal{U}(min = 0.0, max = 1.0)$
L_{obs}	Observed radiance	$\sim \mathcal{N}(\mu =$ [closed form, eq. (4.9)]$, \sigma^2)$
σ	Standard deviation of the $\mathcal{N}$ modeling L_{obs}	$\sim HalfCauchy(\beta = 10)$

Table 4.2: Prior distributions assigned to each of the random variables from our Radiative Transfer Model from (4.9).

the gradient for such duty. In doing so, the more data samples we have, the lower will the uncertainty in the parameter estimation be, and the less the serial correlation of the samples. This will be reflected in narrower monomodal distributions for these estimates. Also, for the same sample size, the uncertainty will grow as the number of parameters we need to estimate increases.

In summary, a complete Bayesian probabilistic model was built, which is sensitive to both the observed data and the theoretical framework comprising the optical calibration transfer function of the device, the proportional misalignment and calibration correction factor of the individual spectrometers, the spectral emissivity and temperature of the radiative sample, and the atmospheric absorption corresponding to the optical path between the device and the observed sample. The joint modeling of these parameters by means of a Bayesian inference model implemented via probabilistic programming makes the developed model remarkably flexible and robust, and enables the simultaneous estimation of all the unknown parameters.

4.5 Experimental validation

In order to validate the proposed method and acquisition system, two different reference materials were chosen: Alumina ($\alpha - Al_2O_3$) and hexagonal boron nitride ($h - BN$), for which the emissivity has already been studied [82, 84, 226, 228]. The spectral emissivity of these materials was carefully measured in laboratory conditions using the HAIRL emissometer and applying a recently upgraded quantification methods [61, 82]. Making use of a dedicated experimental device, infrared spectral directional emissivity measurements were performed accurately in a controlled atmosphere as a function of temperature, emission angle, and *in situ* surface state evolution.These samples have been chosen for its high-temperature structural stability and the presence of a Christiansen wavelength. This is a wavelength in the infrared region that appears in certain ceramic materials at which the emissivity equals to one [226]. It is a very useful feature to obtain the sample temperature at time it is being measured. As the emitted radiation equals to $R = \varepsilon L(T, \lambda)$, where L is the Planck function, knowing the emissivity at a certain wavelength allows calculating the temperature. Since the Christiansen wavelength is temperature-independent, usually it is easily determined at room-temperature by obtaining a spectrum with an integrating sphere.

The spectral emissivity and sample temperature of both materials were estimated under factory industrial conditions by the system and methods proposed in this chapter. The validation of the temperature estimation process was performed by attaching a thermocouple to the sample whereas the spectral emissivities of the

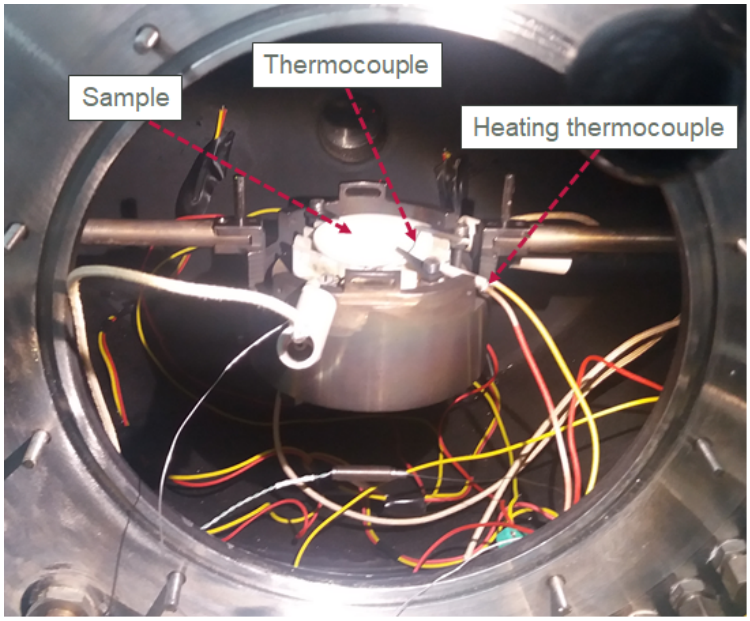


Figure 4.11: Arrangement of the sample in the analysis chamber.

materials were compared with the values measured through the HAIRL emissometer, showing high correlation. This section details the performed experiments.

4.5.1 Experimental setup

Laboratory measurement setup

Normal spectral emissivity measurements were performed in laboratory air using a Fourier-transform infrared spectrometer (FTIR) with a thermal DLaTGS detector (1.43−25 µm spectral range, a reference blackbody (Isotech Pegasus 970-R®) and an optical entrance box that allows switching between the blackbody source and the sample chamber by a rotating plane mirror. As pointed out in [61], a 10° tilting was applied to the sample to avoid spurious signals. The sample was heated with a resistive Kanthal® wire located underneath, as depicted in Fig. 4.11. It was heated at a 0.4 K/s rate in order to avoid cracking from sudden thermal expansion and it was stabilized for 15 minutes to ensure high-temperature stability and constant radiance.

The emissivity measurements were performed between 100°C and 860°C. To determine the actual sample temperature, a non-contact method suitable for materials with low thermal and electrical conductivities was applied. It makes use of the so-called Christiansen wavelength, at which the emissivity is very close to 1 and almost independent of the temperature [82, 226]. The Christiansen wavelength is determined by means of a simple room-temperature reflectivity measurement using an integrating sphere. Then, each measurement temperature is computed by manually forcing the emissivity at the Christiansen wavelength to be the value determined in the reflectance measurement (very close to 1).

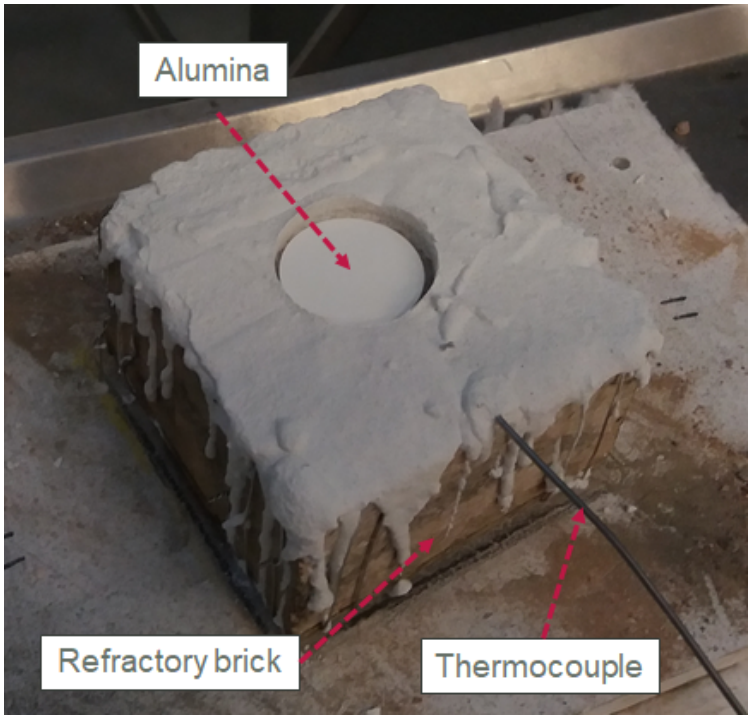


Figure 4.12: Sample holder for the alumina, including the attached thermocouple.

In-field industrial setup

For the industrial conditions measurement, a refractory brick was used to hold the alumina sample. In order to allow time for an accurate temperature measurement, a thermocouple was attached to the sample by means of refractory cement, as depicted in Fig. 4.12. The calibrated acquisition system was positioned at a top position and in-situ calibration was performed following the procedure explained in section 4.3.2 making use of the calibration lamp.

The sample was then heated by means of a LH 15/12 Nabertherm furnace up to 1100°C. After a stabilization time of 10 min the sample was taken off the furnace and placed on the sample stage. The system was monitored simultaneously by an OPTIX PT 50 (1100−1700 nm) external pyrometer and a FLIR T640 (7500−14 000 nm) external thermal camera that were used as control devices. The experimental setup is presented in Fig. 4.13.

Under this configuration, the sample temperature is continuously being monitored by the thermocouple. The sample radiance was also captured by the presented multi-spectrometer device at a rate of 0.75 samples/s, with all radiance in the 200−12 000 nm being captured. The acquired signal was processed by the probabilistic algorithm detailed in section 4.4 and the sample temperature and the spectral emissivities at the anchor wavelengths $\lambda_{ck} = \{0.2, 1.0, 1.5, 2.0, 3.0, 5.0, 6.5, 8.0, 10.0, 12.0\}$µm are estimated. The experiment was stopped when the measured sample temperature reached around 600°C, at which the radiance signal becomes too noisy for our system. This temperature corresponds, in terms of approximate equivalent radiance for the case of the two considered materials, with the lower bound of the calibration range described in 4.3.2, which can be shown shaded in red in Fig. 4.17 and Fig. 4.23.

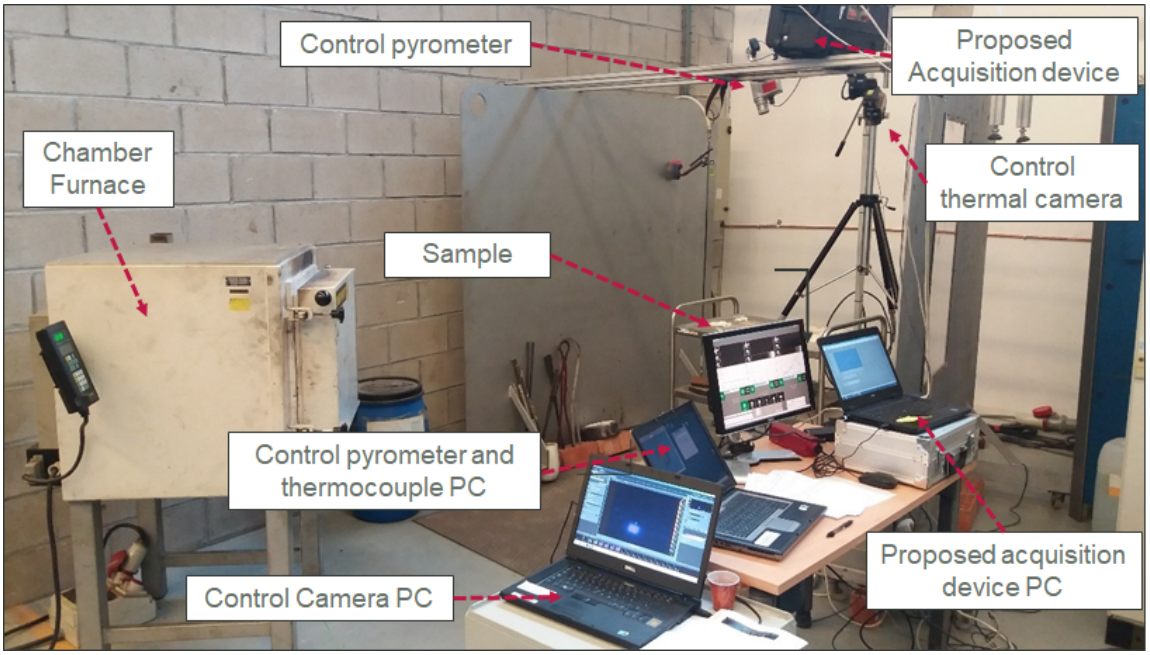


Figure 4.13: Experiment setup. The figure shows the acquisition device mounted on the top part, a control pyrometer and thermal camera, the furnace to heat the sample, the attached thermocouple and the PCs running the acquisition and control software for the different devices.

4.5.2 Analysis of alumina

Alumina ($\alpha - Al_2O_3$) was chosen as a reference material to validate the infrared emissivity measurements performed by the proposed system. Its purity and open porosity were certified by the supplier (McDanel Adv. Ceramic Technologies) to be 99.8% and 0%, respectively. The surface roughness was measured with a profilometer and the average and root mean square values obtained were $R_a = 1.36\,\mu m$ and $R_q = 1.69\,\mu m$, respectively. Then, the infrared reflectance at room temperature was measured (see Fig. 4.14) to locate the Christiansen wavelength, which was thereby at 9.82 µm, with an emissivity value of 0.99. This value is in very good agreement with the expected one according to the data in the literature [226].

The results of the laboratory emissivity measurements following the approach illustrated in section 4.5.1 are shown in Fig. 4.15. The shape of the spectra and the thermal evolution are consistent with previous results in the literature and the typical behaviour of dielectric materials [82, 226]. It is important to note that the emissivity values observed below 4 µm will not be considered in this chapter, because the material becomes semitransparent at those wavelengths and part of the radiation emitted by the heater reaches the detector [82, 112]. The treatment of this effect is out of the scope of the current work. In addition, the feature at 4.18 µm can be explained by the fundamental vibrational mode of CO_2, whereas the H_2O absorption is clearly observed in the $5.5 - 7.5\,\mu m$ range and due to the residual gases present on the optical path. The curve corresponding to the lowest temperature configuration (i.e. 103°C) recorded in this way is also overlaid in Fig. 4.14, showing good accordance with the room temperature reflectance measurement.

Thereafter, an alumina sample was also analyzed following the industrial setup described in section 4.5.1. Fig. 4.16 represents partial results of the algorithm, showing the posterior probability of some of the measured

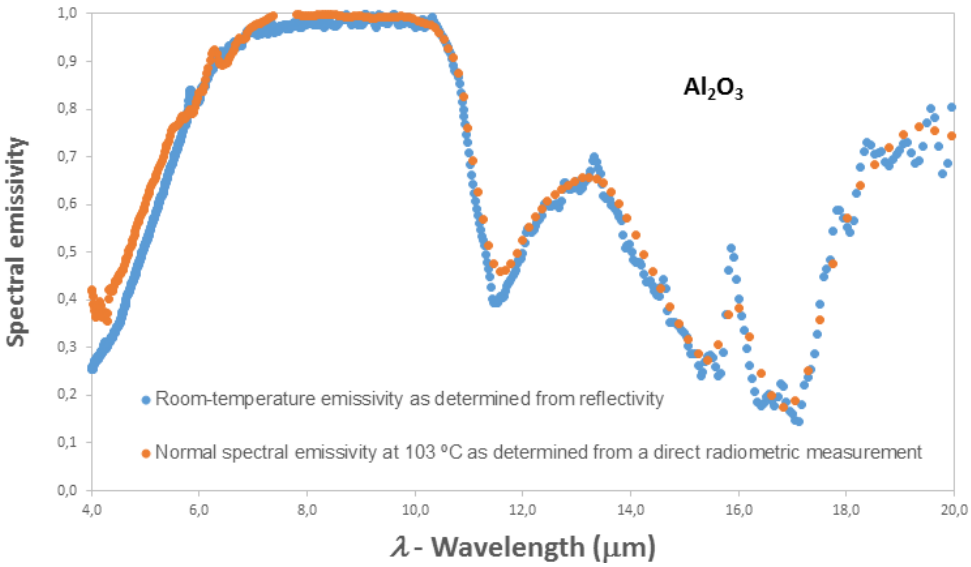


Figure 4.14: Spectral emissivity of the alumina sample measured at room-temperature from reflectivity (blue) and from direct radiometric measurement with the laboratory setup at 103°C (orange).

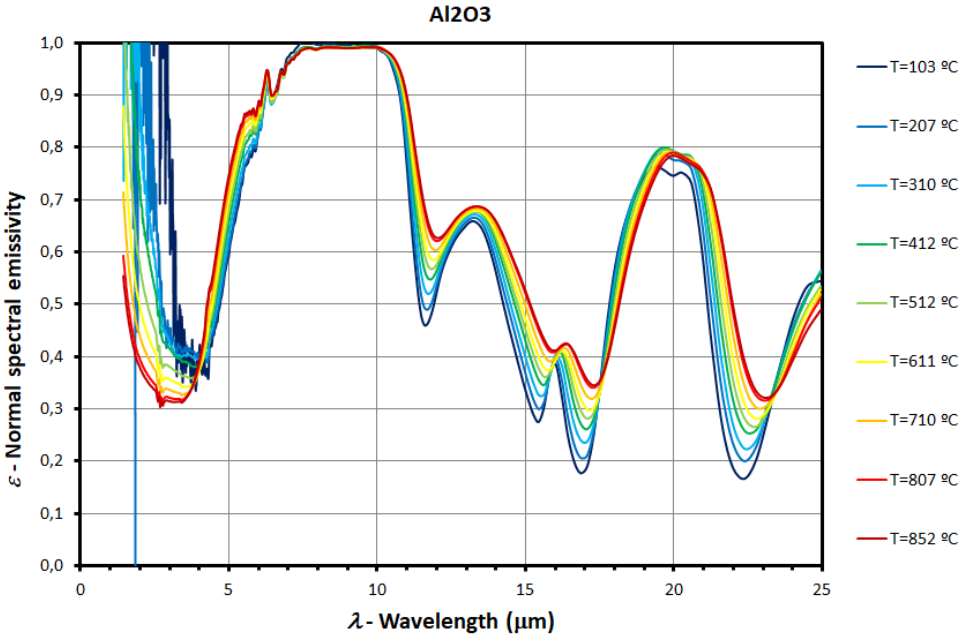


Figure 4.15: Normal spectral emissivity of the Alumina sample as a function of temperature, as measured in laboratory.

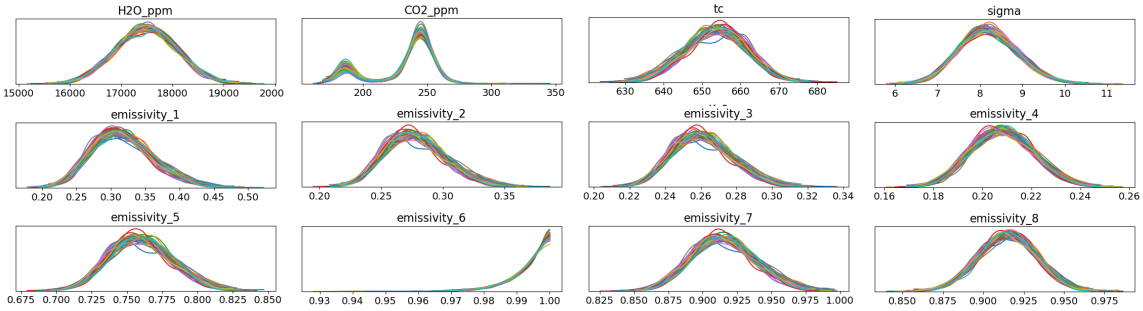


Figure 4.16: Estimated posterior probability of some of the stochastic variables for an alumina sample. Each plot comprises 20 random initializations of the model.

stochastic variables such as sample temperature, spectral emissivities or H_2O and CO_2 concentrations. Note that we obtain full probability density estimations of the considered probabilistic variables, which constitute a more informative outcome than point estimates obtained from other non Bayesian approaches. Fig. 4.17 depicts the remaining components of the output of the algorithm. On the top figure, dotted red points show the signal captured by the device after calibration correction. The blue continuous line represents the theoretical radiation from a blackbody at the temperature given by the thermocouple. The green line represents the radiance of an ideal blackbody $L_{bb}(\lambda, \widehat{T}_{bb})$ at the temperature $\widehat{T}_{bb}$ estimated by the algorithm, whereas the black line represents the estimated radiance $\widehat{L}(\lambda, \widehat{T}_{bb})$ of the sample when applying the spectral emissivity $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm to $L_{bb}(\lambda, \widehat{T}_{bb})$. The magenta line represents the calculated spectrum when applying the estimated attenuation caused by CO_2 and H_2O to the spectrum $\widehat{L}(\lambda, \widehat{T}_{bb})$. On the bottom part of the figure we depict spectral the emissivity $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm as defined in eq. (4.8) and Fig. 4.10, by composition of the posterior probability estimates of the emissivities at the anchor wavelengths ε_k.

Fig. 4.18 shows the comparison between the temperature measured by the thermocouple and the temperature estimated by the proposed system for the alumina at different real temperature values. The comparison yields a Root Mean Square Error (RMSE) of 32.3°C between them and a coefficient of determination $R^2 = 0.96$, thus exhibiting a good correlation across the whole temperature range. Note that the comparatively higher error at the lowest nominal temperatures is consistent with the fact that part of the spectral radiance reaching the system falls below the calibrated range for those temperatures, due to the low emissivity of the sample in such region (see Fig. 4.17). This range also corresponds to the non-zero transmittance region of the alumina.

Finally, Fig. 4.19 represents the emissivities measured under laboratory conditions for the alumina sample at a single temperature setting (see section 4.5.2). The qualitative emissivity behaviour is resembled correctly by the model in the range of 3.6−12.0 µm. The obtained curve is the typical of ceramic materials. From the Christiansen wavelength towards shorter wavelengths, a high-emissivity plateau is observed followed by a steep decrease, whereas towards longer ones the emissivity immediately decreases. For this range it is also observed that there is a more than acceptable quantitative agreement between the algorithm estimation and the laboratory measurement. Therefore, we can conclude that the proposed system is suitable to adequately estimate the temperature and emissivity of alumina. It is worth mentioning here that below 3.6 µm, both curves differ due to the non-zero transmittance of the sample, and part of the radiation from the heating set is therefore detected by the spectrometer. This semi-transparency effect will not be treated here. Besides, in the laboratory-air measurement tracks of CO_2 and H_2O are evident, while the correction applied in the model

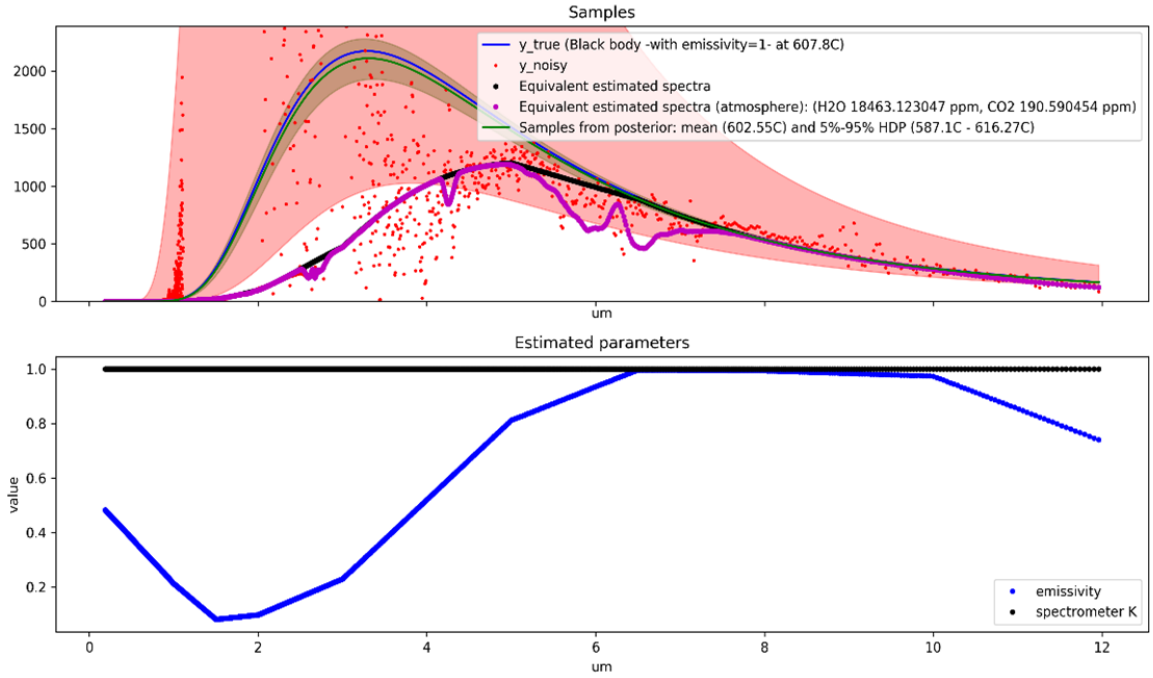


Figure 4.17: Algorithm output for an alumina sample at 607.8°C. Top) Application of probabilistic radiative transfer model to the sample. The blue continuous line represents the theoretical radiation from a blackbody at the temperature given by the thermocouple. The green line represents the radiance of an ideal blackbody $L_{bb}(\lambda, \widehat{T}_{bb})$ at the temperature $\widehat{T}_{bb}$ estimated by the algorithm, whereas the black line represents the estimated radiance $\widehat{L}(\lambda, \widehat{T}_{bb})$ of the sample when applying the spectral emissivity $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm to $L_{bb}(\lambda, \widehat{T}_{bb})$. The magenta line represents the calculated spectrum when applying the estimated attenuation caused by CO_2 and H_2O to the spectrum $\widehat{L}(\lambda, \widehat{T}_{bb})$. Bottom) Emissivities $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm as defined in (4.8) and Fig. 4.10.

.

appropriately removes these "artifacts" from the actual emissivity evolution.

4.5.3 Analysis of boron nitride

A sample of hexagonal boron nitride ($h-BN$) was used as a reference material complementary to alumina in order to validate infrared emissivity measurements by our system. A sample processed by hot pressing of high-purity powders was purchased from Goodfellow, with a range of porosities of 2-15%. The surface roughness was measured with a profilometer and its average value was $R_a = 0.69\,\mu m$. Finally, the infrared emissivity at room temperature was measured indirectly with an integrating sphere in order to locate the Christiansen wavelength, which was found to be at $5.56\,\mu m$, with an emissivity value greater than 0.99 (see Fig. 4.20), which are in agreement with the data reported in the literature [84]. The results of this measurement are shown in Fig. 4.21.

The laboratory set-up was again arranged following the description in section 4.5.1 and the same procedure as with the alumina sample was used. The results of the laboratory emissivity measurements (shown in

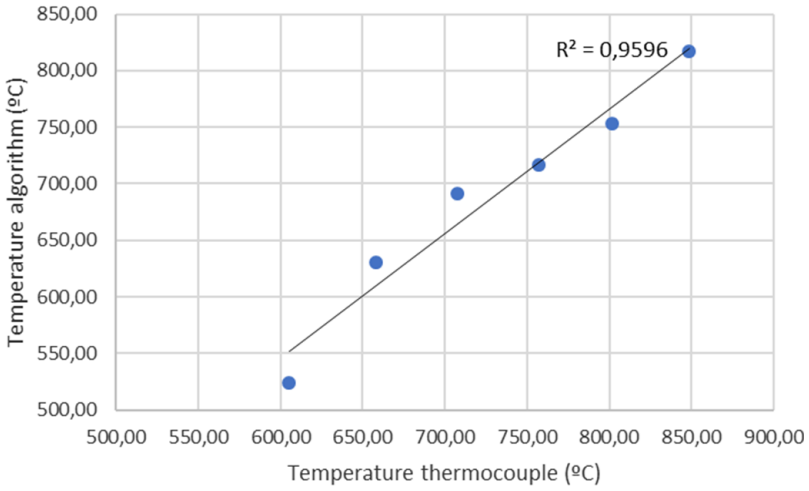


Figure 4.18: Regression graph between the temperature measured by the thermocouple and the temperature estimated by our proposed system and method for the Al_2O_3 sample.

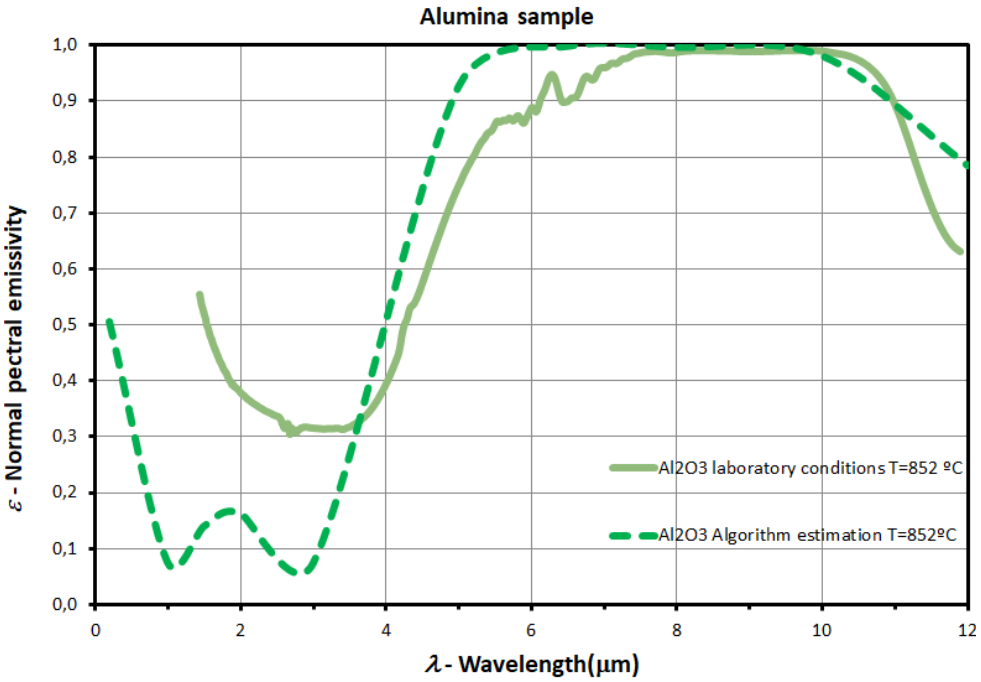


Figure 4.19: Spectral emissivity of the alumina sample determined in laboratory conditions compared to the emissivity estimated by the algorithm under industrial conditions.

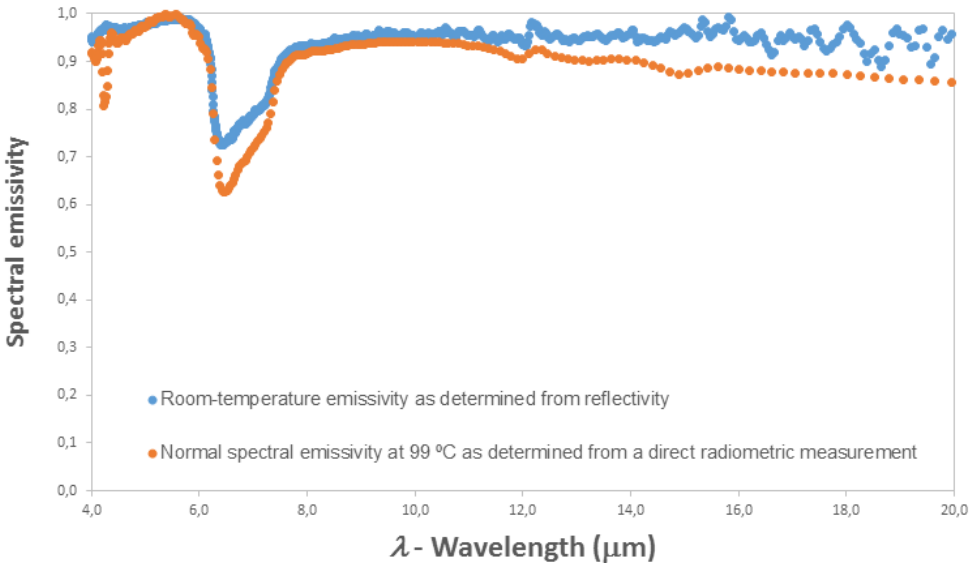


Figure 4.20: Spectral emissivity of the boron nitride sample measured at room-temperature from reflectivity (blue) and from direct radiometric measurement with the laboratory setup at 103°C (orange).

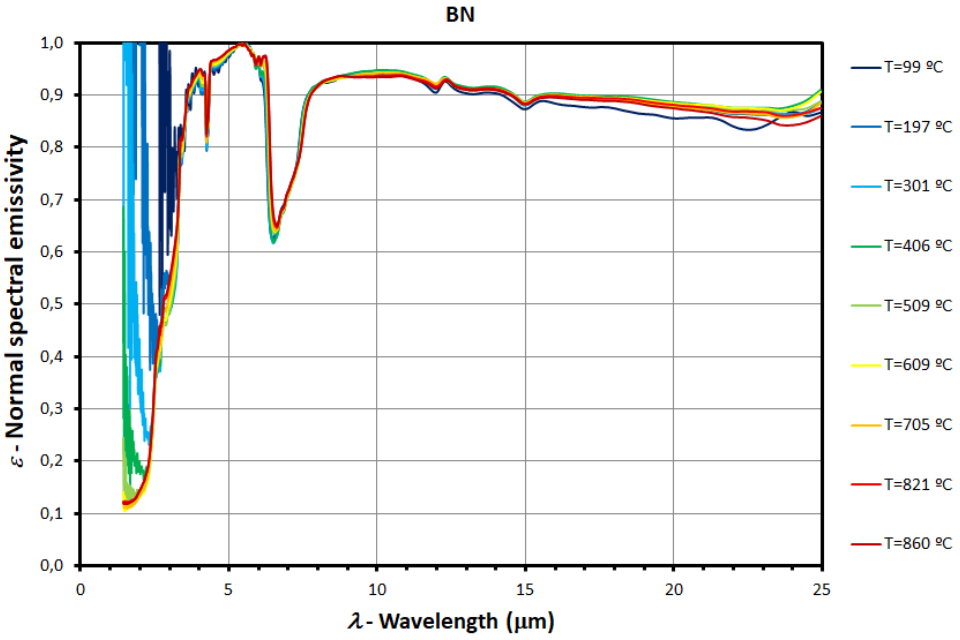


Figure 4.21: Normal spectral emissivity of the hexagonal boron nitride ($h-BN$) sample as a function of temperature, as measured in laboratory.

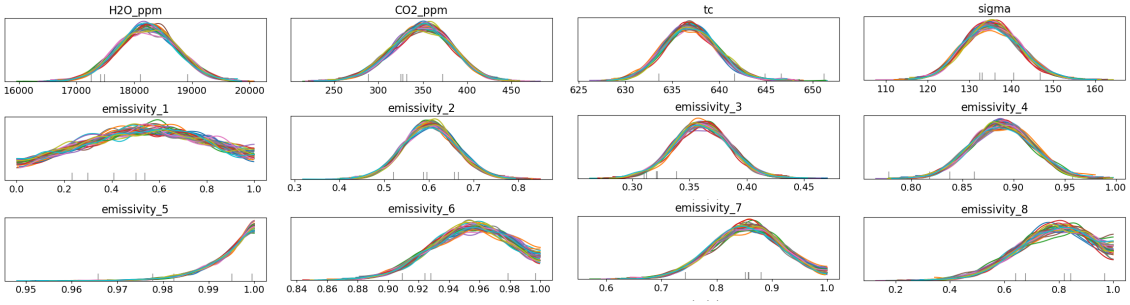


Figure 4.22: Estimated Posterior probability of some of the stochastic variables for a boron nitride sample. Each plot comprises 20 random initializations of the model.

Fig. 4.21 for the 99-860°C range and in Fig. 4.20 for the 99°C case for direct comparison with the reflectivity-based measurement at room-temperature) were again found to be consistent with the literature and the typical behaviour of dielectric materials in terms of shape of the spectra and the evolution as a function of the temperature. As a difference with the measurement acquired from the sample of alumina, the plateau around the Christiansen wavelength is much shorter with decreasing emissivity towards both sides. Of course, as these measurements were also performed under laboratory air the absorption peaks corresponding to $CO2$ and H_2O are also visible. Another particular feature for $h-BN$ in the range of interest for this work, i.e. from 4 to 12 µm in wavelength, is the wide peak observed between 6 and 8 µm, what corrresponds to an infrared active phonon (for details see [84]). Finally, semi-transparency effects for $h-BN$ are not so significant as for alumina below 4 µm, leading to reliable emissivity data down to 2 µm.

For the industrial conditions measurement, the same configuration and procedure as with the alumina experiment was followed, stopping the experiment at 500°C, once the radiance signal became noisy for our system. Fig. 4.22 shows the posterior probabilities of some of the stochastic variables estimated of the algorithm whereas, Fig. 4.23 presents the algorithm's estimations for a boron nitride sample at 599.44°C, following the notation from Fig. 4.17.

The temperature estimation obtained for this sample shows an even better correlation across the whole range with the thermocouple measurements than in the case of the alumina, as can be seen in Fig. 4.24 (RMSE= 5.69°C, R^2 = 0.996). Finally, (Fig. 4.25) compares the emissivities measured under laboratory conditions with those estimated under industrial conditions for a fixed temperature value of 852°C. It can be appreciated that there is a quite good qualitative and quantitative agreement between both in the $2-12\,\mu m$ range, except for the region corresponding to the active infrared phonon (6-8 µm) that is removed from the model. This phenomenon is beyond the scope of the current work.

Furthermore, we can observe that the CO_2 (at 4.2 µm) and H_2O (at 5.8 µm and 6.5 µm) absorptions present in the laboratory measurements are adequately treated by our algorithm and eliminated from the emissivity, thus giving a correct spectra of the actual emissivity.

4.6 Conclusion

In this chapter we presented, to our knowledge, the first capturing device that is able to simultaneously estimate the spectral emissivity and temperature of a hot emissive material under real steel factory conditions. The presented device is capable of capturing 0.2–12 µm range radiance signal at up to 20 m away from the sample, over a 12 mm diameter spot. The device allows for in-situ calibration making use of a stabilized

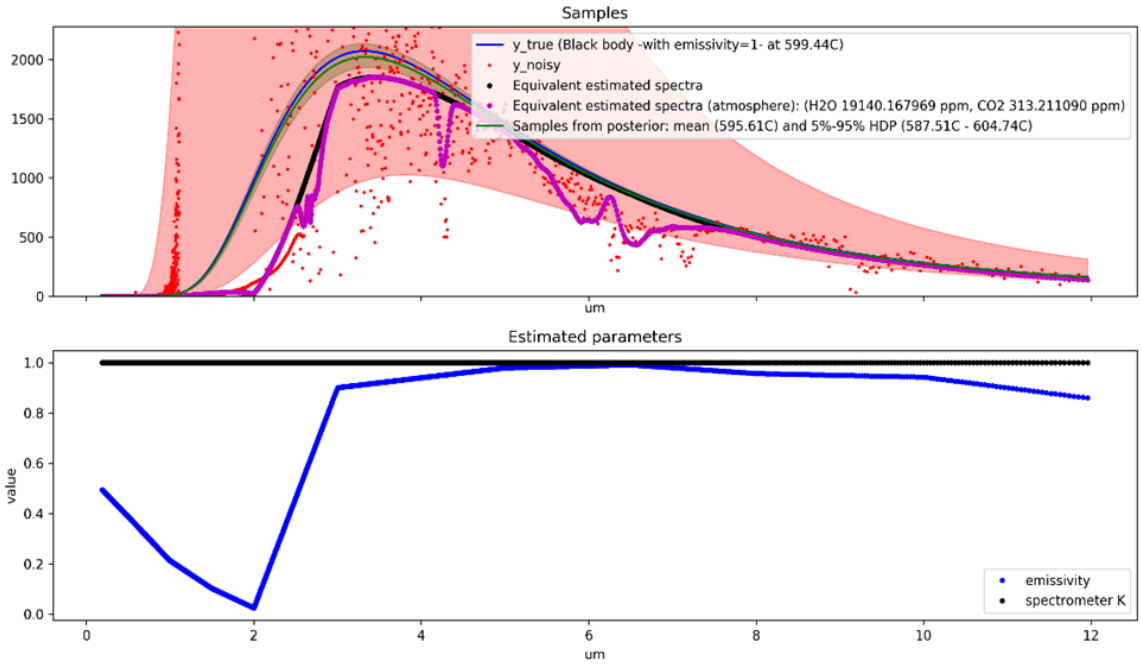


Figure 4.23: Algorithm output for an boron nitride sample at 599.44.8°C. (top) Application of probabilistic radiative transfer model to the sample. The blue continuous line represents the theoretical radiation from a blackbody at the temperature given by the thermocouple. The green line represents the radiance of an ideal blackbody $L_{bb}(\lambda, \widehat{T}_{bb})$ at the temperature $\widehat{T}_{bb}$ estimated by the algorithm, whereas the black line represents the estimated radiance $\widehat{L}(\lambda, \widehat{T}_{bb})$ of the sample when applying the spectral emissivity $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm to $L_{bb}(\lambda, \widehat{T}_{bb})$. The magenta line represents the calculated spectrum when applying the estimated attenuation caused by CO_2 and H_2O to the spectrum $\widehat{L}(\lambda, \widehat{T}_{bb})$. (bottom) Emissivities $\widehat{\varepsilon}(\lambda, \widehat{T}_{bb})$ estimated by the algorithm as defined in (4.8) and Fig. 4.10.

calibration lamp. The system is accompanied by a probabilistic algorithm that is able to produce simultaneous full probability density estimates of the sample temperature and spectral emissivity at different wavelengths, as well as the global concentration of H_2O and CO_2 along the optical path. All these parameters are seamlessly predicted by a Markov Chain Montecarlo-based estimation algorithm as the ones that provide the best explanation for the captured data.

We showed that, by analyzing the radiance between $0.2 - 12\,\mu$m, we were capable of estimating the temperature of alumina and boron nitride samples in a range of 600-1000°C with an RMSE of 32.3°C and 5.69°C, respectively. Spectral emissivity was calculated accurately and the effect of H_2O and CO_2 absorption was also estimated and discounted from the measurement.

These results pave the path for future attempts to estimate the slag composition on the electric arc furnace based on the analysis of the predicted emissivity values.

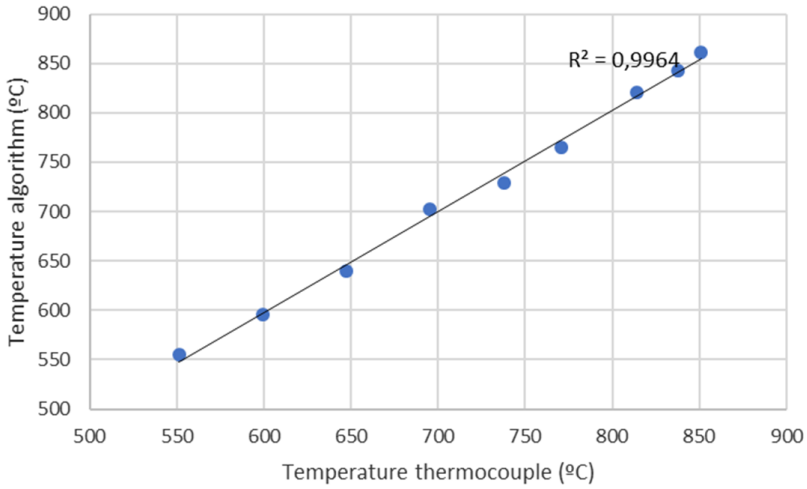


Figure 4.24: Regression graph between the temperature measured by the thermocouple and the temperature estimated by our proposed system and method for the BN sample.

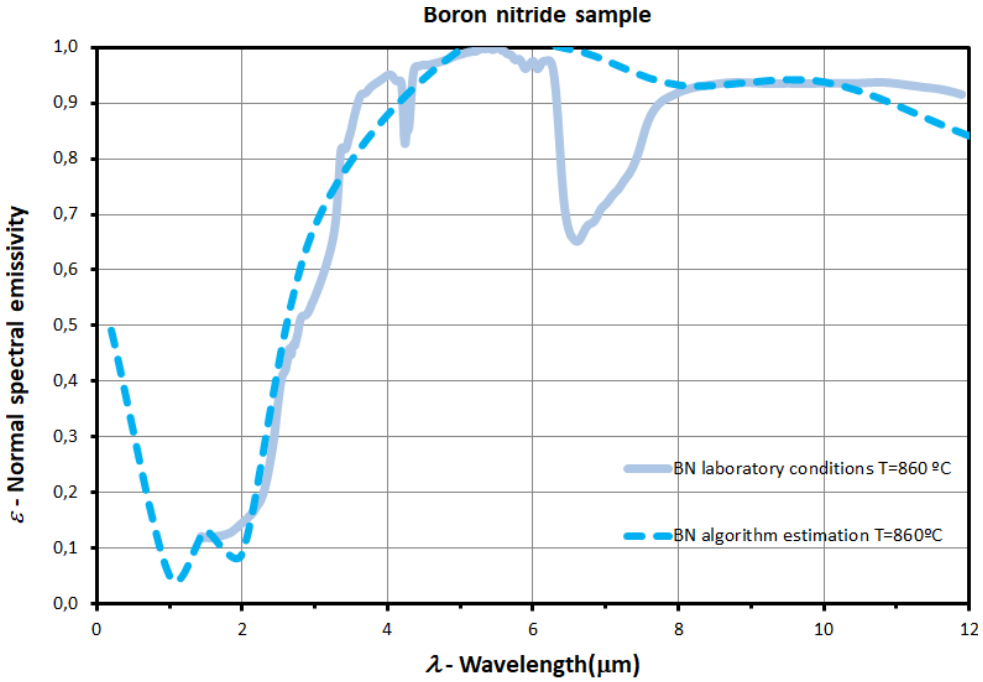


Figure 4.25: Boron nitride emissivity determined in laboratory conditions compared with the emissivity estimated by the algorithm under industrial conditions.

5 MVMO: A Multi-Object dataset for Wide Baseline Multi-View Semantic Segmentation*

5.1 Introduction

The task of *semantic segmentation* [155] aims at, given an input image, performing pixel-wise classification over a predefined set of categories. As in many other dense prediction problems, the end-to-end convolutional neural networks (CNN)-based fully supervised approach to this task has become the *de facto* standard to solve it, leading to robustly performing models [223] at the expense of a large amount of human annotations. Nevertheless, understanding scenes based on a single 2D input is challenging when applied on (i) scenes with significant inter-object and self-occlusions that hide class-distinctive features (ii) scenes covering a wide spatial range, where distant objects can show a small apparent size.

In this context, we hypothesize that posing data-driven models that exploit multi-view camera setups that provide complementary information over the imaged scenes could be of potential interest for improving the results obtained by single-camera baselines. However, so far multi-view semantic segmentation has primarily been approached for close-baseline setups [224] i.e. those where the distance between cameras (and thus, the resulting disparities) are small, whereas solving the aforementioned obstacles requires wide baselines. Scenarios that could benefit from this approach are frequent in real life, in domains as diverse as industry (e.g. conveyor belts), surveillance, or traffic management. While the cost inherent to dense manual annotations is commonly accepted as a necessary toll in monocular segmentation setups, this is hardly the case for multi-view scenarios, which can be efficiently labeled in synthetically modeled scenes.

In this chapter, we introduce **MVMO**, the **Multi-View Multi-Object dataset**, which addresses the current lack of publicly available large-scale datasets of densely annotated wide-baseline multi-view scenes containing multiple objects. MVMO is a synthetic, path tracing-based set of 116,000 scenes with per-view semantic segmentation annotations of 10 object categories. Each scene is observed from a set of 25 camera locations distributed uniformly in the upper hemisphere [55] (see Fig. 5.1). Unlike most existing multi-view image datasets (which are designed to be camera-centric and exhibit very close baselines while sensing their surroundings [224]), MVMO features wide baselines between many camera pairs as a result of a scene-centric design, and a large amount of objects per scene. This leads to large disparities, notable occlusions and variable apparent object geometry, size and surface appearances across views. Therefore, MVMO sets a particularly challenging arrangement that aims at contributing to push research on the fields of multi-view semantic-segmentation and cross-view semantic knowledge transfer, serving as benchmark for the development of viable solutions to be later fine-tuned or domain-adapted and applied to real-life domains. The experiments presented show that simple baselines fail to be of much help in transferring learned models to novel views, hence suggesting the need for novel research in this direction.

*This chapter is based on a conference submission (under review) [5]. The code and dataset are available at: https://aitorshuffle.github.io/projects/mvmo/.

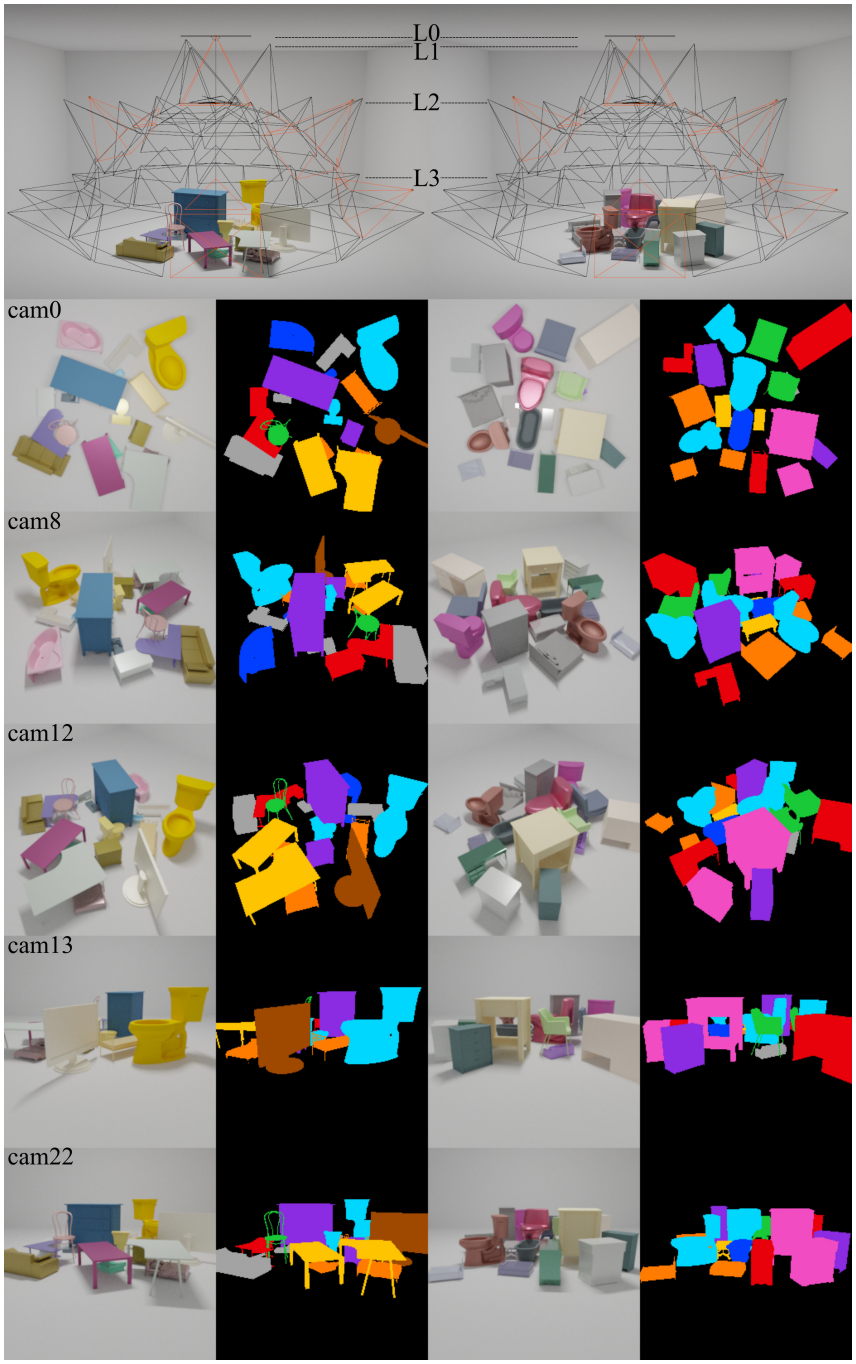


Figure 5.1: Top: two scenes from the proposed 116,000 scene MVMO dataset and the 25 equidistributed camera locations. Bottom: rendered views and semantic ground truth for the 5 camera poses (highlighted) used in our experiments.

Dataset	Wide Baseline	Object Density	Representation	Photorealism	# Scenes	# Views	# Classes
Human3.6M [107]	Yes	Low (1)	2D images	Real	900,000 in 165 sequences	4	24
3Dpeople [201]	Yes	Low (1)	3DM→2D	S: High B: Low	616,000 in 5,600 sequences	4	8(clothes)/14(body)
SYNTHIA [224]	No	N/A	3DM→2D	Low	51,000 in 51 sequences	8	13
ScanNet [48]	⋆	Low	2D→3DS	High	1.5k	⋆	40
House3D [276]	⋆	Low	3DVE	Low	45.6k	⋆	80
Gibson [278]	⋆	Low	3DVE	High (IBR/PCR)	1.4k	⋆	40
CARLA [60]	⋆	⋆	3DVE	Mid-High (RT)	⋆	⋆	12
MVMO (ours)	Yes	High (15-20)	3DM→2D	High (PT, UOM)	116k (uncorrelated)	25	11

Table 5.1: Datasets for multi/cross-view semantic segmentation. The table shows the lack of datasets with wide baseline and high object density addressed by MVMO. **Object Density**: #objects/scene. Does not apply to close baseline scenarios. **Representation**: 2D→3DS: 3D Surface reconstructed from 2D. 3DVE: 3D Virtual Environment. 3DM→2D: 3D Model rendered to 2D images. **Photorealism**: S: Subject. B:Background. IBR: Image-Based Rendering. PCR: Point Cloud Rendering (view synthesis from Point Cloud). RT: Ray-Tracing. PT: Path-Tracing. UOM: Uniform Object Materials. ⋆ Needs to be placed/configured/generated by user; images are not readily available.

5.2 Related work

Our work relates to a number of previous datasets and approaches over them from various research fields. Large scale *single view semantic segmentation datasets* have been of the utmost importance for the development of the field and have been used in diverse domains. The PASCAL VOC 2012 dataset for natural images [65] or the Cityscapes dataset of urban scenes [46] are two representative examples.

At the core of the task of *object co-segmentation* [265], lays the goal of segmenting the object instances of common categories among two or more input images, which do not necessarily correspond to distinct viewpoints of the same scene. Datasets such as iCoseg [17] or [147] have been supporting its evolution for years.

Several classic computer vision tasks have already witnessed various attempts of wide-baseline multi-view datasets, in an effort to improve upon their respective single-view performances. In *multi-view object detection*, [221] introduces a new multi-class, multi object detection dataset with bounding box annotations from 6 calibrated cameras and exploits the perspective diversity to address the detection of pedestrians, cars and buses, hence extending previous datasets (e.g. PETS'09 [67], EPFL [68]), which gather multi-view but single-class annotations from 7 and 4 cameras, respectively.

Advances on *multi-view human pose estimation* were possible by leveraging various wide baseline datasets over RGB [107, 118, 281] and depth [92] images of both groups [114] and individuals. The KTH Multiview Football Dataset II [118] comprises 800 real-life football scenes captured from three views with 2D/3D annotated pose keypoints. [281] scales this to 35-69 views on various single-subject scenes. The ITOP dataset [92] provides front and top view-based human pose keypoint annotations on top of depth input maps. In contrast, in [114] groups of 3 to 8 subjects are captured and their poses annotated through 480 surrounding RGB cameras.

The field of *multi-view semantic segmentation* has been addressed from diverse perspectives. Many early works prior to the irruption of deep learning techniques focused on the binary segmentation of a single static foreground object from a sequence of close-baseline views from a class-agnostic point of view [25, 144], often learning sequence-specific models and being evaluated on just a handful of sequences. They relied on diverse cues, such as modeling of object-background color distributions [25, 144, 245], central object fixation [25], interactive user feedback [245] or stereo geometry constraints [128, 144] that occasionally

led to 3D reconstruction [25, 245]. More recently, [282] used deep self-supervised training to extend the single subject segmentation task to three dynamic scenes in wide-baseline setups. Moving away from the subject/background segmentation, [21] used a few close-baseline stereo image pairs to segment every object from both view references, although without performing any actual semantic classification of the resulting blobs.

Multi-class multi-view semantic segmentation poses harder challenges and calls for larger datasets (see Table 5.1), suitable for data-driven approaches. Different works leverage with varying emphasis the complementary information provided by additional views: [135] extends the Leuven stereo dataset with semantic labels in one of the views to jointly train for segmentation and stereo reconstruction. A few works focus on *cross-view semantic transfer*, with an unsupervised transfer of the semantic annotations to new label-free views, e.g. ground to aerial views [288] or among distinct vehicle-mounted cameras [45] in close-baseline footage from [60]. Both tasks demand datasets comprising two or more 2D RGB views, with annotations in each of them. Our work is most closely related to these. SYNTHIA [224] provides pixel-wise depth and semantic labels for a large synthetic set of scenes captured from a vehicle-mounted 8 RGB camera-rig, thus showing the usual narrow baseline of camera-centric driving setups. The wide baseline scenario has so far only been tackled by the Human3.6M [107] and 3DPeople [201] datasets. They both provide body part [107, 201] or clothing [201] segmentations, but [201] has immutable 2D backgrounds, and they are both restricted to single subjects and thus limited in the severity of the occlusions and subject size variation across views.

Several recent papers [49, 160, 164, 218] leverage the spatial consistencies in temporal sequences of RGB or RGB-D images with small relative baselines among them to address semantic segmentation of either 2D images or their reconstructed 3D representations. The raw sequences of the NYUv2 [243], Camvid, ETHZ RueMonge 2014 [218] or the ScanNet [48] datasets are commonly used to achieve this.

Furthermore, various large scale 3D virtual or reconstructed environments have been released. Their relevance comes from the fact that, through significant user intervention, parts of the 3D model and associated labels could be projected back to 2D to synthesize semantically annotated multi-view image sets from arbitrary camera locations with different degrees of realism. The House3D [276], Gibson [278] and CARLA [60] environments are some relevant examples, although only CARLA, being fully virtual, could yield high object densities via its API. This was shown in [183] for close baseline setups, proposing a multi-view semantic fusion scheme from up to 8 input views onto a new virtual zenithal view.

In conclusion, MVMO covers the lack of a standardised large scale photo-realistic multi-view dataset with wide-baselines (and hence, large disparities and relevant occlusions) across cameras and comprising semantic segmentation annotations for multiple objects of distinct classes.

5.3 MVMO Dataset construction

We set out to construct a synthetic dataset for semantic segmentation, that fills the existing lack of wide-baseline multi-view multi-object datasets. Therefore, we use Blender's Python API for procedural 3D scene construction and image rendering, using the ModelNet10 3D object dataset [277] as repository of well-categorized 3D shapes of 10 common object classes (i.e. *bathtub*, *bed*, *chair*, *desk*, *dresser*, *monitor*, *night_stand*, *sofa*, *table*, *toilet*). We build a basic scene with a grey plane at $z = 0$ and a single zenithal rectangular key light, and define a $2.8 \times 2.8m$ rectangular area for object placement. All cameras are projective cameras with a focal length of $f = 35mm$, oriented to the origin. The camera locations are determined by sampling the surface of a hemisphere of $r = 3m$ regularly so that they are equidistributed [55]. For our set of 25 samples, this yields locations at four levels (Fig. 5.1-top and Fig. 5.2): 1 view at L0 (top, at $z = 3.0m$), 3 views at L1 ($z = 2.90m$), 9 views at L2 ($z = 2.12m$) and 12 views at L3 ($z = 0.78m$).

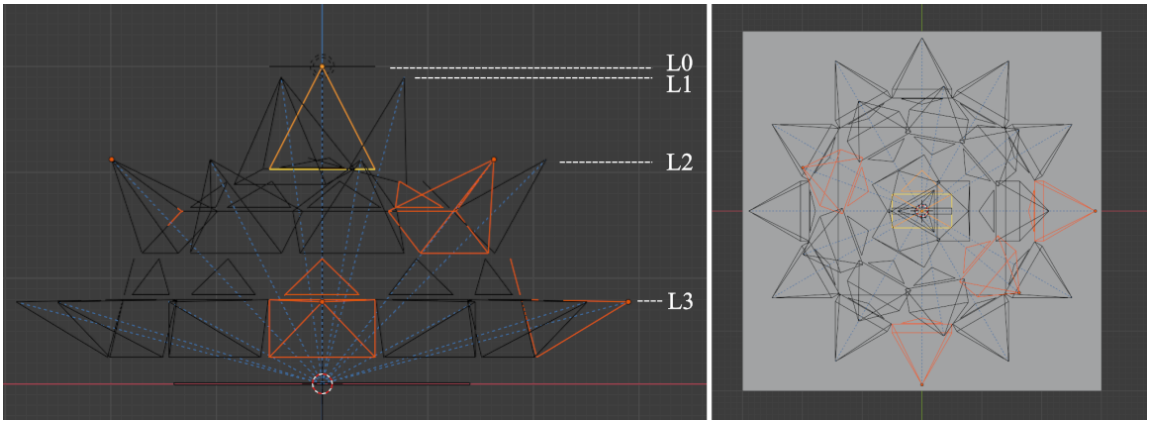


Figure 5.2: Ortogonal lateral and zenithal projections of the set of camera locations distributed in four levels (L0-L3). Highlighted cameras show the poses used in the experiments in section 5.4.

Then, for each scene: (i) we randomly select one of the 10 categories of ModelNet10 and (ii) sample one shape from the selected class, (iii) we normalize its scale so that its largest dimension is $1.0m$, then applying a random scale in the $[0.3-0.8]$ range, (iv) we select a random base-color from a set of 9,284 predefined ones [131] and apply a random combination of the *specularity*, *roughness* and *metallic* material modifier properties that -together with other fixed property values- define the Bidirectional Scattering Distribution Function (BSDF) of the materials applied to the whole shape. (v) we place it on the $z = 0$ plane of our base scene, in a random location (within the designated limit area) and angle, checking that the mesh does not intersect with any previously placed object. (vi) Once $15-20$ objects are placed, the scene and fine-detailed ground truth images are rendered with the *Cycles* engine for each of the 25 views at 256×256 pixels, producing photo-realistic, unbiased and physically consistent shading, reflectance and material effects, including specularities, and interreflections.

The 116,000 created scenes (each with 25 views; see Figs. 5.3 to 5.7 for full samples of five random scenes) were then partitioned in a train set (100,000), two validation and two test sets (4,000 each). The latter are created based on whether the used ModelNet10 shapes were already used for the train set (SO: Same Objects) or come from a held-out set of shapes (OO: Other Objects) from the same categories, which poses a harder problem. Fig. 5.8 shows the resulting distributions of objects per category and scene for the train set.

This proposed wide-baseline multi-object dataset contains many occlusions, making semantic segmentation from a single view difficult. We think MVMO can facilitate research in multiple directions. We highlight two of them: (i) *Multi-view semantic segmentation*: existing close-baseline datasets have only few occlusions. Therefore, the proposed dataset makes for a more interesting setup for multi-view semantic segmentation. (ii) *Cross-view semantic transfer*: this is an especially exciting research direction which can be performed on MVMO. In real-life applications the dense labelling of all views is infeasible. Hence we believe that methods need to be designed that can learn to perform multi-view semantic segmentation based on labels from only a single view.

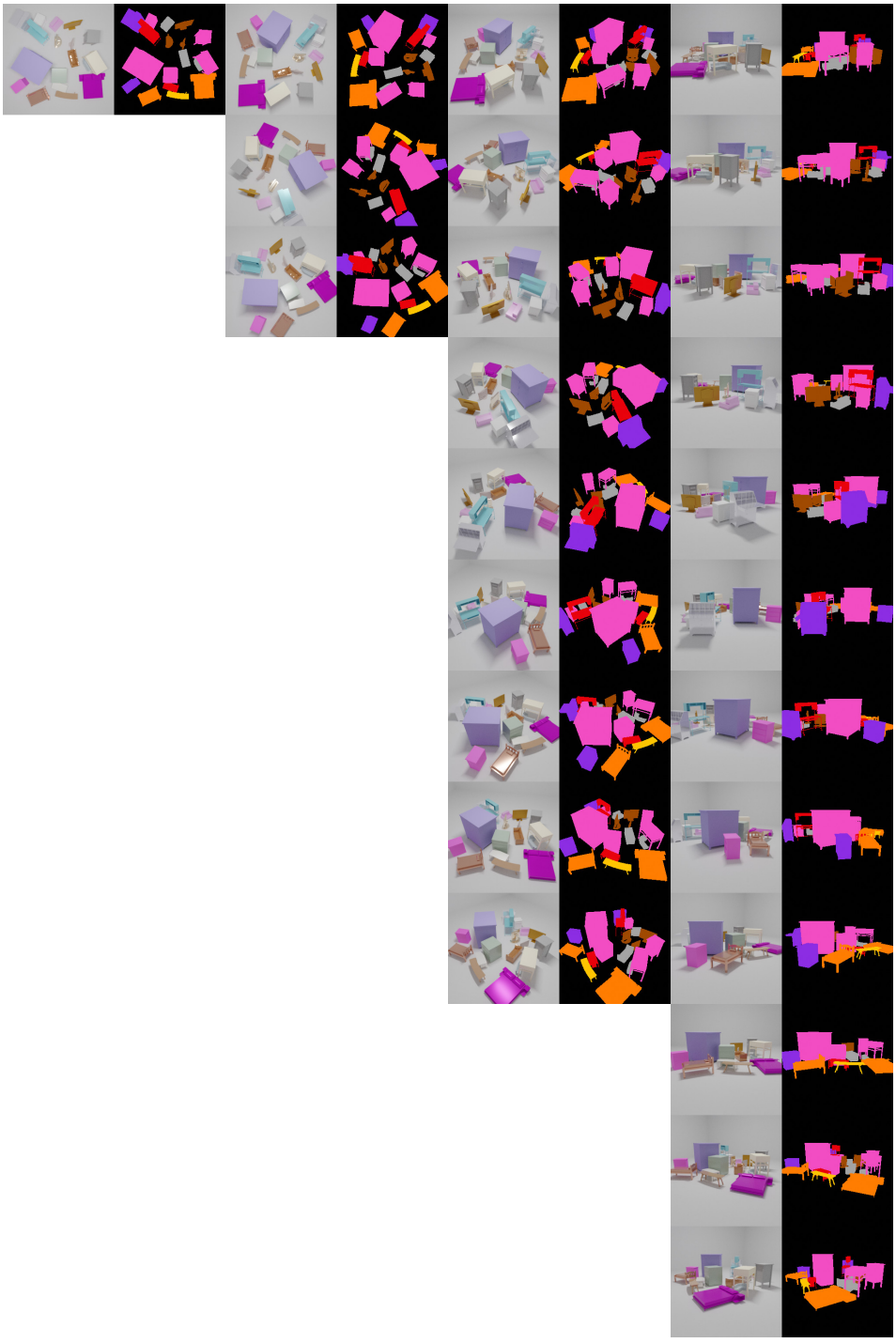


Figure 5.3: The 25 views and semantic maps of a random MVMO scene, sorted from level L0 (left) to L3.

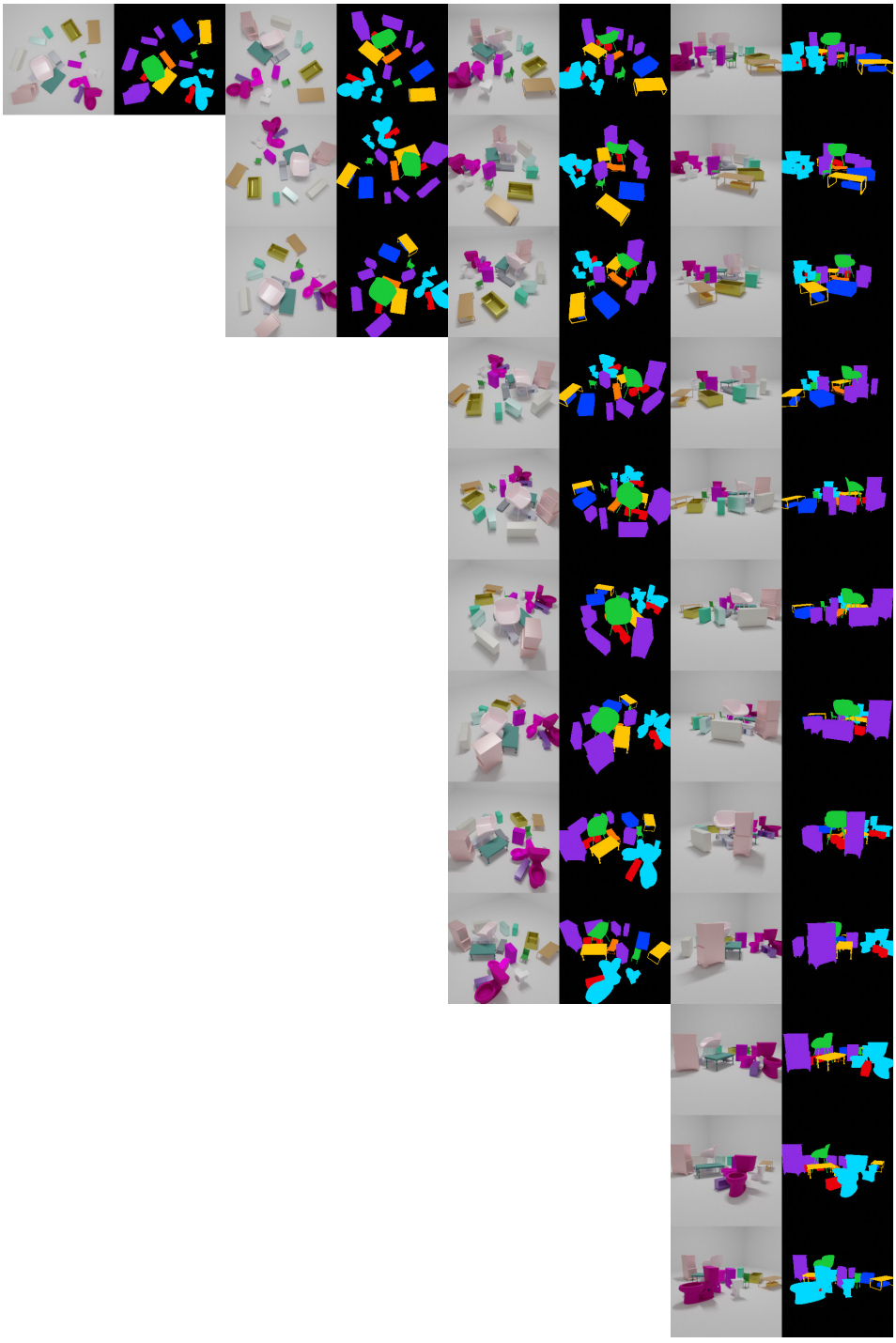


Figure 5.4: The 25 views and semantic maps of a random MVMO scene, sorted from level L0 (left) to L3.

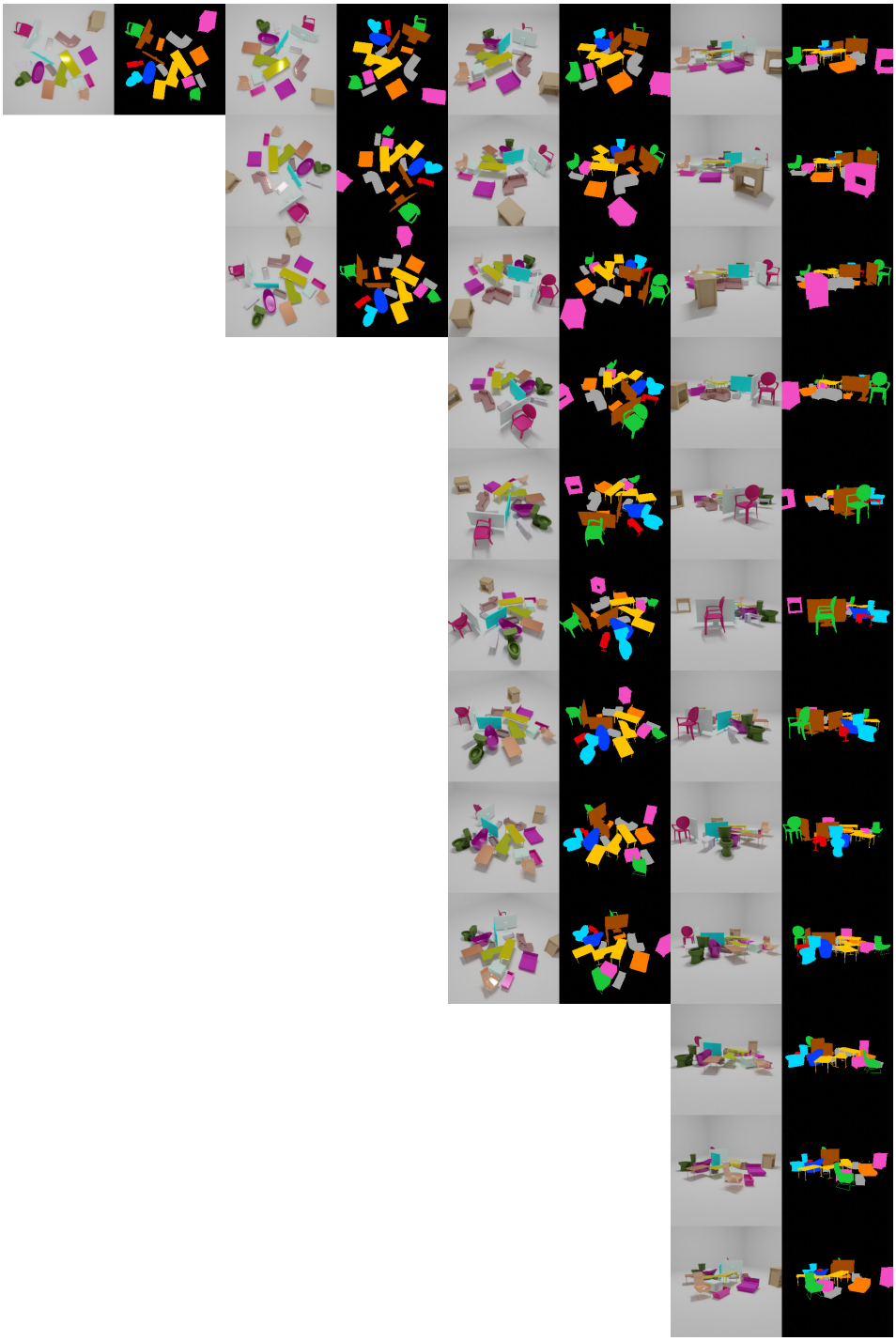


Figure 5.5: The 25 views and semantic maps of a random MVMO scene, sorted from level L0 (left) to L3.

Figure 5.6: The 25 views and semantic maps of a random MVMO scene, sorted from level L0 (left) to L3.

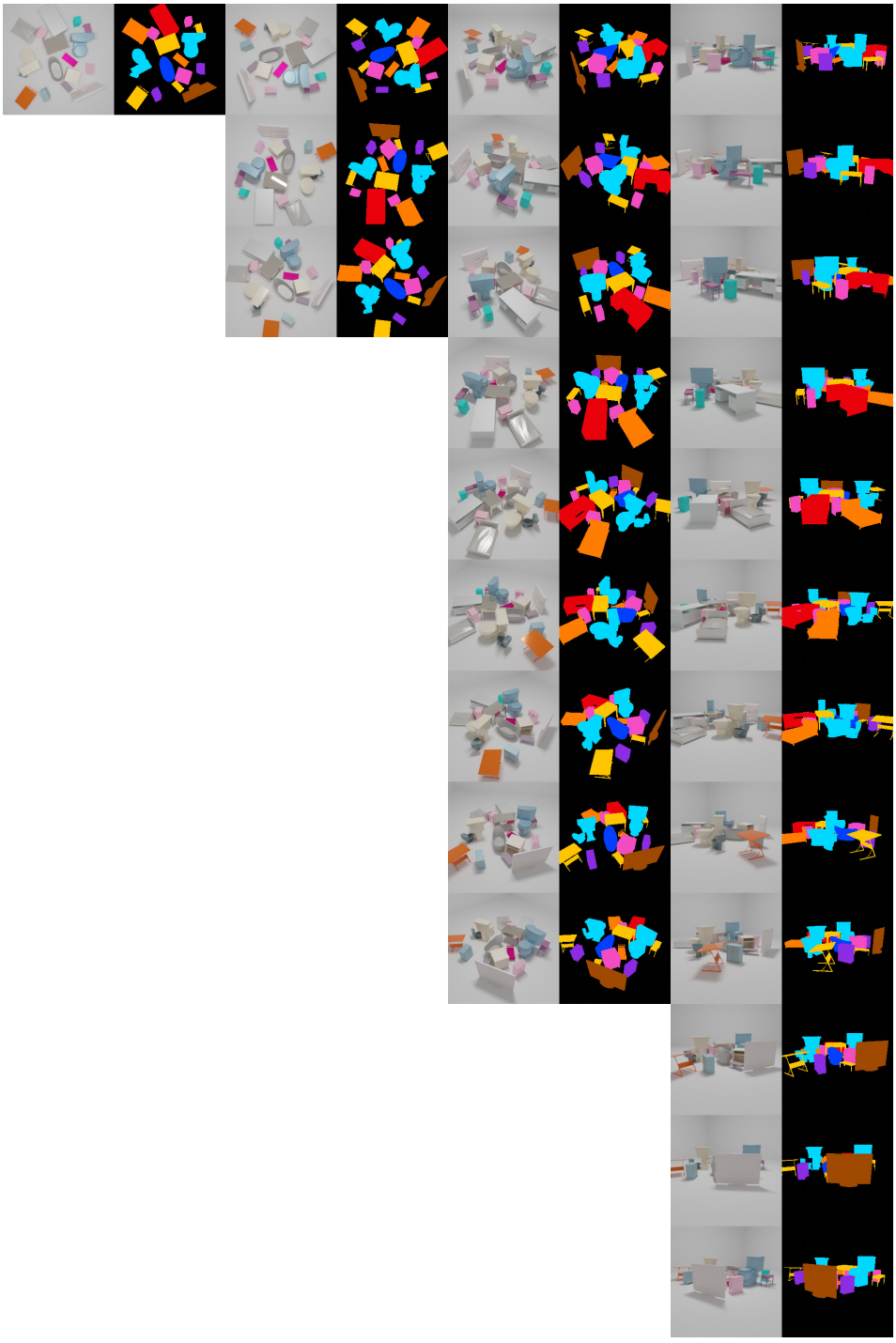


Figure 5.7: The 25 views and semantic maps of a random MVMO scene, sorted from level L0 (left) to L3.

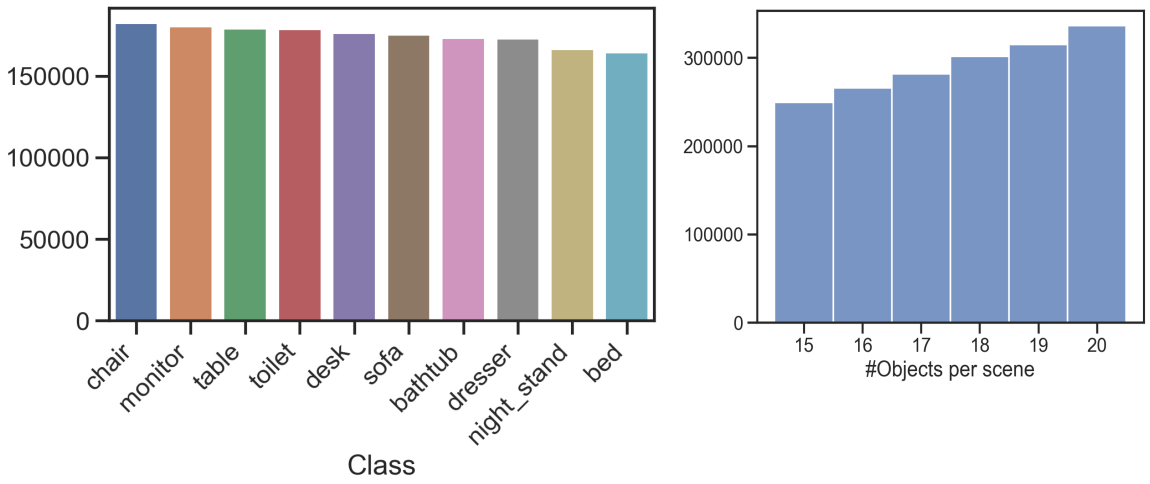


Figure 5.8: Histograms of the train set distributions for (a) Objects per class (total) and (b) Number of objects per scene.

Subset	test (v_t) \ train (v_r)	cam0	cam8	cam12	cam13	cam22
Other objs.	L0.cam0	**71.12**	29.09	29.61	14.28	14.88
	L2.cam8	24.63	**70.21**	**70.16**	28.14	28.54
	L2.cam12	25.14	69.09	70.05	27.73	28.29
	L3.cam13	12.18	31.26	31.46	**59.18**	58.72
	L3.cam22	12.11	30.10	30.59	58.39	**59.41**
Same objs.	L0.cam0	**80.55**	29.92	29.69	14.00	14.51
	L2.cam8	27.11	**77.90**	77.71	27.24	27.46
	L2.cam12	28.01	76.87	**77.97**	26.94	27.52
	L3.cam13	12.90	32.16	32.29	**65.87**	65.69
	L3.cam22	12.76	31.00	31.68	64.84	**66.09**

Table 5.2: IoU results for direct cross-view semantic transfer. Five models trained on 100% of the train set (100k scenes). The models were trained on reference view (v_r) data pairs and tested on target view (v_t) data.

5.4 Experimental baselines

We run two baseline experiments for the cross-view semantic transfer problem. These experiments are included to show that there is no simple solution to this task and it is indeed an open research problem. To conduct them we select 5 representative views from three distinct levels: L0.cam0 (zenithal), L2.cam8, L2.cam12, L3.cam.13 and L3.cam.22 (see Fig. 5.1). In both cases we use a U-Net [223] as our semantic segmentation model, with an Imagenet-pretrained ResNet50 backbone.

Experiment 1. Cross-view semantic transfer via direct testing. We train an independent model with each of the considered views and directly test them against every other camera's test sets, without any specific adaptation. Table 5.2 shows the results in terms of Intersection over Union (IoU): The diagonals correspond

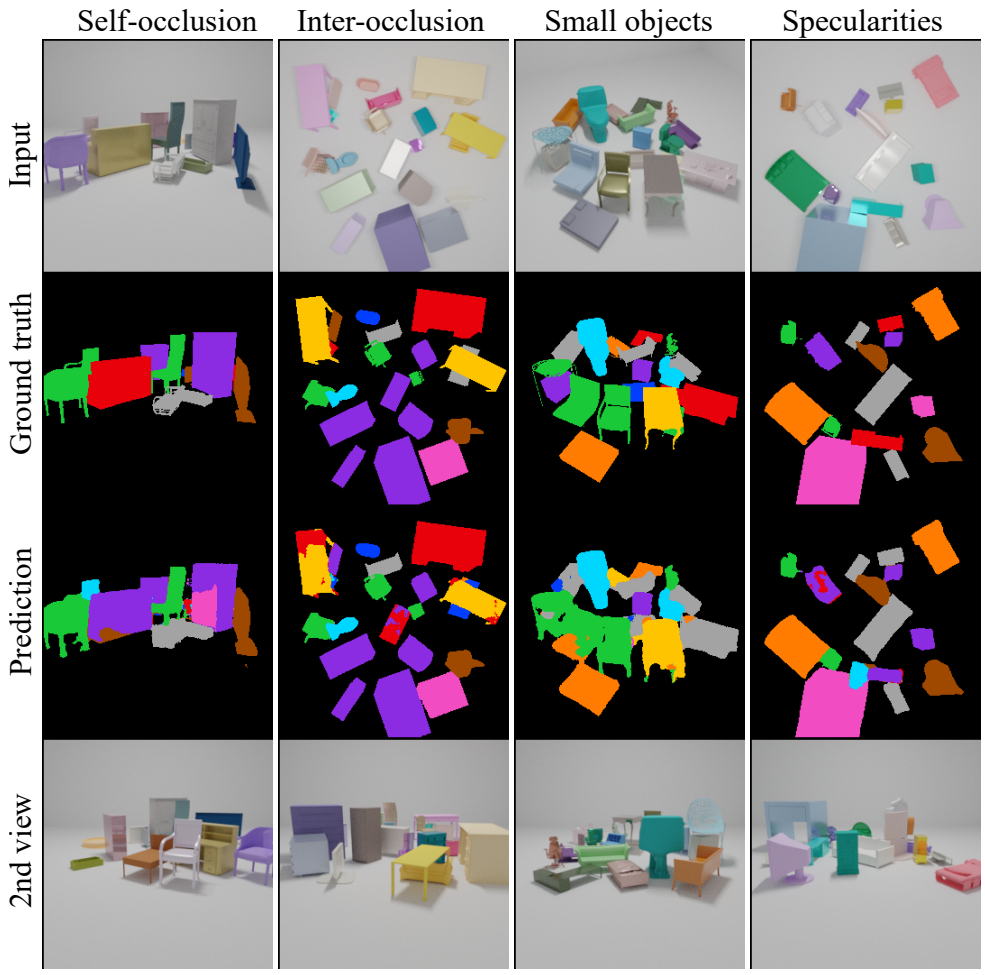


Figure 5.9: Failure cases from monocular models in the diagonals of Table 5.2. a) self-occlusion (golden object) b) inter-object occlusion (sofa under the yellow desk) c) small objects (light pink and dark green objects) d) ambiguity from specular inter-reflection (light blue object with reflections of the cyan one). Last row shows a second view that could help solve the ambiguity.

to standard fully supervised single-view setups. We see that these improve as we adopt a higher perspective of the scene. As one might expect, direct semantic transfer between cameras placed within the same level (e.g. L2.cam8/L2.cam12) yields a minimal performance drop, on account of the quasi-invariance of the learned representations to horizontal camera pose rotations (the objects were placed in the scene with a random rotation, hence the features observed from both views are similar, except for the non-circular symmetry of the placement area). However, the performance across views at distinct levels drops drastically, with the most distant levels yielding the highest differences. Note, finally, the foreseeable performance generalization gap between the OO and SO test subsets that favours the latter.

Fig. 5.9 shows some of the most common failure cases for monocular semantic segmentation models:

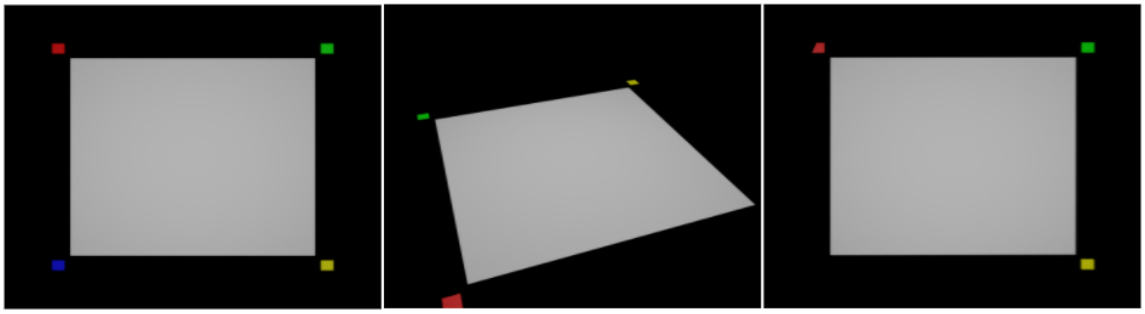


Figure 5.10: Computation of the ground truth homography induced by the $z = 0$ plane that maps cameras v_t to v_r ($H^{z=0}_{t\to r}$). Left: rectangle located at the $z = 0$ plane as viewed by the zenithal camera v_r =L0.cam0. Center: same rectangle as viewed by camera v_t =L2.cam8. We manually log the (x, y) coordinates of each of the vertices in both views (adyacent colored patches were used to ease their identification) and compute $H^{z=0}_{t\to r}$ using four point correspondences by least-squares minimization of the back-projection error. Right: result of using the $H^{z=0}_{t\to r}$ homography to reconstruct the view v_r from v_t.

v_r =L0.cam0→ v_t =L2.cam8		v_r =L2.cam8→ v_t =L0.cam0	
Other objs.	Same objs.	Other objs.	Same objs.
28.72	31.29	24.35	24.84

Table 5.3: IoU results for planar homography-based transfer.

(i) self-occlusions and (ii) partial inter-object occlusions that hide relevant features of the object (resulting in ambiguous geometry and appearances), (iii) distant/small objects and, less prominently, (iv) ambiguities induced by appearance variations (e.g. specularities). All these cases could benefit from the complementary information provided by the additional, significantly distinct perspectives of a multi-view setup. Nevertheless, the way of constructively fusing such multiple-view information sources in data-driven models without explicitly addressing a 3D representation of the scene is far from trivial, both in the multi-view and in the cross-view semantic transfer cases.

Experiment 2. Planar homography-based transfer. Another baseline to model such geometric relation between views in a cross-view semantic transfer scenario is that of a planar 3×3 homography. This model holds well for quasi-planar scenes or relatively distant objects [93]. In this experiment, we first compute the homography $H^{z=0}_{t\to r}$ induced by the $z = 0$ plane that maps cameras v_t (target view) to v_r (reference view) using four point correspondences (Fig. 5.10).

Then, in order to obtain a semantic map estimate from v_t given a model trained on v_r ($f_{v_r\to ss_r}$), we proceed as follows: (i) transform the v_t input to v_r via $H^{z=0}_{t\to r}$ (ii) feed this to $f_{v_r\to ss_r}$ so as to obtain a semantic map referenced to v_r (iii) transform this back to be referenced to v_t with the inverse homography $H^{z=0}_{r\to t} = (H^{z=0}_{t\to r})^{-1}$. We test this on two cameras at distinct levels: L0.cam0 and L2.cam8. The lack of a significant performance gain in the results (see Table 5.3) over the direct transfer baseline from Table 5.2 shows that, as expected, the planar homography fails to help for the general, wide-baseline case, in which a good estimate of pixel-wise depth information from every secondary view is needed for unambiguous matching. Fig. 5.11 shows the associated qualitative results for the transfer between v_r =L0.cam0 and

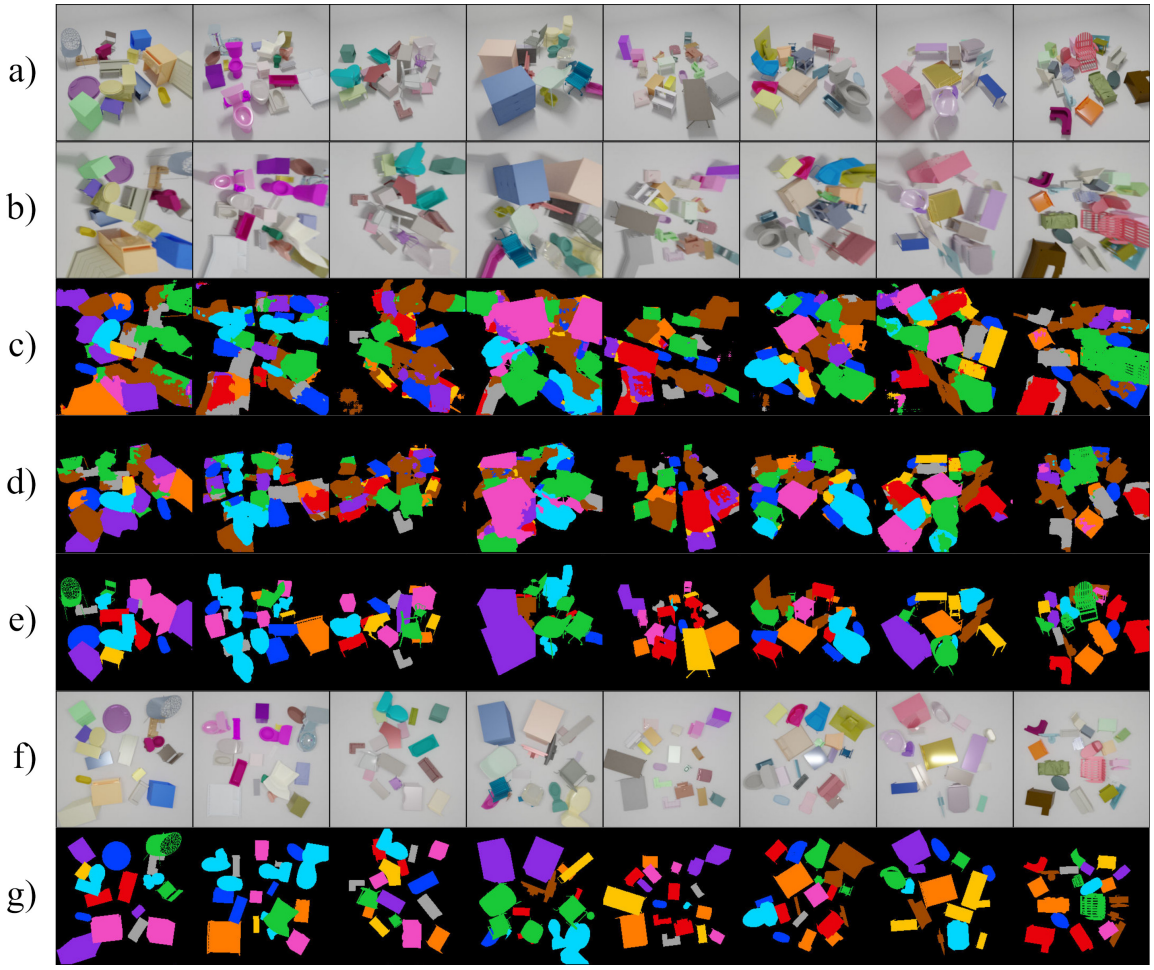


Figure 5.11: Qualitative intermediate and final results for the planar homography-based cross-view semantic transfer case (*i.e.* Experiment 2), for the v_r =L0.cam0→ v_t =L2.cam8 transfer (Table 5.3, left). Given a model, $f_{v_r \to ss_r}$, trained on (v_r, ss_r) pairs, we want to feed it with inputs from view v_t and obtain ss_t segmentation results referenced to v_t (ss_t). Each column represents a random sample scene. From top to bottom rows: a) v_t, input at inference time. b) Result of applying $H_{t\to r}^{z=0}$ to v_t to get a planar homography-based estimate of v_r. Note the significant differences with f). c) Predicted ss_r, result of feeding $f_{v_r \to ss_r}$ with the planar homography-based estimate of v_r. d) Predicted ss_t, result of transforming the predicted ss_r semantic map back to the reference of v_t using the inverse homography $H_{r\to t}^{z=0} = (H_{t\to r}^{z=0})^{-1}$. e) Ground truth for the task, ss_t. f) Ground truth v_r view of the scene, for reference and used for training of the $f_{v_r \to ss_r}$ model. g) Ground truth ss_r semantic map, for reference and used for training of the $f_{v_r \to ss_r}$ model.

v_t =L2.cam8 for eight random samples, also showing every intermediate result. Meanwhile, Fig. 5.12 depicts the same content —for the same samples— for the transfer from v_r =L2.cam8 to v_t =L0.cam0.

The failure of both experimental baselines, along with the fragility of photometric cues in wide baseline

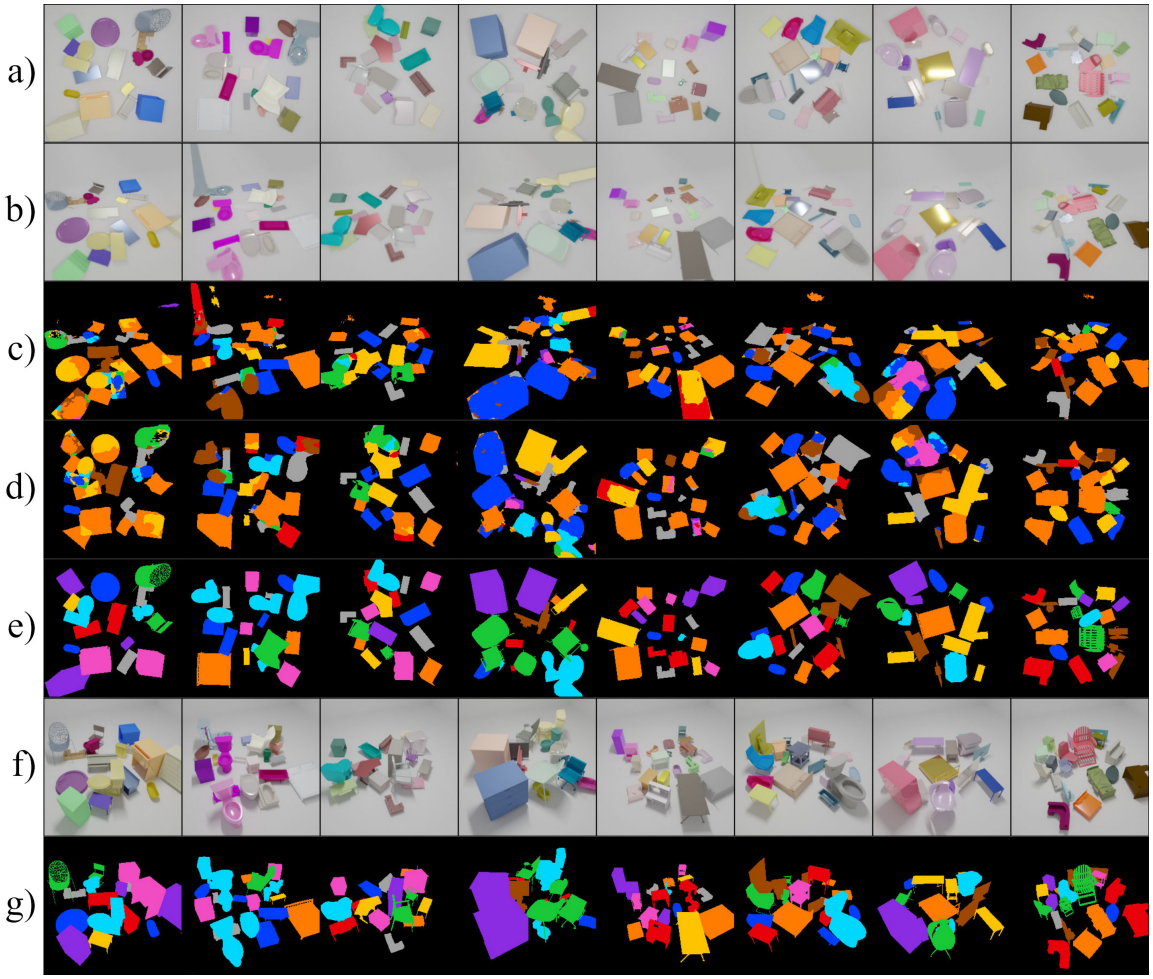


Figure 5.12: Qualitative intermediate and final results for the planar homography-based cross-view semantic transfer case (*i.e.* Experiment 2), for the v_r =L2.cam8→ v_t =L0.cam0 transfer (Table 5.3, right). See Fig. 5.11 for the legend for each row.

scenarios [282], suggests that exploiting the complementary information given by additional views of the scene in a data-driven multi-view learning setup or transferring the knowledge from trained models across views in unsupervised scenarios will require the development of new theoretical approaches.

5.5 Conclusion

We presented MVMO, a wide baseline multi-view synthetic dataset with semantic segmentation annotations that features a high object density and large amount of occlusions. We expect MVMO will propel research in multi-view semantic segmentation and cross-view semantic transfer and, likely through domain adaptation, address the current limitations of monocular setups in heavily-occluded real world scenes.

6 Zero-Pair Semi-Supervised Cross-View Semantic Segmentation

6.1 Introduction

We present the semi-supervised task of *zero-pair, cross-view semantic segmentation* (Fig. 6.1). Given a labeled dataset $\mathscr{D}_l$ of pairs of RGB and semantic segmentation annotations under a certain reference camera viewpoint v_r, we would like to change the camera pose to a new location and orientation and, without any further annotation, be able to keep producing semantic segmentation predictions under the reference point of view v_r, as well as under the new target viewpoint v_t. In addition to this, we assume that we have access to a second -unlabeled- cross-view dataset of RGB image pairs, $\mathscr{D}_u$, comprising synchronized pairs of reference (v_r) and target (v_t) projections of scenes not contained in the first -labeled- dataset, *i.e*. both datasets are disjoint with respect to the scenes they represent.

However, at no moment do we have available any pair of (RGB, dense semantic label) images where the input RGB image is taken under the target viewpoint v_t, regardless of the reference view of the labels. We thus characterize the problem as *zero-pair* [242, 271]. It can also be thought of as a *semi-supervised* learning task, since, even if we can learn an RGB-semantic mapping in the reference view, the problem requires of unsupervised semantic knowledge transfer from source to target viewpoint. Such transfer of information across views is notably challenging even in supervised settings, particularly so whenever the object density is high and when the camera pose change falls within the category of *wide baseline* scenarios, *i.e*. those where the distance between cameras (and thus, the resulting disparities) is significant compared to the distance to the scene objects.

The described scenario naturally fits in industrial contexts, where many in-line production systems rely on semantic segmentation for the inspection of goods carried on conveyor belts, often with remarkable upfront annotation costs. It is often the case that hardware updates need to be carried out on the machine that affect the original placement of the sensor, requiring a new, distinct location for it. In such cases, it may be infeasible -or extremely costly- to capture and re-annotate new scenes from the new perspective. Furthermore, such relocation may well affect downstream vision or manipulation systems, which still expect predicted coordinates referenced to the source camera pose. The problem thus calls for a flexible approach that enables cheap camera relocations without the need for new annotations. In this regard, it would be comparatively cheap to temporarily use a second camera to capture a number of scenes from both perspectives, without any annotation.

These setups cannot be approached by a naive application of models trained on the source domain nor with simple geometric tools as planar homography-based warping (Fig. 6.2). In fact, a successful model should leverage the complementary information provided by cross-view image pairs to fill, at inference time, the occluded parts with the learnt geometric and appearance priors. This also constitutes a potential building block for multi-view semantic segmentation scenarios, where diverse target views contribute to enhanced reference view prediction, bypassing the occlusion-caused ambiguities inherent to monocular systems.

In this chapter, we make the following contributions: (i) we introduce the semi-supervised task of zero-pair cross-view semantic segmentation to tackle cheap fixed camera relocations on established semantic labeling systems. (ii) We present ZPCVNet: a model comprising several deep convolutional neural encoder and decoder modules and a cross-view transformation module to achieve latent space alignment. This enables

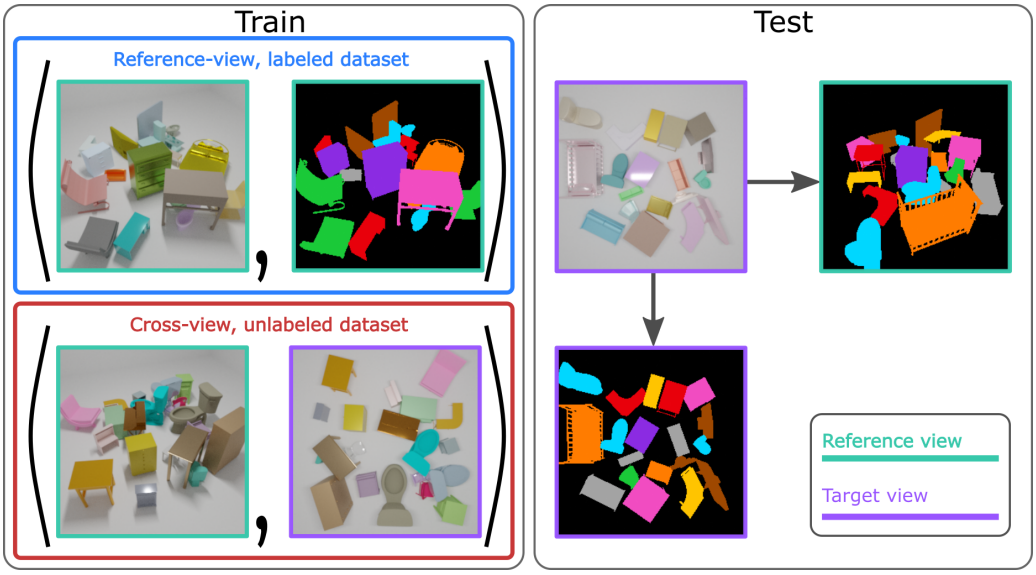


Figure 6.1: The task of zero-pair, semi-supervised cross-view semantic segmentation: given a reference-view, labeled segmentation dataset and an unlabeled cross-view dataset, and given a test RGB input on the target view's frame, predict the semantic labels referenced to i) the source/reference view and ii) the target view.

producing semantic maps referenced to both source and target viewpoints simultaneously from the new camera location, without the requirement of any additional labeling. (iii) Results show that our method can obtain semantic segmentation results for viewpoints that lack any semantic segmentation ground truth.

6.2 Related work

Our work relates to a number of previous approaches across various research fields:

Multi/cross-view computer vision. Several classic computer vision areas have approached their tasks from a wide-baseline multi or cross-view perspective: [221] addressed the detection of pedestrians, cars and buses from 6 calibrated cameras in an attempt to improve the single-view performance. Human pose estimation has been a particularly fertile field, propelled by the existence of large scale, wide baseline multi-view datasets such as Human3.6M [107], ITOP [92] or 3DPeople [201]. [92] learnt viewpoint-invariant features to address pose estimation from diverse viewpoints over depth inputs. [118] relied on 2D part detectors in a simultaneous multi-view RGB input scenario, while [203, 281] exploited multiple-view geometry-related priors and [293] incorporated additional time consistency cues to propagate the annotations across views.

The area of novel view synthesis has also emerged with vigour, impulsed by recent advances in generative modeling [58, 102, 127, 254, 304], and the use of a diverse set of alternatives for scene representation and rendering of novel views that have enabled the use of 2D projections for supervision. Generative Query Networks [64], 3D point clouds [274] and Neural Radiance Fields [166] are representative examples.

Multi/cross-view semantic segmentation. Research on transferring or fusing dense semantic knowledge across views has been hindered until recently by the lack of large scale, photo-realistic, annotated datasets with wide baselines and high object densities. This limited the existing works to (i) cases with a modest

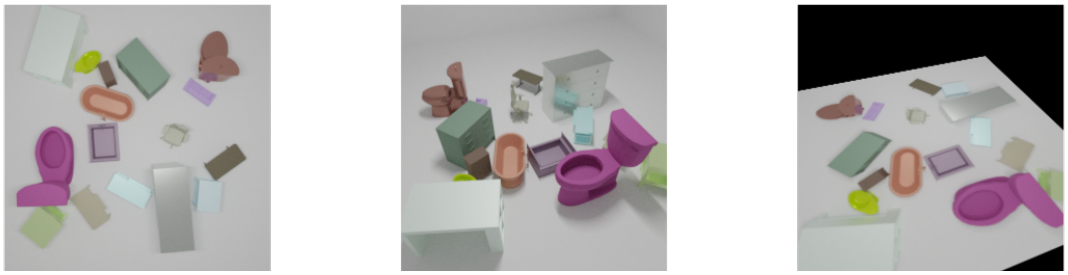


Figure 6.2: Without depth information, planar homography is the only geometrical tool available for cross-view pixel to pixel mapping. However, it fails for wide baseline or non-planar scenes: i) target view ii) reference view iii) ground truth planar homography-based target→reference cross-view transform.

magnitude of the viewpoint change [45] (ii) approaches for 3D model segmentation [49, 90, 125] and/or for segmentation from RGBD or 3D inputs [90, 160, 279] (iii) the domain of aerial to/from ground view transfer [288], in which segmentation was sometimes also addressed as an auxiliary task for mutual benefit with the main goal of novel view synthesis [211]. Such multi-task approach is also often found naturally together with 3D structure reconstruction problems [21, 90, 135]. Besides, only a few of the works in the field produce their semantic predictions in a reference frame other than that of the input RGB image: [183] yield top-view semantic maps from the fusion of up to 8 cameras in a camera-centric, narrow baseline setup, and [288] generates rough ground-view semantic maps from aerial imagery. The first problem, *i.e.* the lack of an adequate dataset, was addressed in Chaper 5 with the release of the wide-baseline, occlusion-heavy MVMO dataset. Meanwhile, this chapter addresses the latter by proposing ZPCVNet as the first model able to obtain predictions in both references.

Multi-view geometry-aware approaches. A handful of the aforementioned works complement data-driven approaches by posing constraints on the network topology or cost function inherited from what we already know from classic multiple-view geometry [93], such as various representations of the epipolar constraints: [95, 203, 281] for body and hand pose estimation, [240] for depth, semantic label and view synthesis, [200] for depth and egomotion prediction or [282] for single class co-segmentation. Unlike our approach, while these inductive biases effectively restrict the solution space, they usually require the calibration of the cameras.

Mix and Match Networks. Our work is most closely related to *Mix and Match Networks* [270, 271], and partially builds upon their idea of aligning the latent representations of multiple encoders and decoders from/to different modalities to achieve zero-pair cross-modal image-to-image translations between unseen modality pairs (*e.g.* depth to semantic labels from RGB to semantic labels and RGB to depth). Functionally, though, their work is limited to domain shifts consisting of perfectly pixel-paired modality changes, while ours addresses simultaneous modality (RGB to semantic labels) and viewpoint (*i.e.* spatial distribution) shifts.

6.3 Zero-pair cross-view semantic segmentation

We address the problem of *cross-view semantic transfer*. In this task, dense ground truth semantic annotations $y_r \in \{1...C\}$ for C classes are available for RGB images x_r acquired from a reference or source view v_r. We consider the scenario where we move the camera location to a new -target- viewpoint v_t, for which we have no annotations but still want to predict the semantic segmentation in either reference or target viewpoint.

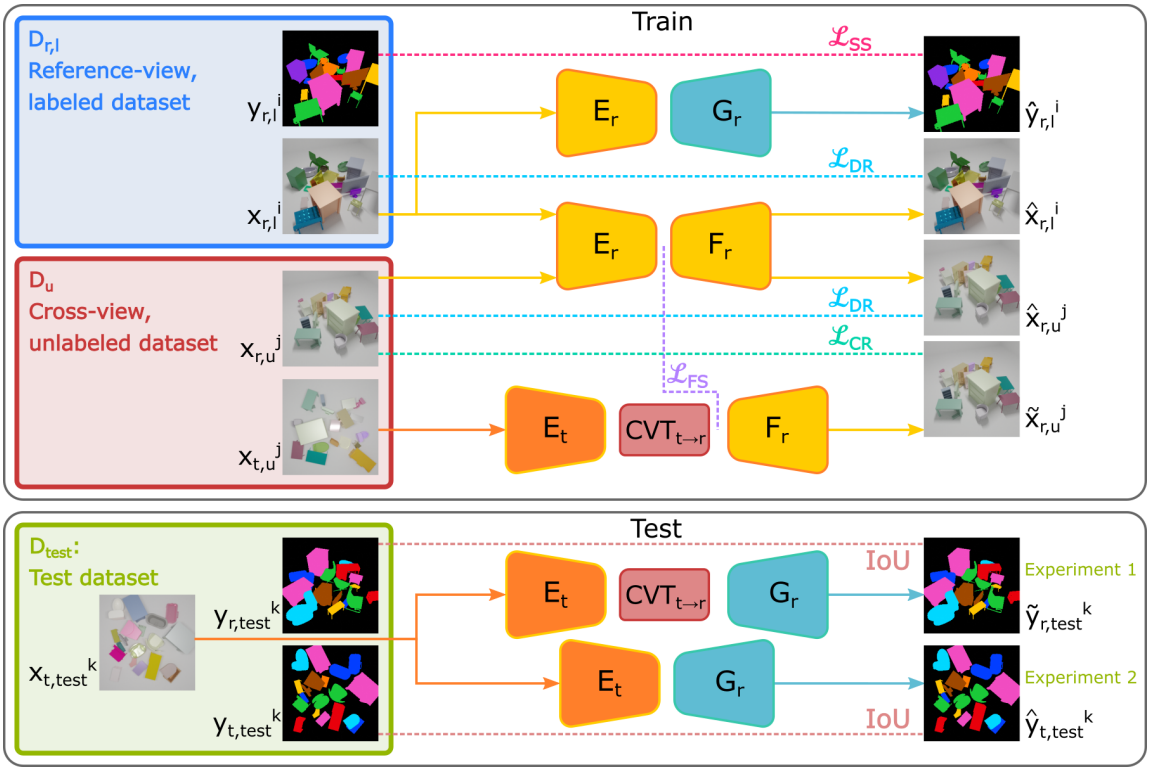


Figure 6.3: The proposed zero-pair, semi-supervised cross view semantic segmentation model, including the Cross View Transformer. Train and test stages are shown, with the test setup yielding predictions referenced to both v_r and v_t. Equal modules share weights.

Direct application of a model f_r learned in a fully supervised way on the source view is expected to achieve sub-optimal results. Another solution would be a fully supervised approach, which implies capturing and annotating a whole new dataset referenced to the new point of view $(x_{t,l}^i, y_{t,l}^i) \forall i \in \mathscr{D}_{t,l}$.

As opposed to this, we propose a setup in which, in addition to the fully labeled reference view dataset $\mathscr{D}_{r,l} = (x_{r,l}^i, y_{r,l}^i)\ \forall i = 1...M$, we collect an additional cross-view image set of unlabeled but synchronized RGB image pairs corresponding to distinct views of the same scene, $\mathscr{D}_u = (x_{r,u}^j, x_{t,u}^j)\ \forall j = 1...N$. Note that these datasets are disjoint; therefore there exists no matching scene contained in both labeled and unlabeled sets; *i.e.* $x_{r,l}^i \neq x_{r,u}^j$ for any possible i, j pair. This is a realistic setting: in practice, in industrial scenarios the need to place the camera in a different pose often occurs due to redesign of the setup. Under these circumstances, a naive approach would require to start the annotation altogether. Instead, our approach is able to exploit the annotations already collected for dataset $\mathscr{D}_{r,l}$, together with paired but unlabeled images of new scenes that should be relatively cheap to obtain when compared to the cost of gathering dense semantic annotations.

In this context, *cross-view semantic segmentation* consists of, given an unseen input RGB image x_t taken under the target viewpoint v_t, predict its corresponding semantic labels, either under the reference or target viewpoint $(\tilde{y}_r, \hat{y}_t)$, respectively (a note on notation: we use tilde˜for cross-view predictions, and hatˆotherwise). Initially, we will take $\tilde{y}_r$ as our primary target task. Section 6.4.2 shows how we can obtain

$\hat{y}_t$ predictions under the new target view.

6.3.1 Zero-pair cross-view: the vanilla model

Our model (see Fig. 6.3) contains a set of convolutional encoders and decoders arranged so that their latent representations get aligned after training, thus enabling the direct coupling of unpaired modules to map previously unseen domains. The first major block ($E_r - G_r$) is a **semantic segmentation net**, composed of a convolutional encoder E_r that takes as input the RGB images from the source view v_r and generates a compact, bottleneck representation of them, and a decoder G_r that takes such latent features and yields dense semantic labels in the same reference viewpoint:

$$\hat{y}^i_{r,l} = G_r(E_r(x^i_{r,l})) \tag{6.1}$$

This block on its own corresponds to the simple monocular fully supervised baseline, and it is supervised with a pixel-wise categorical cross-entropy loss $\mathscr{L}_{SS}$ (see section 6.3.4). To exploit the information of the unlabeled image pairs, and inspired by [270], we place two auxiliary blocks:

A **reference-view autoencoder** ($E_r - F_r$), that aims at reconstructing the input view. This component will operate on inputs sampled from both the labeled and unlabeled datasets, as it requires no annotation:

$$\hat{x}^i_{r,l} = F_r(E_r(x^i_{r,l})) \tag{6.2}$$

$$\hat{x}^j_{r,u} = F_r(E_r(x^j_{r,u})) \tag{6.3}$$

where F_r is a decoder generating the reference view. This component is supervised via a reconstruction loss, $\mathscr{L}_{DR}$.

A **cross-view encoder-decoder** block ($E_t - F_r$) that takes RGB images from the target view v_t and transforms them by generating, at the output, an RGB representation of the same scene as seen from the reference view v_r:

$$\tilde{x}^j_{r,u} = F_r(E_t(x^j_{t,u})) \tag{6.4}$$

We attach a cross-view reconstruction loss $\mathscr{L}_{CR}$ between predicted and ground truth reference views to supervise this.

Although introduced as separate components here for clarity, note that the reference view encoder E_r used in the fully supervised semantic segmenter $E_r - G_r$ and autoencoder $E_r - F_r$ blocks is sharing weights in both tasks. This is equally true for the F_r reference view RGB decoder, used in both ($E_r - F_r$) and ($E_t - F_r$) blocks.

The rationale behind these three encoder-decoder blocks is to seek the alignment of the latent representations at the bottlenecks of the aforementioned encoder-decoder pairs during the training process. By doing so, at inference time we could take the trained target view encoder E_t and reference view semantic labeling decoder G_r out of the training setup, directly connect them and obtain semantic segmentation predictions referenced to the *source view* for RGB inputs captured from the *target view*:

$$\tilde{y}^k_{r,test} = G_r(E_t(x^k_{t,test})) \tag{6.5}$$

where $x^k_{t,test}$ is sampled from the test set $\mathscr{D}_{t\to r,test} = (x^k_{t,test}, y^k_{r,test})\ \forall k = 1...P$. Note that this cross-view semantic transfer is unsupervised, since during training we have no access to the ground truth referenced to

v_t.

Both reference and target view encoders (E_r and E_t) should generate similar latent representations when fed with RGB captures depicting the same scene from both viewpoints, and the reference view decoder F_r should yield consistent reconstructions from them both. In this setup, the latent space may be interpreted as a view-free representation of the scene, containing representative information for both RGB and segmentation reconstructions on the reference view. To further enforce the latent space alignment, we pose an additional feature similarity loss $\mathscr{L}_{FS}$ (see section 6.3.4) between features obtained from both encoders, E_r and E_t. In practice, however, such equilibrium showed to be hardly achievable (see section 6.4). As a consequence, two additional components are discussed in sections 6.3.2 and 6.3.3.

6.3.2 CVT: Cross-View Transformer

In the vanilla model described above, the latent space is conceived as a viewpoint-free one. In consequence, every encoder-decoder block must allocate part of its capacity to disentangle and discard (encoders) or blend (decoders) the viewpoint-dependent information. To discharge the different modules from such duty, we instead place an explicit Cross-View Transformer* ($CVT_{t\to r}$) module at the output of the target view encoder E_t, so that the cross-view encoder-decoder block is now defined as $E_t - CVT_{t\to r} - F_r$.

The Cross-View Transformer module is defined as a function $CVT_{t\to r} : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{h\times w\times c}$ with:

$$CVT_{t\to r} = f^{-1}(W \cdot f(x)) \tag{6.6}$$

where $f : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{hw\times c}$ is a flattening operation on the spatial dimensions of the latent features, $f^{-1} : \mathbb{R}^{hw\times c} \to \mathbb{R}^{h\times w\times c}$ is the reciprocal operation of recovery of the original shape, and $W \in \mathbb{R}^{hw\times hw}$ corresponds to the weights of a single linear layer with dense connections among each of the spatial locations of the features. The intuition of the CVT is that of a layer operating only on the spatial dimensions, thus applying in parallel the same learnt weighted location mixture to every activation map from the latent representation. This allows to spatially reorganize the latent representation to align with the convolutional decoder. An analogous approach was shown to help transfer features across views in [183], where a bank of multi-layer perceptrons was applied over embeddings obtained from depth, RGB and even semantic label images. However, this was applied on a fully supervised setup, while our aforementioned constraint of $\mathscr{D}_{r,l}$ and $\mathscr{D}_u$ being disjoint sets prevents us from being able to perform a straightforward cross-view supervision on (x_t, y_r) pairs.

We can now, therefore, update eq. (6.4) and eq. (6.5):

$$\tilde{x}^j_{r,u} = F_r(CVT_{t\to r}(E_t(x^j_{t,u}))) \tag{6.7}$$

$$\tilde{y}^k_{r,test} = G_r(CVT_{t\to r}(E_t(x^k_{t,test}))) \tag{6.8}$$

while the feature similarity loss $\mathscr{L}_{FS}$ is now applied at the output of the CVT. This enforces the latent representation of the overall model to be referenced to v_r.

We found that, in the absence of geometrically calibrated cameras and depth information from any of the viewpoints (which would allow us to exploit the epipolar constraint [95] or get a pixel-wise projection of the segmentation onto the other view [45], respectively) the data-driven CVT module successfully models a cross-view transfer of the features by learning to predict the corresponding 2D location projection of every

*The usage of the term *Transformer* throughout this chapter differs from its most widespread meaning accross the computer vision community as of 2022, *i.e.* we do not refer to *Transfomers* as in [262].

feature vector (of length c) from the target view v_t in the reference view v_r. Notably, this approach requires no explicit 3D reconstruction of the scene at hand, and we gain interpretability, as the geometric transformation is encapsulated in a single component. Furthermore, factoring out the cross-view transformation in a specific module provides an additional benefit: we can now obtain predictions referenced not only to v_r, but to v_t as well with no need for additional training, by just removing the CVT component. We show this as *Experiment 2* in section 6.4.2 and Fig 6.3.

6.3.3 Pseudolabels

With the presented components, the feature alignment between E_t and G_r, required at inference time, is achieved indirectly via the use of the reference view encoder E_r and decoder F_r. Such setup fails to extract and leverage the shared information between the unpaired domains, *i.e.* between the RGB inputs from the target view and the segmentation outputs from the source view. Following [270], we instead explicitly enforce such representation alignment by introducing a new block at train time: a cross-view semantic segmentation one ($E_t - CVT_{t\to r} - G_r$), which takes RGB inputs from the target view (thus from the unlabeled set) and yields semantic label predictions on the source view reference. This reproduces the inference time scenario:

$$\tilde{y}^j_{r,u} = G_r(CVT_{t\to r}(E_t(x^j_{t,u}))) \tag{6.9}$$

Given the lack of ground truth for the unlabeled dataset, the training of these modules is supervised via the $\hat{y}^j_{r,u}$ predictions obtained by feeding the $E_r - G_r$ block with the reference view input taken from the unlabeled dataset:

$$\hat{y}^j_{r,u} = G_r(E_r(x^j_{r,u})) \tag{6.10}$$

These can be thought of as the pseudo-labels often found in the field of domain adaptation [306], which will be used as ground truth when posing a pixel-wise categorical cross-entropy loss $\mathscr{L}_{PL}(\tilde{y}^j_{r,u}, \hat{y}^j_{r,u})$ between both predictions. This allows us to leverage the correlation between the features of both domains. Note again that the components of this new block are sharing their weights with those named equally in other blocks. We name the resulting full model *ZPCVNet* (Zero-Pair Cross-View Net).

6.3.4 Training process and loss functions

The model training process is as follows: At each iteration, two equal-sized mini-batches are drawn, each one from one of the $\mathscr{D}_{r,l}$ and $\mathscr{D}_u$ datasets: $(x^i_{r,l}, y^i_{r,l})$ and $(x^j_{r,u}, x^j_{t,u})$ $\forall i, j = 1...B$, respectively. After feeding the different encoders with the $(x^i_{r,l}, x^j_{r,u}, x^j_{t,u})$ RGB inputs, the predictions obtained from equations (6.1), (6.2), (6.3), (6.7), (6.9) and (6.10) are assessed by means of the following compound loss:

$$\mathscr{L} = \lambda_{SS}\mathscr{L}_{SS} + \lambda_{DR}\mathscr{L}_{DR} + \lambda_{CR}\mathscr{L}_{CR} + \lambda_{FS}\mathscr{L}_{FS} + \lambda_{PL}\mathscr{L}_{PL} \tag{6.11}$$

where $\lambda_{SS}, \lambda_{DR}, \lambda_{CR}, \lambda_{FS}$ and λ_{PL} are the hyper-parameters to balance the importance of each term. This is a multi-task objective comprising:

- A fully-supervised semantic segmentation loss (pixel-wise class-weighted cross-entropy), which conducts the pixel classification using the paired reference RGB and the semantic segmentation:

$$\mathscr{L}_{SS} = \mathscr{CE}(\hat{y}^i_{r,l}, y^i_{r,l}) = \mathscr{CE}(G_r(E_r(x^i_{r,l})), y^i_{r,l}) \tag{6.12}$$

- An ℓ_2 direct-reconstruction loss on the auto-encoder predictions from both mini-batches, which tries to reconstruct the input RGB for both the label and unlabeled data:

$$\mathscr{L}_{DR} = \frac{1}{2}\left\|\hat{x}^j_{r,l} - x^j_{r,l}\right\|_2 + \frac{1}{2}\left\|\hat{x}^j_{r,u} - x^j_{r,u}\right\|_2 \tag{6.13}$$

- An ℓ_1 cross-view reconstruction loss to perform cross-view prediction from the unlabeled target view RGB:

$$\mathscr{L}_{CR} = \left\|\tilde{x}^j_{r,u} - x^j_{r,u}\right\|_1 = \left\|F_r(CVT_{t\to r}(E_t(x^j_{t,u}))) - x^j_{r,u}\right\|_1 \tag{6.14}$$

- An ℓ_2 feature similarity loss to align the latent representations at the output of the E_r and $E_t - CVT_{t\to r}$:

$$\mathscr{L}_{FS} = \left\|CVT_{t\to r}(E_t(x^j_{t,u})) - E_r(x^j_{r,u})\right\|_2 \tag{6.15}$$

- A pixel-wise class-weighted categorical cross-entropy loss to conduct the pixel classification by using the pseudo-labels for the cross-view semantic predictions from the unlabeled target view input:

$$\mathscr{L}_{PL} = \mathscr{CE}(\tilde{y}^j_{r,u}, \hat{y}^j_{r,u}) = \mathscr{CE}(G_r(CVT_{t\to r}(E_t(x^j_{t,u}))), G_r(E_r(x^j_{r,u}))) \tag{6.16}$$

Although we are able to train our model end-to-end, we found it easier, in practice, to run a two-stage training: we first pretrain the $E_r - G_r$ block alone in a fully supervised fashion. Then, we add the remaining blocks, so that (i) the E_r encoders are initialized from the pretrained weights and unfrozen, (ii) the G_r semantic decoders are initialized from the pretrained weights but kept frozen, and (iii) the F_r decoder and the $CVT_{t\to r}$ are trained from scratch.

6.4 Experimental validation

We conduct two experiments on the MVMO (Multi-View, Multi-Object) dataset (see Chapter 5) to validate our approach, presented in sections 6.4.1 and 6.4.2 for cross-view semantic predictions $\tilde{y}_r$, $\hat{y}_t$, referenced to v_r and v_t, respectively. MVMO is partitioned in one train set (100,000 scenes, each with 25 views), two validation sets and two test sets (4,000 scenes each). In our experiments, we split the train set in two disjoint sets of 50,000 scenes, corresponding to the $\mathscr{D}_{r,l}$ and $\mathscr{D}_u$ datasets, and we evaluate on the *other_objects* (OO) test set. We run our tests on 64×64 inputs focusing on two of the 25 cameras: *L2.cam8* as our reference view v_r, and *L0.cam0* as our target view v_t. The latter provides a perfectly orthogonal (zenithal) view of the scene, while *L2.cam8* captures an oblique perspective (see Fig. 6.4 and Fig. 6.5). The relative pose change involves significant translation and rotation, and can be considered a wide baseline setup.

Network architecture. We design our network topologies taking inspiration from [83]. Both encoders (E_r, E_t) share the same architecture. Their design, for the input size of 64×64 employed in our experiments, is as follows: $C(3,64) - LR - C(64,128) - BN - LR - C(128,256) - BN$, where (i) C is a 2D convolutional layer with the indicated input and output channels, 4×4 kernel size and stride of 2, (ii) LR is a *Leaky ReLU* with a negative slope of 0.2, and (iii) BN is a Batch-Normalization layer with $\epsilon = 10^{-5}$ and momentum=0.1 This architecture leads to a spatial bottleneck feature size of 8×8 and 256 channels. The output of E_t is then forwarded through the CVT module, which is a fully connected layer ($64 \times 64 + 64 = 4160$ parameters). The decoders (F_r, G_r) differ only in the number of output channels (3 and 11, respectively). Their topology can

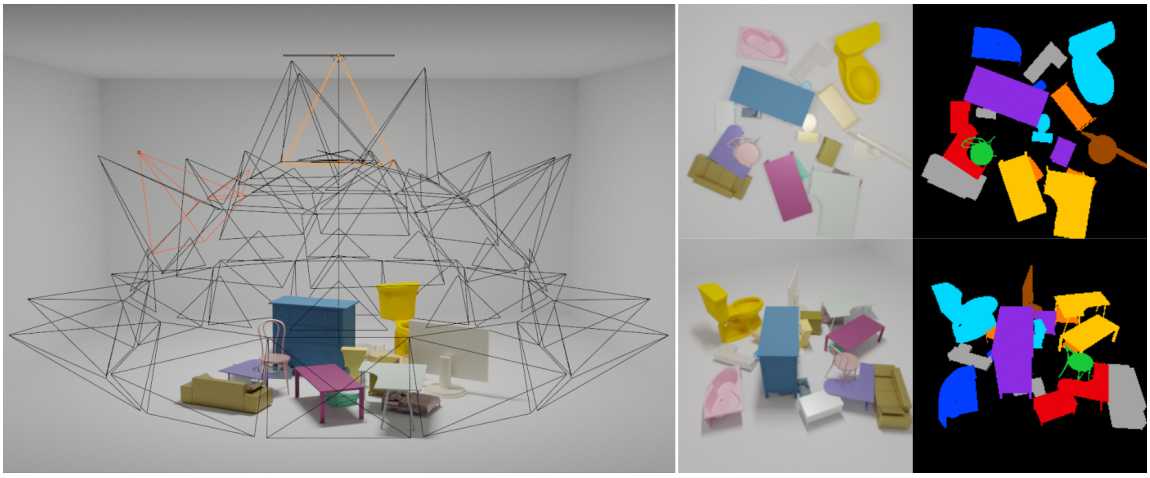
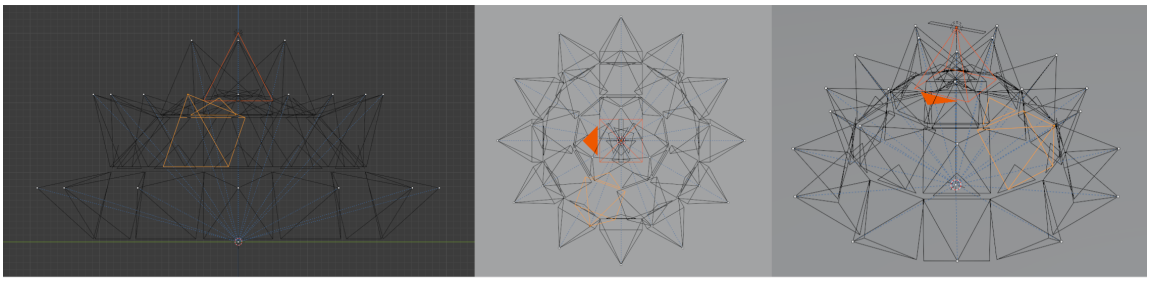


Figure 6.4: (left) sample 3D scene from the MVMO dataset and location of the 25 available cameras. The two camera locations used in our experiments are highlighted. (right) rendered views (256×256) and ground truth for the selected cameras: v_t =L0.cam0 (top) and v_r =L2.cam8 (bottom).



(a) (b) (c)

Figure 6.5: Three additional views of the location of the two cameras employed in our experiments (highlighted) in the surface of the upper hemisphere of a sphere of radius $r = 3m$: (a) Frontal view, orthogonal projection, showing the location of L0.cam0 and L2.cam8 in levels L0 and L2, respectively. (b) Zenithal view, orthogonal projection. (c) Oblique view, perspective projection.

be summarized as: $R - TC(256,128) - BN - D - R - TC(128,64) - BN - D - R - TC(64,3/11) - BN$, where (i) R is a regular *ReLU* (ii) TC represents a 2D transposed convolution with the indicated input and output channels, a 4×4 kernel size and a stride of 2 (iii) BN is a Batch-Normalization layer with $\epsilon = 10^{-5}$ and momentum=0.1 (iv) D is a *Dropout* layer with a 0.5 dropout rate.

Training details. For each experiment run, we kept the model that optimized the Intersection over Union (IoU) on the validation subset, and conducted a hyperparameter search over these, then reporting over MVMO's OO test subset. We used a batch size of 128 for all our reported results, where 64 images were sampled from each of the datasets. The experiments run for a maximum of 500 epochs, using a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 10^{-3}, a weight decay of 10^{-4} and a momentum of 0.9. The weighting factors found by cross-validation are: $\lambda_{SS} = 10, \lambda_{DR} = 1, \lambda_{CR} = 1, \lambda_{FS} = 10$ and $\lambda_{PL} = 100$.

Method	Trained on	bathtub	bed	chair	desk	dresser	monitor	nightstand	sofa	table	toilet	noBG	Avg.
FSCV	$x^i_{t,l}, y^i_{r,l} \sim \mathscr{D}^+_{r,l}$	04.81	03.54	04.43	00.00	00.00	00.00	05.64	00.00	04.43	00.00	02.29	07.45
CV$_{test}\upsilon_r$	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$	02.53	03.28	00.82	01.14	01.27	02.42	00.59	02.92	01.56	00.55	01.71	07.04
CV$_{test}\upsilon_t$	$x^i_{t,l}, y^i_{t,l} \sim \mathscr{D}^+_{r,l}$	02.22	02.65	02.82	03.11	03.28	02.58	03.58	02.47	02.08	02.81	02.76	07.81
Homography	$x^i_{t,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$	06.25	09.48	01.22	03.14	01.68	04.59	01.03	07.76	03.79	01.72	04.06	10.65
Mix&Match [270]	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$ $x^j_{r,u}, x^j_{t,u} \sim \mathscr{D}_u$	02.04	03.28	03.47	04.32	02.46	03.84	01.63	03.00	02.57	03.95	3.06	08.33
ZPCVNet (ours)	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$ $x^j_{r,u}, x^j_{t,u} \sim \mathscr{D}_u$	**17.73**	**21.47**	**14.33**	**10.81**	**18.52**	**22.82**	**22.73**	**12.81**	**14.29**	**20.42**	**17.59**	**23.10**

Table 6.1: IoU for experiment 1 on MVMO's OO test set: cross-view semantic transfer with RGB test set inputs captured from target view υ_t and predictions referenced to source view υ_r. **noBG**: Average of all classes but background.

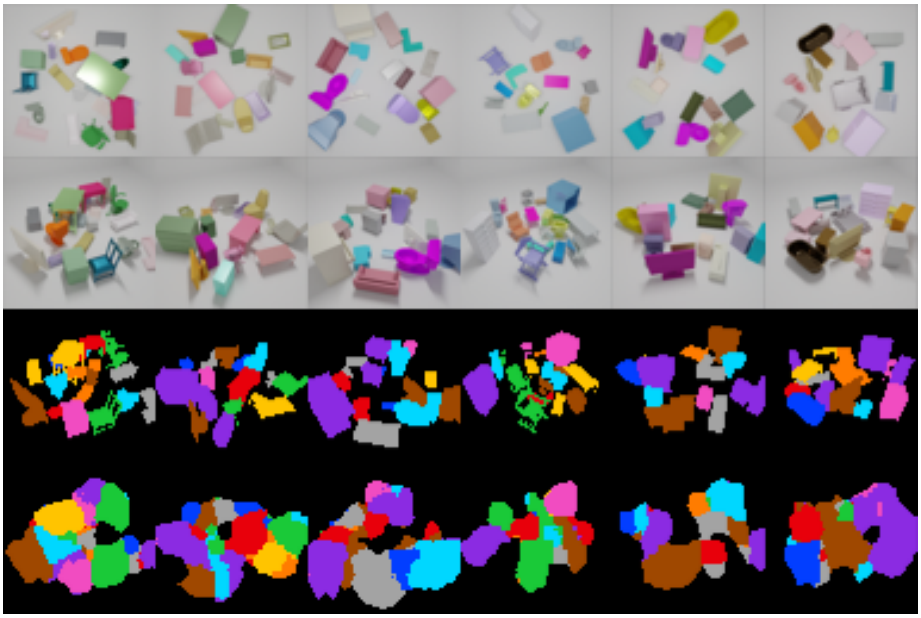


Figure 6.6: Experiment 1 results for ZPCVNet. Rows: i) υ_t input ii) corresponding υ_r iii) ground truth in υ_r iv) ZPCVNet prediction (ours) in υ_r.

6.4.1 Experiment 1: cross view with output in υ_r

Our first cross-view semantic transfer experiment corresponds to the case where we place the new camera on the υ_t pose but we wish to obtain the prediction of our model still referenced to the source view, υ_r. All the testing is therefore run on the $(x^k_{t,test}, y^k_{r,test})$ pairs of the $\mathscr{D}_{t\rightarrow r,test}$ test set.

Baselines. In order to validate the suitability of our approach, we consider the following baselines: (i) *FSCV* (Fully-Supervised Cross-View): this represents the supervised solution for the task. We use the $E_t - G_r$ block to obtain a directly supervised cross-view model, in charge of learning both the semantic labels and the cross-view knowledge transfer, which has been shown to be a difficult task [183]. We train this on $(x^i_{t,l}, y^i_{r,l})$ pairs drawn from the $\mathscr{D}^+_{r,l}$ dataset, an extension of $\mathscr{D}_{r,l}$ including the target view and ground truth of the

same scenes but under v_t: $x^i_{t,l}, y^i_{t,l} \sim \mathscr{D}^+_{r,l}$. (ii) $CV_{test}\, v_r$: we run a fully supervised training (using only the $E_r - G_r$ block with $\mathscr{L}_{SS}$) on the 50,000 $(x^i_{r,l}, y^i_{r,l})$ pairs of the $\mathscr{D}_{r,l}$ dataset, *i.e.* training and testing are carried out on the reference view. We then directly evaluate the learnt model on $(x^k_{t,test}, y^k_{r,test})$ pairs, without any adaptation. This corresponds to a lower bound for our setup, *i.e.* the system's performance evaluated on the original viewpoint's reference, should we just move the camera and do nothing (we would expect this to fail). (iii) $CV_{test}\, v_t$: we repeat this for the target view, training on $(x^i_{t,l}, y^i_{t,l})$ pairs from $\mathscr{D}^+_{r,l}$ and evaluating on $(x^k_{t,test}, y^k_{r,test})$. (iv) *Homography*: a 4×4 homography can serve as a deterministic mapping between two 2D scene projections. However, in order to exploit it, we would require having per-pixel depth information from the source of the projection. In the absence of such dense depth values, a planar 3×3 homography can work as a baseline tool to model the relation between views. Unfortunately, such model holds well only for quasi-planar scenes or relatively distant objects [93], neither of them being the case. Nevertheless, in this experiment we compute the homography induced by the $z = 0$ plane that maps cameras v_t to v_r ($H^{z=0}_{t\to r}$) using four point correspondences [215] (we could, alternatively, learn the 9 parameters of the matrix as part of the model from $(x^j_{r,u}, x^j_{t,u})$ pairs and a photometric loss, but pre-computing it provides an upper bound to this homography-based approach). We then compute our prediction as: $\tilde{y}^k_{r,test} = G_r(E_r(H^{z=0}_{t\to r}(x_t)))$, where the $E_r - G_r$ block can be pretrained in advance. (v) *Mix&Match Networks* [270]: We implement and run a Mix&Match variant by extending the vanilla ZPCVNet with bidirectional transforms and pseudolabels.

Results. Table 6.1 shows the Intersection over Union (IoU) obtained both with our model and the presented baselines, while Fig. 6.6 shows some samples. We observe that, with IoU values of 23.10, our full ZPCVNet model outperforms every other baseline by a large margin, homography being the runner-up 12.45 points behind. Note, for reference, that the CV_{test,v_r} model evaluated on source view pairs (*i.e.* the original monocular system) scored an IoU of 33.43. Interestingly, the FSCV upper bound fails to learn any significant mapping, and yields unusable results that are close to the naive CV_{test,v_r} approach, evidencing the difficulty of the proposed problem. Our Mix&Match implementation is also unsuccessful in modeling the cross-view transformation, and scores below the homography, which does not use the unlabeled dataset. These results confirm the importance of the CVT module (absent in Mix&Match and FSCV) for cross-view semantic segmentation. Fig. 6.7 shows a visual comparison of our results for some additional sample scenes, together with the results from every other considered method. We observe that our method is the only one providing reasonable predictions both in terms of spatial structure and predicted classes. For the specific case of the planar homography-based baseline (*Homography*), Fig. 5.10 depicts the scene and image pair employed in the offline computation of the 3×3 homography (induced by the $z = 0$ plane) relating both views. Meanwhile, Fig. 6.8 shows the intermediate output at each of the steps taken during the execution of such baseline. We can observe that, as expected, the mapping of non-planar objects is not correctly achieved due to the wide-baseline setup: the distance between cameras and the distance between camera and scene objects are in the same order of magnitude, while the planar homography approach would only perform satisfactory mappings on planar objects or on setups where the inter-camera distance is much shorter than the camera-objects distance. The large amount of occlusions from MVMO renders the task even more challenging.

Ablation study. We now conduct an ablation study (Table 6.2) so as to determine the contribution of the pseudo-labels and CVT to the results. Surprisingly, none of them are, on their own, able to significantly boost the predictive performance of the vanilla model. In fact, it is the simultaneous action of both components that enables the success of our full ZPCVNet model in the setup from experiment 1.

Method	CVT	PL	noBG	Avg.
Vanilla	No	No	03.17	09.26
Vanilla+PL	No	Yes	02.80	08.70
Vanilla+CVT	Yes	No	05.73	11.41
ZPCVNet (full model)	Yes	Yes	**17.59**	**23.10**

Table 6.2: IoU results for the ablation study of our approach over experiment 1 on MVMO's OO test set. First two rows correspond to basic variants of Mix&Match [270], in which only one directional transformations are leveraged. **PL**: Pseudo-labels.

Method	Trained on	bathtub	bed	chair	desk	dresser	monitor	nightstand	sofa	table	toilet	noBG	Avg.
FSv_t $\star$	$x^i_{t,l}, y^i_{t,l} \sim \mathscr{D}^+_{r,l}$	30.86	33.11	24.70	18.88	25.19	29.39	33.06	22.14	11.80	32.35	26.15	31.67
CV$_{test}$	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$	09.87	10.50	03.39	04.67	03.67	09.16	03.15	10.91	05.96	04.31	06.56	13.29
Homography	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$	06.49	08.49	01.12	03.10	02.03	05.11	01.46	06.96	05.97	02.35	04.31	10.42
ZPCVNet (ours)	$x^i_{r,l}, y^i_{r,l} \sim \mathscr{D}_{r,l}$ $x^j_{r,u}, x^j_{t,u} \sim \mathscr{D}_u$	**16.40**	**17.07**	**09.57**	**10.13**	**06.54**	**14.87**	**06.73**	**14.95**	**11.20**	**11.80**	**11.93**	**18.10**

Table 6.3: IoU for experiment 2 on MVMO's OO test set: cross-view semantic transfer with RGB test set inputs captured from target view v_t and predictions referenced to v_t. $\star$: Upper bound. **noBG**: Average of all classes but background.

6.4.2 Experiment 2: cross view with output in v_t

In the second cross-view semantic transfer experiment, the goal after modifying the camera pose from v_r to v_t is to yield predictions referenced to v_t. Hence, we extend our $\mathscr{D}_{t\to r,test}$ test set to include the target view ground truth $y^k_{t,test}$ as well, so that we can use $(x^k_{t,test}, y^k_{t,test}) \sim \mathscr{D}^+_{t\to r,test}$ for evaluation. Instead of posing a new model aimed at achieving this, we take our setup trained from *experiment 1* (section 6.4.1) and keep only the E_r and G_r components, plugging the CVT module in between them out and feeding the output of E_r directly into G_r.

Baselines. We run the following baselines and bounds: (i) FSv_t: we take the model trained in the CV$_{test}v_t$ case from section 6.4.1 and test it on $(x^k_{t,test}, y^k_{t,test})$, target view pairs. This is a fully supervised upper bound to our task, should we have available the ground truth for such target view. (ii) CV_{test}: we run the simple baseline of directly applying the fully supervised $E_r - G_r$ model trained on reference (RGB, semantic) pairs to the $(x^k_{t,test}, y^k_{t,test})$ pairs. (iii) *Homography*: we extend the approach from section 6.4.1, with the additional step of projecting the semantic prediction $\tilde{y}^k_{r,test}$ back to the target viewpoint using the inverse homography $H^{z=0}_{r\to t} = (H^{z=0}_{t\to r})^{-1}$.

Results. Table 6.3 shows that, with an IoU of 18.10, our approach yields a 36.2% performance gain over the best baseline, the naive application of the model trained in the reference view. The discrete results of the homography (see Fig. 6.8 for visual outcomes), consistent with those presented in Table 6.1, can be explained by the fact that the model fed with the homography-projected images never saw such kind of features during training. Fig. 6.9 shows some corresponding visual results as compared with the rest of the proposed baselines and upper bounds. Our model, trained as explained to yield its output in the frame of the reference view (section 6.3), can simultaneously produce the results shown here (outperforming every other baseline except for the provided fully supervised upper bound) with the sole removal of the Cross-View Transformer component. In such case, the inference configuration is reduced to a $E_t - G_r$ module, as can be

seen in Fig. 6.3.

This also confirms the enormous challenge that represents the transfer of semantic knowledge across wide baseline camera locations, as both simple baselines and learned approaches that do not explicitly handle the non-aligned nature of the data fail to provide useful solutions that can leverage the complementary information provided by both cameras. Even though ZPCVNet outperforms such baselines, we are still far from the upper bounds defined by the straight fully-supervised approaches within each view (IoU=33.43 for v_r and IoU=31.67 for v_t), suggesting that this is an interesting research problem to be pursued.

6.5 Conclusions

In this chapter we introduced and proposed an initial solution for the semi-supervised task of *zero-pair, cross-view semantic segmentation*, in which a trained semantic segmentation system requires relocating its fixed camera placement but keep producing predictions on either original or new viewpoint. The proposed ZPCVNet method exploits the complementary information provided by the cross-view RGB image pairs and, as a result, outperforms other reasonable baselines by a large margin over MVMO, without requiring any labelling from the new viewpoint. This paves the way for further research in the direction of dense semantic knowledge transfer across views that can narrow the current gap with the associated (in practice, infeasible) fully supervised performance.

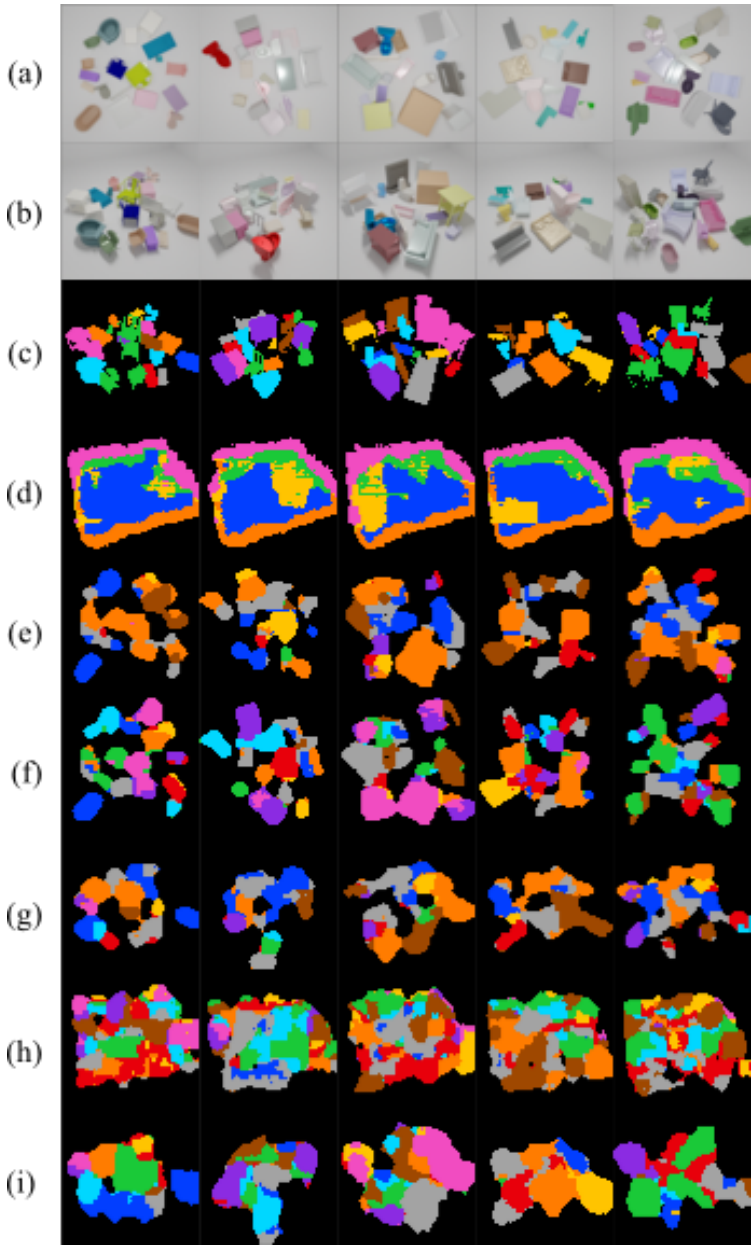


Figure 6.7: Qualitative results for all methods in Experiment 1: (a) $x^k_{t,test}$: input view on target frame v_t (b) Ground truth view of the scene from the reference frame v_r (not available at test time) (c) Ground truth semantic segmentation on the reference of the reference view v_r. This is the task ground truth for Experiment 1. (d) Outcome for method *FSCV* (Fully-Supervised Cross-View) (e) Outcome for method $\text{CV}_{test} v_r$ (f) Outcome for method $\text{CV}_{test} v_t$ (g) Outcome for planar homography. See also Figs. 5.10 and 6.8 for detailed steps and intermediate results of this method. (h) Outcome for *Mix&Match Networks* [270] (i) Outcome for ZPCVNet (ours)

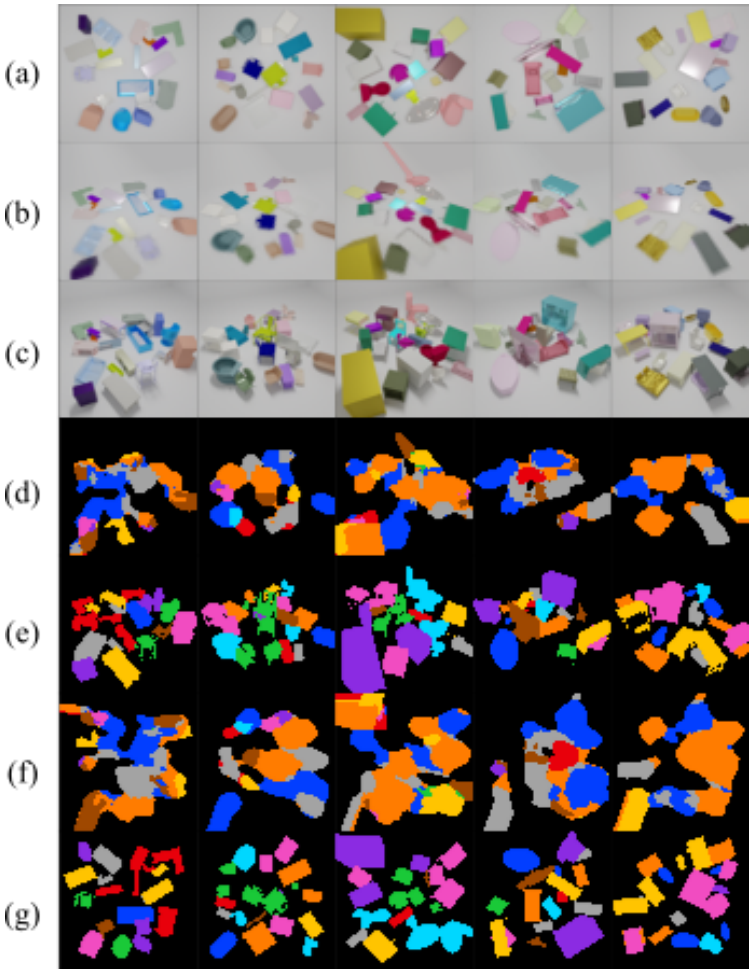


Figure 6.8: Detailed steps for the planar homography-based baseline in experiments 1 and 2: (a) $x^k_{t,test}$: input view on target frame v_t (b) $H^{z=0}_{t\to r}(x_t)$: output of the planar homography induced by the $z=0$ plane mapping target and reference views, applied to $x^k_{t,test}$ (c) Ground truth view of the scene from the reference frame v_r (not available at test time) (d) $G_r(E_r(H^{z=0}_{t\to r}(x_t)))$: result of applying the semantic segmentation model trained in a fully supervised way on the reference frame input and semantic ground truth pairs. This is the output of the homography-based baseline for Experiment 1 (e) Ground truth for Experiment 1. The previous prediction should resemble this. (f) [only for Experiment 2] $H^{z=0}_{r\to t}(G_r(E_r(H^{z=0}_{t\to r}(x_t))))$: result of applying the inverse homography to the prediction for Experiment 1. (g) [only for Experiment 2] Ground truth for Experiment 2. The previous prediction should resemble this.

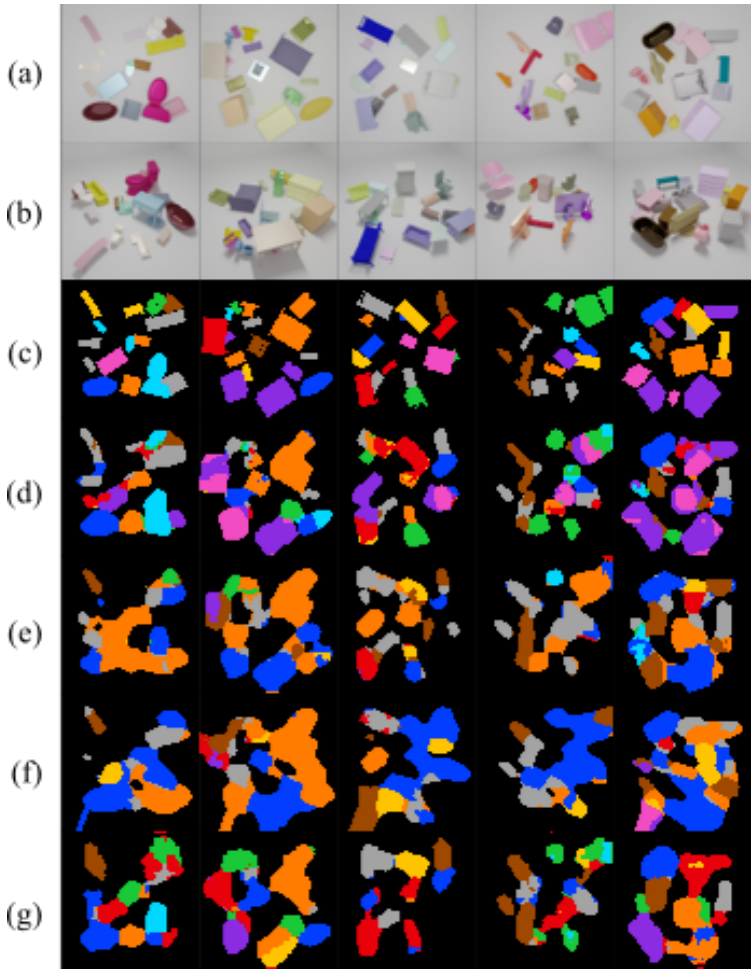


Figure 6.9: Qualitative results for all methdos in Experiment 2: (a) $x^k_{t,test}$: input view on target frame v_t (b) Ground truth view of the scene from the reference frame v_r (not available at test time, shown here for reference.) (c) Ground truth semantic segmentation on the reference of the target view v_t. This is the task ground truth for Experiment 2. (d) Outcome for method FSv_t. This corresponds to a fully-supervised upper bound, using pairs of images from rows (a), (c). (e) Outcome for method CV_{test} (f) Outcome for planar homography. See also Figs. 5.10 and 6.8 for detailed steps and intermediate results of this method. (g) Outcome for ZPCVNet (ours).

7 Conclusions

This thesis comprises different approaches to handle several particular scenarios with the common context of the scarcity —or even absence— of labeled data to be used for model learning. This problem, inherent to many applied computer vision problems, has been overcome throughout the different chapters using a variety of self and semi-supervised methods to handle mostly different image-to-image mapping problems, where some edge —in the form of prior knowledge and/or synthetic data derived from it— could be exploited. With these tools, we addressed the tasks of blur detection, hyperspectral reconstruction, temperature-spectral emissivity separation and cross-view semantic segmentation. We hereby summarize the main contributions of the thesis, and present potential directions for future work.

7.1 Summary of contributions

Chapter 1 discusses specific research objectives with respect to each of the identified challenges associated with the aforementioned tasks. We now reproduce them and discuss the solutions proposed in the respective chapters.

> **Objective 1— Self-supervised blur detection through synthetic blurring of scenes**: propose a self-supervised framework to bypass the need for large-scale blur localization datasets. Study the combined effectiveness of unsupervised object proposals and synthetic modeling of blur degradations to produce augmented training data.

In Chapter 2 we step aside the use of fully supervised end-to-end learning of image-to-image models for blur localization from medium-sized datasets, which led to limited success in the past, and instead leverage a set of physics-based models for defocus and motion blur generation to produce synthetically blurred images. The use of a class-agnostic object proposals, together with halo artifact removal inpainting techniques, allows us to create semantically plausible object masks that enable the generation of an online stream of augmented images with partial blurs and associated masks. This leads to the construction of a framework that can be employed in purely self-supervised, weakly supervised or semi-supervised configurations. We show that our framework improves state-of-the-art results for two distinct test datasets even without ever observing any real blurred image, and that the addition of a small number of annotated images consistently outperforms the fully supervised approach to the problem.

> **Objective 2— Spatio-spectral feature fusion for hyperspectral reconstruction**: current approaches for RGB to hyperspectral image recovery disregard the potentially useful information provided by nearby pixels. Hence, we propose the use of a Deep Convolutional Generative Adversarial Network to exploit the spatio-textural information from the local neighborhood while producing plausible reconstructions.

In Chapter 3 we address the significantly underconstrained problem of the color to hyperspectral image reconstruction. We show how traditional techniques used to exploit diverse statistical properties of the spectral signal in an attempt to construct informative priors for real surface reflectances in order to be able to build

such mappings. We note, however, how most of them treat each sample independently, thus ignoring the information from their local vicinity when trying to obtain high spectral resolution versions of the input. Interestingly, no CNN-based method had ever been tried for this task. Therefore, we formulate the task as an image to image mapping learning problem, and solve it using a conditional generative adversarial framework in an attempt to capture spatial semantics and yield plausible and consistent reconstructions. The absence of simultaneously captured RGB-HSI pairs is circumvented through the colorimetrically rigorous, synthetic generation of RGB versions of real HSI images. Quantitative evaluation shows a Root Mean Squared Error (RMSE) drop of 44.7% and a Relative RMSE drop of 47.0% on the ICVL natural hyperspectral image dataset, therefore setting a new state of the art for the task.

> **Objective 3— Robust multi-spectrometer hardware system for remote spectral radiance acquisition**: there is a lack of adequate rugged commercial hardware able to obtain wide range remote spectral radiance raw data from hot remote sources. Thus, we tackle the design and development of our own capture device, which is required to provide fast, reliable and calibrated wide range radiance readings over a reduced size spot from a safe distance from the EAF.

> **Objective 4— Self-supervised model for simultaneous temperature and spectral emissivity estimation**: we aim at defining and validating a method that can leverage a well-established radiative transfer model within a blind inverse problem modeling framework without extensive available ground truth, so that we are able to estimate both sample temperature and spectral emissivity over captures obtained with the described hardware.

The two preceding objectives are tackled in Chapter 4. There, we contextualize the addressed problem of temperature and spectral emissivity separation for hot samples in the broader context of the quest for the —to date still unattainable— full online remote monitoring of the steel production process parameters in EAF-based steelmaking plants. Estimating the temperature of emissive samples (e.g. liquid slag) in such harsh industrial environments is indeed a crucial yet challenging task, which is typically addressed by means of methods that require physical contact. Current remote methods require information on the emissivity of the sample. However, the spectral emissivity is dependent on the sample composition and temperature itself, and it is hardly measurable unless under controlled laboratory procedures. In Chapter 4, we present a portable device and associated probabilistic model that can simultaneously produce quasi real-time estimates for temperature and spectral emissivity of hot samples in the [0.2, 12.0$\boldsymbol{\mu} m$] range at distances of up to 20$\boldsymbol{m}$. The model is robust against variable atmospheric conditions, and the device is presented together with a quick calibration procedure that allows for in field deployment in rough industrial environments, thus enabling in line measurements. We validate the temperature and emissivity estimates of our device against laboratory equipment under controlled conditions in the [550, 850°$\boldsymbol{C}$] temperature range for two solid samples with well characterized spectral emissivities: alumina ($\boldsymbol{\alpha - Al_2O_3}$) and hexagonal boron nitride ($\boldsymbol{h - BN}$). The analysis of the results yields Root Mean Squared Errors of 32.3°$\boldsymbol{C}$ and 5.7°$\boldsymbol{C}$ respectively, and well correlated spectral emissivities. This development led to a European patent application [196] being filed by ArcelorMittal, our client and end user of the resulting asset.

> **Objective 5— Multi-view, multi-object dataset for semantic segmentation**: create a new synthetic path-tracing based dataset, comprising scenes with multiple, varied objects per-scene and multiple, wide-baseline views where each of them is annotated in terms of semantic segmentation. Release the code and dataset to the community for public access.

In chapter 5, we introduce MVMO (Multi-View Multi-Object dataset), a large scale multi-view photorealistic synthetic dataset with multi-class semantic segmentation annotations that, unlike existing alternatives,

features wide baselines between many camera pairs, a high object density and a large amount of occlusions. MVMO is a synthetic, path tracing-based set of 116,000 scenes, observed from 25 equally distributed camera poses, with fine-detailed per-view semantic segmentation annotations of 10 object categories. Given its challenging setting, evidenced by the provided experimental baseline results, we expect that MVMO will drive further research in the largely overlooked fields of cross and multi-view semantic parsing as approaches that are called to overcome the fundamental limitations of monocular semantic segmentation setups. Moreover, the fact that both the images and the code have been publicly released* enables the extension to further dense prediction tasks as well (*e.g.* cross-view depth prediction, novel view synthesis).

> **Objective 6— Semi-supervised cross-view semantic segmentation**: define and validate a data-efficient approach for semantic transfer of dense predictions across wide-baseline views upon the event of a forced camera-relocation.

Finally, in chapter 6 we take advantage of the MVMO dataset to handle one particular data-efficient instantiation of the cross-view semantic knowledge transfer problem: the relocation of the camera in a working monocular semantic segmentation system will certainly cause a substantial degradation of the predictive performance. Therefore, we introduce the novel task of *zero-pair cross-view semantic segmentation*, in which we are allowed to obtain an additional set of unlabeled but synchronized image pairs of new scenes from both original and new camera poses. Under such assumption, we present ZPCVNet, a CNN model composed of a set of encoders and decoders that are trained jointly with the available data so that they share common latent representations of the scenes. Such alignment enables the combination of encoder and decoders of arbitrary viewpoints at inference time so as to yield segmentation results from the desired reference. The envisioned model comprises a cross-view transformation module as well, which facilitates the cross-view alignment. Our experiments over the MVMO dataset evidence the failure of classic geometry-based baselines for this task, and show that ZPCVNet outperforms these and other learning-based baselines.

Software packages: as mentioned through the respective chapters, the code supporting part of our work has been made available, and can be found in our *Summary of published code*.

7.2 Future research directions

Self-supervised learning of visual representations has rapidly evolved since the early works, *i.e.* those which leveraged pretraining over handcrafted pretext tasks of little practical value [79, 137, 179] —for which automatic labels could be generated— in an attempt to learn semantically rich features that could be transferred for subsequent downstream image understanding tasks. The second —ongoing— wave of self-supervised representation learning works relies on varied forms of contrastive [39, 181, 253], clustering-based [28–30] or teacher-student learning (*i.e.* feature reconstruction) based [30, 41, 78, 89] approaches, often tightly coupled with the application of strong augmentation techniques to generate distinct versions (views) of one image. These have recently achieved significant landmarks, surpassing even the performance of the fully-supervised pretraining for various tasks [29, 40, 89], and thus posing themselves as potential candidates to become the dominant paradigm for neural net pretraining.

However, an overwhelming majority of these works are biased, either by design or indirectly through the use of Imagenet-1000-like [229] datasets, towards domains where there exists one single predominant object in the —unlabeled— images used for training. Furthermore, they cannot be directly applied (*e.g.* through k-NN assignment) to pixel-dense prediction tasks, unless complemented with freshly initialized

*The MVMO dataset and code are available at https://aitorshuffle.github.io/projects/mvmo/

decoders that must be trained on densely labeled data. Even in that case, they are likely to produce mediocre results due to the aforementioned bias. Consequently, there is a large void to be filled by expanding the current emergence of self-supervised learning methods to dense, image-to-image mapping problems, as only object-detection oriented extensions have been attempted so far [104, 280]. This is of particular interest for the fields of cross-view and multi-view semantic segmentation, as they provide a natural notion of "views" of a 3D scene in the geometric interpretation of the term that could be potentially exploited by contrastive or feature-reconstruction methods.

With regard to our particular proposed setting from Chapter 6, we envision two additional potentially fruitful research directions: one would be advancing in the inclusion of explicit physics-inspired inductive biases from the field of classic multi-view geometry (particularly so as some form of the epipolar constraint) in order to obtain a significant restriction of the solution search space [95]. The other is the inclusion of attention in the framework as a mechanism capable of providing scene-dependent inference-time matches between features across views in substitution of the current Cross-View Transformer module. The latter lies in perfect alignment with the recent rise of attention-based transformers [262] in computer vision [59] (paradoxically, advancing towards inductive bias-free architectures) and their excellent behavior in conjunction with self-supervised learning objectives [30], even showing the emergence of dense features for multiple objects in the image.

Actually, the first works employing these ingredients are already being issued [232]. These evidence the shrinkage of the gap of multi-view semantic segmentation with other related fields as well, especially so with that of novel view synthesis. Along these lines, the explosion of new forms of implicit 3D scene representations and neural radiance fields [166] hints at a more than probable cross-fertilization between both predictive tasks.

The stated ubiquitous presence of transformers (and self-attention as the most fundamental building block) supports the extended belief in the community that they are called to represent the prevailing paradigm as for architecture design choices across a variety of tasks and settings. In fact, recent trends in both blur detection [113] and hyperspectral reconstruction [193] problems point towards such direction, together with their evaluation under more severe, real-world conditions [8].

Crucial to the development of the aforesaid strategies for cross and multi-view scene understanding is the availability of adequate synthetic, labeled datasets —such as MVMO— that can stand in for the absent real ones. Our future plans comprise the release of the code necessary to extend current semantic segmentation annotations in MVMO to additional tasks and modalities (*e.g.* instance/panoptic segmentation, dense depth), in order to facilitate the development of multimodal approaches and undertaking new, rapidly spreading complex computer vision tasks, such as multi-view amodal segmentation [146], cross-view and cross-modal translations, or object-centric learning [154] related ones, among others. The very recent explosion of new multi-view, multi-object synthetic datasets [256] and dataset generators [88] reveal the growing interest of the aforesaid research lines.

Moving on to our approach for temperature and spectral emissivity separation in Chapter 4, it is apparent that there is a long way forward until the desired full monitoring of online process parameters is achieved in EAF plants. Accurate remote and online multi-variate regression of slag components' concentration would be a major step towards such final goal, one which would first require yielding even more accurate predictions on temperature and emissivity values than those obtained from the presented system. In particular, the choice of discrete, fuzzy membership function-based representation of the spectral emissivity in our radiative transfer model is one key element that calls for a modeling alternative providing better spectral resolution while keeping the number of parameters low. To that respect, the integration of Gaussian processes [207] into the workflow as a continuous representation mechanism for the spectral emissivity is one remarkably promising direction of future research to be pursued from the modeling perspective.

On the other hand, from the experimental point of view, the validation of such improved model not only on hot solid, but also on molten samples would constitute a significant milestone. However, there currently exists hardly any laboratory-grade equipment able to perform such emissivity measurements for molten samples under controlled atmospheric and sample conditions for the obtention of a reliable ground truth. A wider availability of such rare piece of laboratory equipment would definitely represent a major highlight and would contribute to significant breakthroughs in the field. Finally, an orthogonal yet exciting research direction would be that of extending the domain of application of the device and, especially, the method (which is agnostic to the particular spectrometer being used) to other fields out of steelmaking. Vulcanology, for instance, has recently gained massive public attention due to the 2021 eruptions of Fagradalsfjall (Iceland) and Cumbre Vieja (Spain) [267]. The extensive monitorization of the expelled hot magma through a variety of techniques (*e.g.* IR cameras, remote FTIR spectrometers, physical samples analyzed in laboratory, etc.) in such events represents a unique opportunity to validate the proposed model out of the context it was initially intended for and to contribute to additional scientific research fields.

Summary of published works

1. **A. Alvarez-Gila**, A. Galdran, E. Garrote, and J. van de Weijer, "Self-supervised blur detection from synthetically blurred scenes," Image and Vision Computing, 92:103804, December 2019

2. **A. Alvarez-Gila**, J. van de Weijer, and E. Garrote, "Adversarial Networks for Spatial Context-Aware Spectral Image Reconstruction from RGB," in IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

3. B. Arad, O. Ben-Shahar, R. Timofte, L. Van Gool, L. Zhang, M.-H. Yang, Z. Xiong, C. Chen, Z. Shi, D. Liu, F. Wu, C. Lanaras, S. Galliani, K. Schindler, T. Stiebel, S. Koppers, P. Seltsam, R. Zhou, M. El Helou, F. Lahoud, M. Shahpaski, K. Zheng, L. Gao, B. Zhang, X. Cui, H. Yu, Y. B. Can, **A. Alvarez-Gila**, J. van de Weijer, E. Garrote, A. Galdran, M. Sharma, S. Koundinya, A. Upadhyay, R. Manekar, R. Mukhopadhyay, H. Sharma, S. Chaudhury, K. Nagasubramanian, S. Ghosal, A. K. Singh, A. Singh, B. Ganapathysubramanian, and S. Sarkar, "NTIRE 2018 Challenge on Spectral Reconstruction from RGB Images," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.

4. A. Picon⋆, **A. Alvarez-Gila**⋆, J. A. Arteche, G. A. López, and A. Vicente, "A Probabilistic Model and Capturing Device for Remote Simultaneous Estimation of Spectral Emissivity and Temperature of Hot Emissive Materials," IEEE Access, 9:100513–100529, 2021. ⋆ Equal contribution.

5. A. Picon, **A. Alvarez-Gila**, Arteche, Jose Antonio and A. Vicente, "System and method for determining the emitting temperature and emissivity in a wavelength range of metallurgical products," PCT/IB2019/061335, filed December 24, 2019.

6. **A. Alvarez-Gila**, J. van de Weijer, Y. Wang, and E. Garrote, "MVMO: A Multi-Object Dataset for Wide Baseline Multi-View Semantic Segmentation," arXiv:2205.15452 [cs], May 2022.

Summary of published code

1. **SynthBlur**, pretrained models resulting from the experimental section of Chapter 2, presented in "Self-supervised blur detection from synthetically blurred scenes", Image and Vision Computing, 2019: https://github.com/aitorshuffle/synthblur
2. **MVMO**, code and dataset presented in Chapter 5 and in [5]: https://aitorshuffle.github.io/projects/mvmo/

Bibliography

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[2] Farnaz Agahian, Seyed Ali Amirshahi, and Seyed Hossein Amirshahi. Reconstruction of reflectance spectra using weighted principal component analysis. *Color Research & Application*, 33(5):360–371, October 2008.

[3] Aitor Alvarez-Gila, Adrian Galdran, Estibaliz Garrote, and Joost van de Weijer. Self-supervised blur detection from synthetically blurred scenes. *Image and Vision Computing*, 92:103804, December 2019.

[4] Aitor Alvarez-Gila, Joost van de Weijer, and Estibaliz Garrote. Adversarial Networks for Spatial Context-Aware Spectral Image Reconstruction from RGB. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[5] Aitor Alvarez-Gila, Joost van de Weijer, Yaxing Wang, and Estibaliz Garrote. MVMO: A Multi-Object Dataset for Wide Baseline Multi-View Semantic Segmentation. *arXiv:2205.15452 [cs]*, May 2022.

[6] Boaz Arad and Ohad Ben-Shahar. Sparse Recovery of Hyperspectral Signal from Natural RGB Images. In *European Conference on Computer Vision (ECCV)*, 2016.

[7] Boaz Arad, Ohad Ben-Shahar, Radu Timofte, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Zhiwei Xiong, Chang Chen, Zhan Shi, Dong Liu, Feng Wu, Charis Lanaras, Silvano Galliani, Konrad Schindler, Tarek Stiebel, Simon Koppers, Philipp Seltsam, Ruofan Zhou, Majed El Helou, Fayez Lahoud, Marjan Shahpaski, Ke Zheng, Lianru Gao, Bing Zhang, Ximin Cui, Haoyang Yu, Yigit Baran Can, Aitor Alvarez-Gila, Joost van de Weijer, Estibaliz Garrote, Adrian Galdran, Manoj Sharma, Sriharsha Koundinya, Avinash Upadhyay, Raunak Manekar, Rudrabha Mukhopadhyay, Himanshu Sharma, Santanu Chaudhury, Koushik Nagasubramanian, Sambuddha Ghosal, Asheesh K. Singh, Arti Singh, Baskar Ganapathysubramanian, and Soumik Sarkar. NTIRE 2018 challenge on spectral reconstruction from RGB images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[8] Boaz Arad, Radu Timofte, Ohad Ben-Shahar, Yi-Tun Lin, and Graham D. Finlayson. NTIRE 2020 Challenge on Spectral Reconstruction From an RGB Image. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 2017.

[10] Worldsteel Association et al. http://www.worldsteel.org. *accessed on*, 19, 2020.

[11] Fernando Ayala, José F. Echávarri, Pilar Renet, and Angel I. Negueruela. Use of three tristimulus values from surface reflectance spectra to calculate the principal components for reconstructing these

spectra by using only three eigenvectors. *Journal of the Optical Society of America A*, 23(8):2020–2026, August 2006.

[12] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481 – 2495, 2017.

[13] Soonmin Bae and Frédo Durand. Defocus Magnification. *Computer Graphics Forum*, 26(3):571–579, September 2007.

[14] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. PixelNet: Representation of the pixels, by the pixels, and for the pixels. *arXiv:1702.06506 [cs]*, February 2017.

[15] Alessandro Barducci, Donatella Guzzi, Cinzia Lastri, Paolo Marcoionni, Vanni Nardino, and Ivan Pippi. Emissivity and Temperature Assessment Using a Maximum Entropy Estimator: Structure and Performance of the MaxEnTES Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):738–751, February 2015.

[16] Alessandro Barducci, Donatella Guzzi, Cinzia Lastri, Vanni Nardino, Ivan Pippi, and Valentina Raimondi. Emissivity spectra estimated with the MaxEnTES algorithm. In *SPIE Sensors, Systems, and Next-Generation Satellites*, 2014.

[17] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[18] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, June 2021.

[19] Candace Berrett, Gustavious Paul Williams, Todd Moon, and Jacob Gunther. A Bayesian Nonparametric Model for Temperature-Emissivity Separation of Long-Wave Hyperspectral Images. *Technometrics*, 56(2):200–211, April 2014.

[20] Marcin Blachnik, Krystian Mączka, and Tadeusz Wieczorek. A model for temperature prediction of melted steel in the electric arc furnace (eaf). In *International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, 2010.

[21] Michael Bleyer, Carsten Rother, Pushmeet Kohli, Daniel Scharstein, and Sudipta Sinha. Object stereo — Joint stereo matching and object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[22] S. Bohnes, V. Scherer, S. Linka, M. Neuroth, and H. Brüggemann. Spectral Emissivity Measurements of Single Mineral Phases and Ash Deposits. In *ASME Summer Heat Transfer Conference (SHTC)*, 2005.

[23] Christoph C. Borel. Surface emissivity and temperature retrieval for a hyperspectral sensor. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1998.

[24] Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, and William T. Freeman. Computational Imaging for VLBI Image Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Automatic 3D object segmentation in multiple views using volumetric graph-cuts. *Image and Vision Computing*, 28(1):14–25, January 2010.

[26] Xun Cao, Xin Tong, Qionghai Dai, and Stephen Lin. High resolution multispectral video capture with a hybrid camera system. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[27] Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J. Brady. Computational Snapshot Multispectral Cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108, September 2016.

[28] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In *European Conference on Computer Vision (ECCV)*, 2018.

[29] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[30] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[31] Ayan Chakrabarti and Todd Zickler. Statistics of Real-World Hyperspectral Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[32] Ayan Chakrabarti, Todd Zickler, and William T. Freeman. Analyzing spatially-varying blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[33] Ding-Jie Chen, Hwann-Tzong Chen, and Long-Wen Chang. Fast defocus map estimation. In *International Conference on Image Processing (ICIP)*, 2016.

[34] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, June 2018.

[35] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *International Conference on Learning Representations (ICLR)*, 2015.

[36] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.05587 [cs]*, June 2017.

[37] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[38] Shu-Jie Chen and Hui-Liang Shen. Multispectral Image Out-of-Focus Deblurring Using Interchannel Correlation. *Transactions on Image Processing*, 24(11):4433–4445, November 2015.

[39] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 2020.

[40] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[41] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep Colorization. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[43] CIE. CIE 142-2001 Improvement to Industrial Colour-Difference Evaluation. Technical Report CIE 142-2001, Commission Internationale de L'éclairage, Vienna, 2001. ISBN 978 3 901906 08 4.

[44] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[45] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. NoVA: Learning to see in novel viewpoints and domains. In *International Conference on 3D Vision (3DV)*, 2019.

[46] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abràmoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards Adversarial Retinal Image Synthesis. *arXiv:1701.08974 [cs, stat]*, January 2017.

[48] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[49] Angela Dai and Matthias Niessner. 3DMV: Joint 3D-Multi-View Prediction for 3D Semantic Scene Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[50] Jifeng Dai, Kaiming He, and Jian Sun. Instance-Aware Semantic Segmentation via Multi-Task Network Cascades. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Doug Dailey. Netezza and IBM Cloud Pak for Data: A knockout combo for tough data, June 2020.

[52] P. Dash, F.-M. Göttsche, F.-S. Olesen, and H. Fischer. Land surface temperature and emissivity estimation from passive sensor data: Theory and practice-current trends. *International Journal of Remote Sensing*, 23(13):2563–2594, January 2002.

[53] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, January 2022.

[54] Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[55] Markus Deserno. How to generate equidistributed points on the surface of a sphere. Technical report, 2004.

[56] C. Dickinson, H. K. Eriksen, A. J. Banday, J. B. Jewell, K. M. Górski, G. Huey, C. R. Lawrence, I. J. O'Dwyer, and B. D. Wandelt. Bayesian Component Separation and Cosmic Microwave Background Estimation for the Five-Year WMAP Temperature Data. *The Astrophysical Journal*, 705(2):1607, 2009.

[57] Xiaoying Ding and Zhenzhong Chen. Improving Saliency Detection Based on Modeling Photographer's Intention. *IEEE Transactions on Multimedia*, 21(1):124–134, January 2019.

[58] Chris Donahue, Akshay Balsubramani, Julian McAuley, and Zachary C. Lipton. Semantically Decomposing the Latent Spaces of Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[59] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[60] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, 2017.

[61] Telmo Echániz, Raul B. Pérez-Sáez, and Manuel J. Tello. IR radiometer sensitivity and accuracy improvement by eliminating spurious radiation for emissivity measurements on highly specular samples in the 2–25μm spectral range. *Measurement*, 110:22–26, November 2017.

[62] Jia Eckhard, Timo Eckhard, Eva M. Valero, Juan Luis Nieves, and Estibaliz Garrote Contreras. Outdoor scene reflectance measurements using a Bragg-grating-based hyperspectral imager. *Applied Optics*, 54(13):D15–D24, May 2015.

[63] H. K. Eriksen, J. B. Jewell, C. Dickinson, A. J. Banday, K. M. Górski, and C. R. Lawrence. Joint Bayesian Component Separation and CMB Power Spectrum Estimation. *The Astrophysical Journal*, 676(1):10, 2008.

[64] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, June 2018.

[65] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[66] Paolo Favaro, Stefano Soatto, Martin Burger, and Stanley J. Osher. Shape from Defocus via Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):518–531, March 2008.

[67] James Ferryman and Ali Shahrokni. PETS2009: Dataset and challenge. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009.

[68] François Fleuret, Jérôme Berclaz, Richard Lengagne, and Pascal Fua. Multi-Camera People Tracking with a Probabilistic Occupancy Map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2008.

[69] David H. Foster and Kinjiro Amano. Hyperspectral imaging in color vision research: tutorial. *Journal of the Optical Society of America A*, 36(4):606–627, April 2019.

[70] David H. Foster, Kinjiro Amano, Sérgio M. C. Nascimento, and Michael J. Foster. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 23(10):2359, October 2006.

[71] David H. Foster, Kinjiro Amano, and Sérgio M.C. Nascimento. Time-lapse ratios of cone excitations in natural scenes. *Vision Research*, 120:45–60, March 2016.

[72] Silvano Galliani, Charis Lanaras, Dimitrios Marmanis, Emmanuel Baltsavias, and Konrad Schindler. Learned Spectral Super-Resolution. *arXiv:1703.09470 [cs]*, March 2017.

[73] Victor Garcia Satorras, Zeynep Akata, and Max Welling. Combining Generative and Discriminative Models for Hybrid Inference. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[74] Jochen Gast, Anita Sellent, and Stefan Roth. Parametric Object Motion From Blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[75] Bert Geelen, Nicolaas Tack, and Andy Lambrechts. A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. In *SPIE Advanced Fabrication Technologies for Micro/Nano Optics and Photonics*, 2014.

[76] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, September 2006.

[77] Theo Gevers, Arjan Gijsenij, Joost van de Weijer, and Jan-Mark Geusebroek. *Color in Computer Vision: Fundamentals and Applications*, volume 23. John Wiley & Sons, September 2012.

[78] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, and Matthieu Cord. Learning Representations by Predicting Bags of Visual Words. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[79] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[80] Mayank Goel, Eric Whitmire, Alex Mariakakis, T. Scott Saponas, Neel Joshi, Dan Morris, Brian Guenter, Marcel Gavriliu, Gaetano Borriello, and Shwetak N. Patel. HyperCam: Hyperspectral Imaging for Ubiquitous Computing Applications. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2015.

[81] S. Alireza Golestaneh and Lina J. Karam. Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[82] Iñigo González de Arrieta, Telmo Echániz, Raquel Fuente, Jose M. Campillo-Robles, Josu M. Igartua, and Gabriel A. López. Updated measurement method and uncertainty budget for direct emissivity measurements at the University of the Basque Country. *Metrologia*, 57(4):045002, July 2020.

[83] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[84] Iñigo González de Arrieta, Telmo Echániz, Raquel Fuente, Leire del Campo, Domingos De Sousa Meneses, Gabriel A. López, and Manuel J. Tello. Mid-infrared optical properties of pyrolytic boron nitride in the 390–1050°C temperature range using spectral emissivity measurements. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 194:1–6, June 2017.

[85] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[86] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, December 2016.

[87] I. E. Gordon, L. S. Rothman, C. Hill, R. V. Kochanov, Y. Tan, P. F. Bernath, M. Birk, V. Boudon, A. Campargue, K. V. Chance, B. J. Drouin, J. M. Flaud, R. R. Gamache, J. T. Hodges, D. Jacquemart, V. I. Perevalov, A. Perrin, K. P. Shine, M. A. H. Smith, J. Tennyson, G. C. Toon, H. Tran, V. G. Tyuterev, A. Barbe, A. G. Császár, V. M. Devi, T. Furtenbacher, J. J. Harrison, J. M. Hartmann, A. Jolly, T. J. Johnson, T. Karman, I. Kleiner, A. A. Kyuberis, J. Loos, O. M. Lyulin, S. T. Massie, S. N. Mikhailenko, N. Moazzen-Ahmadi, H. S. P. Müller, O. V. Naumenko, A. V. Nikitin, O. L. Polyansky, M. Rey, M. Rotger, S. W. Sharpe, K. Sung, E. Starikova, S. A. Tashkun, J. Vander Auwera, G. Wagner, J. Wilzewski, P. Wcisło, S. Yu, and E. J. Zak. The HITRAN2016 molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 203:3–69, December 2017.

[88] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti, Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[89] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[90] Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurelien Plyer, and David Filliat. SnapNet-R: Consistent 3D Multi-View Semantic Labeling for Robotics. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[91] Shir Gur and Lior Wolf. Single Image Depth Estimation Trained via Depth From Defocus Cues. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[92] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2016.

[93] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, April 2004.

[94] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[95] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I. Yu. Epipolar Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[96] Patrick Heasler, Christian Posse, Jeff Hylden, and Kevin Anderson. Nonlinear Bayesian Algorithms for Gas Plume Detection and Estimation from Hyper-spectral Thermal Image Data. *Sensors*, 7(6):905–920, June 2007.

[97] Ville Heikkinen, Reiner Lenz, Tuija Jetsu, Jussi Parkkinen, Markku Hauta-Kasari, and Timo Jääskeläinen. Evaluation and unification of some methods for estimating reflectance spectra from RGB images. *Journal of the Optical Society of America A*, 25(10):2444–2458, October 2008.

[98] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *Advances in Neural Information Processing Systems Workshops (NIPSW)*, 2014.

[99] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[100] M. S. Hosseini and K. N. Plataniotis. Convolutional Deblurring for Natural Imaging. *Transactions on Image Processing*, 29:250–264, 2020.

[101] John R. Howell. The Monte Carlo Method in Radiative Heat Transfer. *Journal of Heat Transfer*, 120(3):547–560, August 1998.

[102] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[103] Clive Humby. In *Association of National Advertisers Senior marketer's summit, Kellogg School*, 2006.

[104] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient Visual Pretraining With Contrastive Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[105] Francisco H Imai, Mitchell R Rosen, and Roy S Berns. Comparative study of metrics for spectral match quality. In *European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*, 2002.

[106] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, 2015.

[107] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, July 2014.

[108] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[109] Frédéric Jacob, François Petitcolin, Thomas Schmuge, Eric Vermote, Andrew French, and Kenta Ogawa. Comparison of land surface emissivity and radiometric temperature derived from MODIS and ASTER sensors. *Remote Sensing of Environment*, 90(2):137–152, March 2004.

[110] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative Models as a Data Source for Multiview Representation Learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[111] Simon Jenni and Paolo Favaro. Self-Supervised Feature Learning by Learning to Spot Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[112] Sangho Jeon, Seung-Nam Park, Yong Shim Yoo, Jisoo Hwang, Chul-Woung Park, and Geun Woo Lee. Simultaneous measurement of emittance, transmittance, and reflectance of semitransparent materials at elevated temperature. *Optics Letters*, 35(23):4015–4017, December 2010.

[113] Zeyu Jiang, Xun Xu, Le Zhang, Chao Zhang, Chuan Sheng Foo, and Ce Zhu. MA-GANet: A Multi-Attention Generative Adversarial Network for Defocus Blur Detection. *Transactions on Image Processing*, 31:3494–3508, 2022.

[114] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[115] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A General Platform for Intelligent Agents. *arXiv:1809.02627 [cs, stat]*, May 2020.

[116] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), October 2017.

[117] Rei Kawakami, Yasuyuki Matsushita, John Wright, Moshe Ben-Ezra, Yu-Wing Tai, and Katsushi Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[118] Vahid Kazemi, Magnus Burenius, Hossein Azizpour, and Josephine Sullivan. Multi-view body part recognition with random forests. In *BMVA British Machine Vision Conference (BMVC)*, 2013.

[119] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations (ICLR)*, 2017.

[120] Jack Kiefer and Jacob Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[121] Beomseok Kim, Hyeongseok Son, Seong-Jin Park, Sunghyun Cho, and Seungyong Lee. Defocus and Motion Blur Detection with Deep Contextual Features. *Computer Graphics Forum*, 37(7):277–288, October 2018.

[122] Byungsoo Kim, Vinicius C. Azevedo, Nils Thuerey, Theodore Kim, Markus Gross, and Barbara Solenthaler. Deep Fluids: A Generative Network for Parameterized Fluid Simulations. *Computer Graphics Forum*, 38(2):59–70, 2019.

[123] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[124] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2013.

[125] Vladimir V. Kniaz, Vladimir A. Knyaz, Fabio Remondino, Artem Bordodymov, and Petr Moshkantsev. Image-to-Voxel Model Translation for 3D Scene Reconstruction and Segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

[126] R. V. Kochanov, I. E. Gordon, L. S. Rothman, P. Wcisło, C. Hill, and J. S. Wilzewski. HITRAN Application Programming Interface (HAPI): A comprehensive approach to working with spectroscopic data. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 177:15–30, July 2016.

[127] Amit Kohli, Vincent Sitzmann, and Gordon Wetzstein. Semantic Implicit Neural Scene Representations with Semi-supervised Training. In *International Conference on 3D Vision (3DV)*, 2020.

[128] Adarsh Kowdle, Sudipta N. Sinha, and Richard Szeliski. Multiple View Object Cosegmentation Using Appearance and Stereo Cues. In *European Conference on Computer Vision (ECCV)*, 2012.

[129] Philipp Krahenbuhl and Vladlen Koltun. Learning to Propose Objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[130] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[131] Martin Krzywinski. A Large (9,284) List of Named Colors, October 2017.

[132] Tejas D. Kulkarni, Pushmeet Kohli, Joshua B. Tenenbaum, and Vikash Mansinghka. Picture: A Probabilistic Programming Language for Scene Perception. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[133] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[134] Herbert Köchner. Sensors and measurement techniques for process control. *Workshop on Electrical Arc Furnace: state-of-the-art of RFCS-supported projects*, 2015.

[135] Lubor Ladický, Paul Sturgess, Chris Russell, Sunando Sengupta, Yalin Bastanlar, William Clocksin, and Philip H. S. Torr. Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction. *International Journal of Computer Vision*, 2(100):122–133, 2012.

[136] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*, 2016.

[137] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[138] Steven Le Moan and Philipp Urban. Image-Difference Prediction: From Color to Spectral. *Transactions on Image Processing*, 23(5):2058–2068, May 2014.

[139] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[140] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989.

[141] Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence, March 2021.

[142] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *arXiv:1609.04802 [cs]*, September 2016.

[143] Rachel J. Lee, Michael S. Ramsey, and Penelope L. King. Development of a new laboratory technique for high-temperature thermal emission spectroscopy of silicate melts. *Journal of Geophysical Research: Solid Earth*, 118(5):1968–1983, May 2013.

[144] Wonwoo Lee, Woontack Woo, and Edmond Boyer. Silhouette Segmentation in Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1429–1441, July 2011.

[145] Chuan Li and Michael Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[146] Ke Li and Jitendra Malik. Amodal Instance Segmentation. In *European Conference on Computer Vision (ECCV)*, 2016.

[147] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep Object Co-segmentation. In *Asian Conference on Computer Vision (ACCV)*, 2018.

[148] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.

[149] S. Linka, S. Wirtz, and V. Scherer. Spectral Thermal Radiation Characteristics of Coal Ashes and Slags: Influence of Chemical Composition and Temperature. In *ASME Summer Heat Transfer Conference (SHTC)*, 2003.

[150] Junchi Liu, Hongwen Li, Jianli Wang, Hongzhuang Li, Limei Yin, and Zhenduo Zhang. Inversion of emissivity spectrum and temperature in the TIR waveband based on the Maximum Entropy. *Infrared Physics & Technology*, 72:179–190, September 2015.

[151] Renting Liu, Zhaorong Li, and Jiaya Jia. Image partial blur detection and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[152] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. RankIQA: Learning From Rankings for No-Reference Image Quality Assessment. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[153] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[154] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[155] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[156] Xavier Moles Lopez, Etienne D'Andrea, Paul Barbot, Anne-Sophie Bridoux, Sandrine Rorive, Isabelle Salmon, Olivier Debeir, and Christine Decaestecker. An Automated Blur Detection Method for Histological Whole Slide Imaging. *PLOS ONE*, 8(12):e82710, December 2013.

[157] Miguel A. López-Álvarez, Javier Hernández-Andrés, Javier Romero, F. J. Olmo, A. Cazorla, and L. Alados-Arboledas. Using a trichromatic CCD camera for spectral skylight estimation. *Applied Optics*, 47(34):H31–H38, December 2008.

[158] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[159] Kede Ma, Huan Fu, Tongliang Liu, Zhou Wang, and Dacheng Tao. Deep Blur Mapping: Exploiting High-Level Semantics by Deep Neural Networks. *Transactions on Image Processing*, 27(10):5155–5166, October 2018.

[160] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[161] Sajal Maheshwari, Pranjal Kumar Rai, Gopal Sharma, and Vineet Gandhi. Document Blur Detection Using Edge Profile Mining. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2016.

[162] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[163] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016.

[164] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[165] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[166] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[167] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv.1411.1784 [cs]*, November 2014.

[168] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S. Feris. Task2Sim : Towards Effective Pre-training and Transfer from Synthetic Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[169] Travis J. Moore, Darron P. Cundick, Matthew R. Jones, Dale R. Tree, R. Daniel Maynes, and Larry L. Baxter. In situ measurements of the spectral emittance of coal ash deposits. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(12):1978–1986, August 2011.

[170] John A. Morgan. Bayesian estimation for land surface temperature retrieval: the nuisance of emissivities. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1279–1288, June 2005.

[171] John A. Morgan. A Bayesian Estimator for Linear Calibration Error Effects in Thermal Remote Sensing. *IEEE Geoscience and Remote Sensing Letters*, 3(1):117–119, January 2006.

[172] John A. Morgan. Comparison of Bayesian land surface temperature algorithm performance with Terra MODIS observations. *International Journal of Remote Sensing*, 32(23):8139–8159, December 2011.

[173] Steve Morgan. The 2020 Data Attack Surface Report. Technical report, 2020.

[174] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A Metric Learning Reality Check. In *European Conference on Computer Vision (ECCV)*, 2020.

[175] Andrew Ng. A Chat with Andrew on MLOps: From Model-centric to Data-centric AI, March 2021.

[176] Rang M. H. Nguyen, Dilip K. Prasad, and Michael S. Brown. Training-Based Spectral Reconstruction from a Single RGB Image. In *European Conference on Computer Vision (ECCV)*, 2014.

[177] Shijie Nie, Lin Gu, Yinqiang Zheng, Antony Lam, Nobutaka Ono, and Imari Sato. Deeply Learned Filter Response Functions for Hyperspectral Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[178] Sergey I. Nikolenko. *Synthetic Data for Deep Learning*. June 2021.

[179] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *European Conference on Computer Vision (ECCV)*, 2016.

[180] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-Supervised Learning of Geometrically Stable Features Through Probabilistic Introspection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[181] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019.

[182] Tuomas Paloposki and Leif Liedquist. Steel emissivity at high temperatures, 2005.

[183] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-View Semantic Segmentation for Sensing Surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, July 2020.

[184] Jinshan Pan, Zhe Hu, Zhixun Su, Hsin-Ying Lee, and Ming-Hsuan Yang. Soft-Segmentation Guided Object Motion Deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[185] Yanwei Pang, Hailong Zhu, Xinyu Li, and Xuelong Li. Classifying Discriminative Features for Blur Detection. *IEEE Transactions on Cybernetics*, 46(10):2220–2227, October 2016.

[186] Yanwei Pang, Hailong Zhu, Xuelong Li, and Jing Pan. Motion blur detection with an indicator function for surveillance machines. *IEEE Transactions on Industrial Electronics*, 63(9):5592–5601, 2016.

[187] Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. A Unified Approach of Multi-Scale Deep and Hand-Crafted Features for Defocus Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[188] Jong-Il Park, Moon-Hyun Lee, Michael D. Grossberg, and Shree K. Nayar. Multispectral Imaging Using Multiplexed Illumination. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[189] Manu Parmar, Steven Lansel, and Brian A. Wandell. Spatio-spectral reconstruction of the multispectral datacube using sparse recovery. In *International Conference on Image Processing (ICIP)*, 2008.

[190] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[191] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[192] Henri Pauna, Matti Aula, Jonas Seehausen, Jens-Sebastian Klung, Marko Huttula, and Timo Fabritius. Optical emission spectroscopy as an online analysis method in industrial electric arc furnaces. *Steel Research International*, page 2000051, 2020.

[193] Hao Peng, Xiaomei Chen, and Jie Zhao. Residual Pixel Attention Network for Spectral Reconstruction From RGB Images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[194] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[195] Said Pertuz, Miguel Angel Garcia, and Domenec Puig. Focus-aided scene segmentation. *Computer Vision and Image Understanding*, 133:66–75, April 2015.

[196] Artzai Picon, Aitor Alvarez-Gila, Jose Antonio Arteche, and Asier Vicente. System and method for determining the emitting temperature and emissivity in a wavelength range of metallurgical products. PCT/IB2019/061335, 2019.

[197] Artzai Picon, Asier Vicente, Sergio Rodriguez-Vaamonde, Jorge Armentia, Jose Antonio Arteche, and Iñaki Macaya. Ladle furnace slag characterization through hyperspectral reflectance regression model for secondary metallurgy process optimization. *IEEE Transactions on Industrial Informatics*, 14(8):3506–3512, 2017.

[198] Artzai Picon⋆, Aitor Alvarez-Gila⋆, Jose A. Arteche, Gabriel A. López, and Asier Vicente. A Probabilistic Model and Capturing Device for Remote Simultaneous Estimation of Spectral Emissivity and Temperature of Hot Emissive Materials. *IEEE Access*, 9:100513–100529, 2021. ⋆Equal contribution.

[199] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, January 2017.

[200] Vignesh Prasad and Brojeshwar Bhowmick. SfMLearner++: Learning Monocular Depth & Ego-Motion using Meaningful Geometric Constraints. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[201] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[202] Kuldeep Purohit, Anshul B. Shah, and A. N. Rajagopalan. Learning Based Single Image Blur Detection and Segmentation. In *International Conference on Image Processing (ICIP)*, 2018.

[203] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross View Fusion for 3D Human Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[204] Weichao Qiu and Alan Yuille. UnrealCV: Connecting Computer Vision to Unreal Engine. In *European Conference on Computer Vision Workshops (ECCVW)*, 2016.

[205] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[206] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, 2021.

[207] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Mass, November 2005.

[208] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Honglak Lee, and Bernt Schiele. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, 2016.

[209] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning What and Where to Draw. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[210] Krishna Regmi and Ali Borji. Cross-View Image Synthesis Using Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[211] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional GANs. *Computer Vision and Image Understanding*, 187:102788, October 2019.

[212] Salvador Rego Barcena. *A Passive Mid-Infrared Sensor to Measure Real-Time Particle Emissivity and Gas Temperature in Coal-Fired Boilers and Steelmaking Furnaces*. Ph.d. dissertation, University of Toronto, 2008.

[213] Salvador Rego-Barcena, Rebecca Saari, Reza Mani, Sameh El-Batroukh, and Murray J. Thomson. Real time, non-intrusive measurement of particle emissivity and gas temperature in coal-fired power plants. *Measurement Science and Technology*, 18(11):3479, November 2007.

[214] David Reinsel, John Gantz, and John Rydning. Data Age 2025: The Evolution of Data to Life-Critical. Technical report, IDC, April 2017.

[215] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[216] G. H. Richards, J. N. Harb, L. L. Baxter, S. Bhattacharya, R. P. Gupta, and T. F. Wall. Radiative heat transfer in pulverized-coal-fired boilers—Development of the absorptive/reflective character of initial ash deposits. *Symposium (International) on Combustion*, 25(1):511–518, 1994.

[217] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV)*, 2016.

[218] Hayko Riemenschneider, András Bódis-Szomorú, Julien Weissenberg, and Luc Van Gool. Learning Where to Classify in Multi-view Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[219] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[220] Antonio Robles-Kelly. Single Image Spectral Reconstruction for Multimedia Applications. In *ACM International Conference on Multimedia (ICMM)*, 2015.

[221] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional Random Fields for multi-camera object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[222] Javier Romero, Antonio Garcıa-Beltrán, and Javier Hernández-Andrés. Linear bases for representation of natural and artificial illuminants. *Journal of the Optical Society of America A*, 14(5):1007–1014, 1997.

[223] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[224] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[225] Patricia Rosales, Carmen Blanco, Mónica Flores, Rosario Pareja, Arturo Revuelta, and Fernando Marquez. Comparison between the "grey body emissivity" and "Bayesian inference" methods to retrieve temperature and emissivity from FTIR spectroradiometer measurements. In *SPIE Electro-Optical and Infrared Systems: Technology and Applications*, 2010.

[226] Benoit Rousseau, Jean F. Brun, Domingos de Sousa Meneses, and Patrick Echegut. Temperature Measurement: Christiansen Wavelength and Blackbody Reference. *International Journal of Thermophysics*, 26(4):1277–1286, July 2005.

[227] Soumali Roychowdhury, Michelangelo Diligenti, and Marco Gori. Regularizing deep networks with prior knowledge: A constraint-based approach. *Knowledge-Based Systems*, 222:106989, June 2021.

[228] Olivier Rozenbaum, Domingos de Sousa Meneses, and Patrick Echegut. Texture and Porosity Effects on the Thermal Radiative Behavior of Alumina Ceramics. *International Journal of Thermophysics*, 30(2):580–590, April 2009.

[229] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[230] John Rydning. Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2021–2025. Market Forecast US47998321, IDC: The premier global market intelligence company, July 2021.

[231] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators. *Image and Vision Computing*, 96:103898, 2020.

[232] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[233] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[234] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016.

[235] Steven A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, December 1985.

[236] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2016.

[237] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[238] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative Blur Detection Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[239] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[240] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3D scene reconstruction with multi-layer depth and epipolar transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[241] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning From Simulated and Unsupervised Images Through Adversarial Training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[242] Samarth Shukla, Andrés Romero, Luc Van Gool, and Radu Timofte. Zero-pair image to image translation using domain conditional normalization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[243] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision (ECCV)*, 2012.

[244] Sho Sonoda, Noboru Murata, Hideitsu Hino, Hiroshi Kitada, and Manabu Kano. A Statistical Model for Predicting the Liquid Steel Temperature in Ladle and Tundish by Bootstrap Filter. *ISIJ International*, 52(6):1086–1091, 2012.

[245] Mario Sormann, Christopher Zach, and Konrad Karner. Graph Cut Based Multiple View Segmentation for 3D Reconstruction. In *International Symposium on 3D Data Processing, isualization, and Transmission (3DPVT)*, 2006.

[246] Bolan Su, Shijian Lu, and Chew Lim Tan. Blurred Image Region Detection and Classification. In *ACM International Conference on Multimedia (ICMM)*, 2011.

[247] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[248] Chang Tang, Chunping Hou, and Zhanjie Song. Defocus map estimation from a single image via spectrum contrast. *Optics Letters*, 38(10):1706–1708, May 2013.

[249] Chang Tang, Jin Wu, Yonghong Hou, Pichao Wang, and Wanqing Li. A Spectral and Spatial Approach of Coarse-to-Fine Blurred Image Region Detection. *IEEE Signal Processing Letters*, 23(11):1652–1656, November 2016.

[250] Alexandru Telea. An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools*, 9(1):23–34, January 2004.

[251] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016.

[252] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, November 2020.

[253] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[254] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N. Metaxas. CR-GAN: Learning complete representations for multi-view generation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[255] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[256] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. RTMV: A Ray-Traced Multi-View Synthetic Dataset for Novel View Synthesis. *arXiv:2205.07058 [cs]*, May 2022.

[257] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training Deep Networks With Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[258] Matthew Trentacoste, Rafal Mantiuk, and Wolfgang Heidrich. Blur-Aware Image Downsampling. *Computer Graphics Forum*, 30(2):573–582, 2011.

[259] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to Adapt Structured Output Space for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[260] Shinji Tsuda, Teruyuki Okazaki, and Alfred Gwosdz. Proposal for Predictive Expressions of Emissivity Spectra for Powdery Coal Ash. *Journal of Power and Energy Systems*, 6(3):360–377, 2012.

[261] Manik Varma and Andrew Zisserman. Classifying Images of Materials: Achieving Viewpoint and Illumination Independence. In *European Conference on Computer Vision (ECCV)*, 2002.

[262] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[263] Abinaya Priya Venkataraman, Aiswaryah Radhakrishnan, Carlos Dorronsoro, Linda Lundström, and Susana Marcos. Role of parafovea in blur perception. *Vision Research*, 138:59–65, September 2017.

[264] Asier Vicente, Artzai Picon, Jose Antonio Arteche, Miguel Linares, Arturo Velasco, and Jose Angel Sainz. Magnetic field-based arc stability sensor for electric arc furnaces. *Measurement*, 151:107134, 2020.

[265] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[266] Sebastian H. Völkel, Christian J. Krüger, and Kostas D. Kokkotas. Bayesian inverse problem of rotating neutron stars. *Physical Review D*, 103(8):083008, April 2021.

[267] Fabian B. Wadsworth, Edward W. Llewellin, Jamie I. Farquharson, Janina K. Gillies, Ariane Loisel, Léon Frey, Evgenia Ilyinskaya, Thor Thordarson, Samantha Tramontano, Einat Lev, Matthew J. Pankhurst, Alejandro Galdeano Rull, María Asensio-Ramos, Nemesio M. Pérez, Pedro A. Hernández, David Calvo, M. Carmen Solana, Ulrich Kueppers, and Alejandro Polo Santabárbara. Crowd-sourcing observations of volcanic eruptions during the 2021 Fagradalsfjall and Cumbre Vieja events. *Nature Communications*, 13(1):2611, May 2022.

[268] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, October 2018.

[269] Xiaolong Wang and Abhinav Gupta. Generative Image Modeling Using Style and Structure Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[270] Yaxing Wang, Luis Herranz, and Joost van de Weijer. Mix and Match Networks: Cross-Modal Alignment for Zero-Pair Image-to-Image Translation. *International Journal of Computer Vision*, June 2020.

[271] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and Match Networks: Encoder-Decoder Alignment for Zero-Pair Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[272] M Wiecek, R Strakowski, B Wiecek, R Olbrycht, T Świątczak, W Wittchen, and M Borecki. Estimation of steel slag parameters using thermal imaging and neural networks classification. In *International Conference on Quantitative InfraRed Thermography (QIRT)*, 2010.

[273] Thomas Wiecki. Probabilistic Programming in Quantitative Finance, April 2015.

[274] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end View Synthesis from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[275] Cornelius J. Willers. *Electro-Optical System Analysis and Design: A Radiometry Perspective*. SPIE Press, Bellingham, Wash, April 2013.

[276] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building Generalizable Agents with a Realistic and Rich 3D Environment. *arXiv:1801.02209 [cs]*, January 2018.

[277] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[278] Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-World Perception for Embodied Agents. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[279] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3D semi-supervised learning with uncertainty-aware multi-view co-training. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[280] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[281] Yuan Yao, Yasamin Jafarian, and Hyun Soo Park. MONET: Multiview Semi-Supervised Keypoint Detection via Epipolar Divergence. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[282] Yuan Yao and Hyun Soo Park. Multiview Co-segmentation for Wide Baseline Images using Cross-view Supervision. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[283] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized Assorted Pixel Camera: Postcapture Control of Resolution, Dynamic Range, and Spectrum. *Transactions on Image Processing*, 19(9):2241–2253, September 2010.

[284] Xin Yi and Mark Eramian. LBP-Based Segmentation of Defocus Blur. *Transactions on Image Processing*, 25(4):1626–1638, April 2016.

[285] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[286] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations (ICLR)*, 2016.

[287] Kai Zeng, Yaonan Wang, Jianxu Mao, Junyang Liu, Weixing Peng, and Nankai Chen. A Local Metric for Defocus Blur Detection Based on CNN Feature Learning. *Transactions on Image Processing*, 28(5):2107–2115, May 2019.

[288] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting Ground-Level Scene Layout From Aerial Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[289] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful Image Colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[290] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Networks as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[291] Shanghang Zhang, Xiaohui Shen, Zhe Lin, Radomír Měch, João P. Costeira, and José M. F. Moura. Learning to Understand Image Blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[292] Xinxin Zhang, Ronggang Wang, Xiubao Jiang, Wenmin Wang, and Wen Gao. Spatially variant defocus blur map estimation and deblurring from a single image. *Journal of Visual Communication and Image Representation*, 35:257–264, February 2016.

[293] Yilun Zhang and Hyun Soo Park. Multiview Supervision By Registration. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[294] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus Blur Detection via Multi-Stream Bottom-Top-Bottom Fully Convolutional Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[295] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus Blur Detection via Multi-Stream Bottom-Top-Bottom Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[296] Yonghui Zhao and Roy S. Berns. Image-based spectral reflectance reconstruction using the matrix R method. *Color Research & Application*, 32(5):343–351, October 2007.

[297] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[298] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[299] Changyin Zhou, Oliver Cossairt, and Shree Nayar. Depth from Diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[300] Changyin Zhou, Stephen Lin, and Shree Nayar. Coded aperture pairs for depth from defocus. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[301] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative Visual Manipulation on the Natural Image Manifold. In *European Conference on Computer Vision (ECCV)*, 2016.

[302] Tong Zhu and Lina J. Karam. Efficient perceptual-based spatially varying out-of-focus blur detection. In *International Conference on Image Processing (ICIP)*, 2016.

[303] Xiang Zhu, Scott Cohen, Stephen Schiller, and Peyman Milanfar. Estimating Spatially Varying Defocus Blur From A Single Image. *Transactions on Image Processing*, 22(12):4879–4891, December 2013.

[304] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative Adversarial Frontal View to Bird View Synthesis. In *International Conference on 3D Vision (3DV)*, 2018.

[305] Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, September 2011.

[306] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision (ECCV)*, 2018.