Detecció de persones en imatges IR aplicada a l'àmbit del rescat

Oriol Rovira Vacas

Resum— Aquest projecte estudia el procés d'aplicació de la detecció de persones en imatges de càmeres tèrmiques, amb l'objectiu de proporcionar una prova de concepte per a la seva aplicació en l'àmbit del rescat d'emergència. Les càmeres tèrmiques són utilitzades de forma freqüent, ja que proporcionen grans avantatges en situacions de baixa visibilitat com pot ser el fum o la vegetació densa. Al llarg del projecte s'han realitzat diversos entrenaments basats en l'arquitectura YOLOv3 per a assolir un model robust en la solució del problema. S'han realitzat processos de transferència de coneixement entre diferents versions de models entrenats. També s'ha dut a terme un etiquetatge per habilitar datasets tèrmics originalment creats per altres aplicacions per a poder ser usats en casos de detecció d'objectes.

Paraules clau—Intel·ligència Artificial, Aprenentatge Automàtic, Aprenentatge Profund, Transferència de coneixement, Visió per Computador, Detecció d'objectes, Detecció de persones, CNN, YOLO, Darknet

Abstract—This project studies the proccess of applying person detection models in thermal imaging with the objective of providing a proof of concept for its aplication in the field of search and rescue. Thermal càmeres are being used frequently due to the advantatges they provide in low visibility settings like smoke or desne vegatition. Throughout this project, múltiple YOLO-based models have been trained in order to achieve a model accurate enough to solve the stablished problem reliably. Transfer Learning has been used between diferent trained model versions. There's also been a labeling process carried out in order to make certain thermal datasets, created for other purposes, be used for object detection cases.

Index Terms—Artificial Inteligence, Machine Learning, Deep Learning, Transfer Learning, Computer Vision, Object detection, Person Detection, CNN, YOLO, Darknet

1 Introducció

Automatitzar la detecció d'objectes suposa una eina increïblement útil per a multitud d'aplicacions. Des de l'inici del camp, el qual engloba la visió per computador i el processament d'imatges, s'han realitzat grans avenços pel que fa a tècniques i tecnologies utilitzades per a millorar la seva eficàcia.

La combinació de l'aprenentatge computacional (popularment conegut com a Machine Learning o ML) i la visió per computador és un àmbit que ja porta un bon nombre d'anys en alça, amb aplicacions funcionals sent ja emprades en molts àmbits del nostre dia a dia. Des del control de desperfectes i errors en cadenes de producció, fins als cotxes autònoms i la seguretat, l'ús de ML per al reconeixement d'objectes (i especialment persones i les seves característiques) és ja una realitat més que establerta.
Gràcies a la naturalesa de l'aprenentatge computacional, els models resultants poden ser aplicats a qualsevol format d'entrada (sempre que sigui una imatge i pugui ser redimensionada adequadament), per això proposem estudiar els beneficis de fer servir imatges infraroges (IR), les quals proporcionen una major qualitat d'informació en el problema concret de la detecció de persones en àmbits amb poca visibilitat. Com que les imatges IR capturen les diferències en temperatura, persones i animals acostumen a aparèixer perfectament contrastades amb el fons de la imatge gràcies al con-trast de temperatura entre el terra i la persona, facilitant enormement la tasca de detecció.

En aquest context, es planteja la utilització d'aquestes tècniques per al desenvolupament d'un model orientat a la detecció de persones en l'àmbit del rescat. L'escenari objectiu del nostre model és el de poder-lo aplicar en casos en les imatges convencionals perden eficàcia, com poden ser escenaris amb vegetació o fums que obstrueixen la línia de visió, afectant negativament la precisió dels models de detecció clàssics que utilitzen l'espectre visible de la llum.

Avui en dia els equips d'emergències ja fan servir càmeres tèrmiques (IR) per millorar la seva capacitat operativa en entorns de baixa visibilitat com pot ser un incendi d'un habitatge. Tot i això, les imatges obtingudes son de baixa qualitat, i això, juntament amb la representació en escala de grisos que s'utilitza per a la visualització, fan de la interpretació de les imatges de la càmera una tasca difícil en plena actuació a causa de la complexitat de manipular una d'aquestes càmeres amb una mà mentre s'executen les tasques d'intervenció. Creiem que proporcionar una identificació clara i ràpidament llegible sobre els elements detectats a l'entorn, permetrà millorar considerablement l'efectivitat d'aquestes càmeres tèrmiques, possiblement ajudant a detectar persones que d'altres maneres no hauríem pogut localitzar fàcilment

- *E-mail de contacte: oriolrovira96@gmail.com*
- *Menció realitzada: Enginyeria Computació*
- *Treball tutoritzat per: Robert Benavente Vidal (DCC)*
- *Curs 2021/22*

2 OBJECTIUS

En aquest s'analitza el problema de construir un model robust per a la detecció de persones en imatges IR. Per tal de ser capaç de prioritzar de manera correcta els requeriments i tasques del projecte, es divideixen els objectius entre principals i secundaris.

2.1 Objectius principals

Els objectius principals són la base essencial del projecte, els quals han de ser assolits amb prioritat màxima.

-Construir un model d'arquitectura YOLO capaç d'identificar persones en escenaris de baixa visibilitat mitjançant imatges IR amb un grau de fiabilitat superior al 70%.

-Realitzar una comparativa de resultats de diferents models per estudiar els beneficis de l'ús d'imatges tèrmiques en vers a imatges RGB en la detecció de persones.

-Obtenir un model prou robust per a detectar persones en els escenaris plantejats independentment de la perspectiva de la imatge (arran de terra/aèria).

2.2 Objectius secundaris

Els objectius secundaris del projecte o bé s'alineen i suplementen algun objectiu principal, o es tracta d'objectius interessants per a ampliar lleugerament l'abast del projecte, amb la voluntat de ser realitzats sempre que no comprometin el compliment dels objectius principals.

-Proporcionar possibles noves eines als cossos d'emergències per a la localització i rescat de víctimes.

-Estudiar i definir els requeriments per a la implantació d'un model de ML en un cas operatiu real.

-Desenvolupar una metodologia ben definida per al desenvolupament i entrenament de models per a la detecció de persones mitjançant xarxes ja existents.

-Identificar i analitzar els reptes per a la implementació d'un model per a la detecció d'objectes aplicat a un cas d'ús real.

-Potenciar els meus coneixements en els àmbits de ML i CV en un cas d'ús pràctic, sent capaç d'entrenar models per a tasques concretes partint d'arquitectures preestablertes.

3 ESTAT DE L'ART

Aquest apartat proporciona una introducció al context del projecte dins del camp de la visió per computador i la detecció d'objectes mitjançant aprenentatge màquina.

3.1 Detecció d'objectes

La detecció d'objectes és una tasca clau en la visió per computador utilitzada per a detectar instàncies visuals de diferents tipus d'objectes en imatges estàtiques o fotogrames de vídeo.
L'objectiu de la detecció d'objectes en imatges és respondre a les preguntes "Quins objectes hi ha a la imatge i a on es troben?".

La capacitat de detectar objectes és essencial per a moltes aplicacions de visió per computador com poden ser la segmentació, etiquetatge, tracking, detecció de text i nombres, etc.

La detecció d'objectes ha anat evolucionant ràpidament en els últims 20 anys gràcies als avenços en capacitat de computació. Generalment, es fa una divisió en l'evolució del camp entre abans i després de l'aparició del Deep Learning l'any 2014.

La detecció de persones forma part del camp de la detecció d'objectes, on la classe principal a detectar és la de "persona".

Dins dels algorismes de xarxes neuronals, es fa la divisió entre algorismes d'una etapa i algorismes de dues etapes.

3.1.1 Detectors de dues capes:

Aquest tipus de detectors aproximen les regions dels objectes fent servir característiques (features) de deep learning.

Primer es fa una proposta de regions d'interès on es pot haver detectat un objecte emprant mètodes convencionals de deep learning.
Posteriorment, es du a terme una classificació dels objectes basant-se en una extracció de les features de les regions proposades, tot realitzant la corresponent regressió per a la col·locació de la caixa delimitadora (bounding boxes) de l'objecte reconegut.

Aquests tipus de detectors proporcionen un alt nivell de precisió en la detecció, però l'existència de les dues capes i el nombre elevat de passos a realitzar per a cada imatge fa que siguin comparativament més lents.
L'última versió dels detectors de dues capes més popular és el G-RCNN[5].

3.1.2 Detectors d'una capa:

Els detectors d'una capa determinen les posicions de les bounding boxes saltant-se el pas de suggerir regions que fan els de dues capes.

Aquests detectors prioritzen la velocitat d'inferència, però sacrifiquen capacitat per a identificar múltiples objectes petits o objectes molt irregulars.

Els detectors d'una capa més populars actualment són el YOLO[6], SSD i RetinaNet

Gràcies a la gran velocitat d'aquests detectors d'una capa, han guanyat molta popularitat en aplicacions de reconeixement d'imatges en vídeo a temps real. Sent aplicats ja en

l'àmbit comercial en sectors com l'automoció, la logística i la seguretat. Els quals requereixen deteccions a temps real.

4 METODOLOGIA

En aquest apartat s'estableix la proposta d'entrenament a seguir, les eines i tecnologies utilitzades, i els datasets utilitzats per a la realització del projecte.

4.1 Proposta d'entrenament models detecció

Aquest projecte, al tractar-se d'un Treball de Fi de Grau, es veu limitat temporalment. Per tal de mantenir el desenvolupament dintre del període de temps establert, s'ha realitzat una planificació basada en sprints setmanals. Aquesta planificació pot ser visualitzada mitjançant el diagrama de Gantt recollit a l'apartat 3 de l'apèndix.

La proposta inicial es la de dur a terme entrenaments concatenats utilitzant les propietats del transfer learning per a escurçar els temps d'entrenaments totals. Partirem dels pesos del model de yolov3 entrenat sobre el dataset coco, que suposa el model de referència per a la xarxa utilitzat i inclou la classe persona.
A partir d'aquests pesos es realitzara un primer entrenament amb un dataset tèrmic inicial, el qual estudiarem i n'extreurem els resultats. Finalment, realitzar un entrenament amb un segon dataset tèrmic orientat a la baixa visibilitat i exclusiu a la detecció de persones. Aquesta proposta es basa en intentar establir una adaptació progressiva de la detecció de persones en RGB a la detecció de persones en entorns de baixa visibilitat, on no tota la silueta de la persona és sempre visible.

L'objectiu final sent el de generar un model robust per a la detecció de persones en imatges tèrmiques, tot i estudiant els límits de la transferència de coneixement.

4.2 Eines

Per tal de realitzar el desenvolupament de la manera més àgil possible i sense dependre d'un hardware en concret s'ha pres la decisió d'utilitzar Google Colab com a entorn d'execució del projecte. Aquest entorn proporciona un espai tipus notebook on executar comandes de bash fent servir els recursos computacionals que proporciona Google de manera gratuïta. Això permet usar hardware considerablement potent sense necessitat de comprar-lo, permetent així l'ús de CUDA per a executar el model a la GPU proporcionada.

L'únic límit que existeix en aquesta eina és un temps màxim d'execució continua de 12 hores, temps més que suficient per a les execucions del nostre model YOLO. En cas que els entrenaments del model sobrepassin aquesta franja de 12 hores, es dividirà l'entrenament en blocs, guardant els pesos entre sessions i continuant on s'hagi quedat.

Per a l'etiquetatge de datasets s'ha utilitzat YoloLabel, amb la qual s'ha pogut etiquetar ràpidament grans quantitats d'imatges. Aquest programari està disponible al repositori GitHub del seu creador i és d'ús lliure.

Al llarg del desenvolupament de la pràctica també s'ha utilitzat Roboflow, un servei en linea que proporciona eines de conversió de format per datasets. Aquesta eina però, tot i que a primera vista semblava molt útil, ha acabat suposant un mal de cap considerable, i el dataset en que l'hem utilitzat ha sigut descartat de cara a realitzar els entrenaments finals.

4.3 Detecció d'objectes mitjançant YOLO

El nom de YOLO ve de les sigles de "You Only Look Once", i es basa en l'ús de xarxes convolucionals per a realitzar detecció i classificació de múltiples objectes en una sola imatge, generant les corresponents bounding boxes i la seva posició relativa a la imatge.

El que fa especial a YOLO es que, a diferència de anteriors tècniques basades en classificació, YOLO aplica una sola xarxa neuronal a la imatge completa, la qual divideix en regions per a les quals prediu les corresponents probabilitats de que continguin una bounding box. Aquest mètode permet execucions molt ràpides, permetent processar imatges a fins a 30 FPS, sent viable per a aplicacions de video.

YOLOv3[15], la versió que serà utilitzada en la implementació de la part pràctica del treball, està formada per 53 capes CNN (Darknet-53[8]) i 53 capes addicionals per a tasques de detecció, resultant en un total de 106 capes que configuren YOLOv3.

Per potenciar la capacitat del model d'identificar objectes de diferents mides, s'analitzen les imatges en 3 diferents escales. Aquests redimensionats es realitzen a les capes 82, 91 i 106. Això es fa en funció d'un hiperparàmetre anomenat stride. Per tal que es pugui aplicar aquest redimensionat, les imatges han de ser divisibles per 32. En el cas de YOLOv3, l'input típic acostuma a ser el de 416x416. Tot i això, YOLO realitza un redimensionat de les imatges d'entrade a la resolució establerta, per tant no cal realitzar les transformacions abans de passar-les al model.
Aquesta resolució té un impacte directe en el nivell de detall que serà capaç de detectar el model, al igual que el cost computacional del mateix. El nivell de detall també es veu afectat per les subdivisions aplicades. Aquestes subdivisions, que han de ser una potència de dos, especifiquen el nombre de regions en les quals dividir la imatge i realitzar deteccions en cada regió.

Tot i que ja existeixen noves versions de YOLO (v5), YOLOv3 proporciona major accessibilitat i recursos gràcies al temps que porta establerta com a referent. Motiu principal per al qual la utilitzarem en aquest projecte.

4.4 Transferència de coneixement

La transferència de coneixement, anomenada Transfer Learning[18] en anglès, és el procés d'entrenar un model a

partir de pesos ja entrenats anteriorment amb altres dades. A través d'aquesta tècnica, el temps d'entrenament es pot reduir considerablement. Per exemple, si volem que una xarxa sigui capaç de detectar gossos de la raça Pastor Alemany, si es comença l'entrenament a partir d'un model ja capaç de detectar gossos, aquesta xarxa ja comença l'entrenament sabent detectar les característiques generals d'un gos, i només haurà d'aprendre a distingir les concretes als Pastors Alemanys. Permetent així començar l'entrenament amb una loss molt menor a fer-ho amb pesos aleatoris.

Amb aquesta transferència de coneixement, s'espera obtenir un model capaç de detectar persones en imatges IR de manera consistent, utilitzant com a punt de partida la xarxa de demo de YOLOv3 proporcionada per darknet, que ha estat entrenada sobre el dataset COCO.

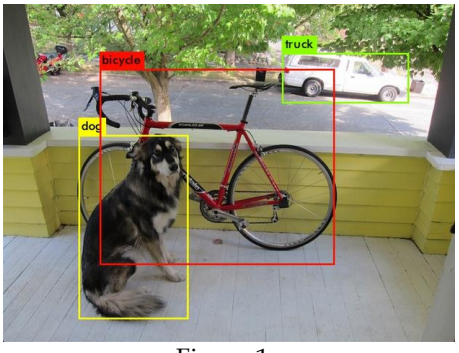


Figura 1:
Imatge de proba del dataset COCO amb prediccions del model YOLOv3 de mostra per aquest dataset.

Aquest model de YOLOv3 detecta 80 classes diferents, desde objectes quotidians fins a diferents vehicles i animals. Sent capaç de detectar objectes amb una precisió mitjana del 52% a 45 FPS.[15]

4.5 Datasets

Depenent de la xarxa i el framework utilitzats, certs datasets seran més adequats que altres. Per assolir els requeriments del cas d'ús objectiu, les dades a utilitzar han de complir dos requeriments principals. Que siguin imatges tèrmiques on apareguin persones, i que hi hagi un cert grau d'ofuscació sobre les siluetes de les persones sigui per vegetació o fum. Cal tenir en compte que en les imatges tèrmiques el fum no és visible, per tant, en aquest sentit no és absolutament necessari per a l'entrenament dels models, però si molt valuós per a la valoració dels resultats de cara als resultats dels pesos originals.

Per altra banda, existeix la divisió entre imatges a nivell de terra o imatges aèries. Per a l'aplicació plantejada al projecte, la idea principal és el cas d'ús del POV, tot i que s'intentarà explorar el fet de crear un model capaç de funcionar en ambdós escenaris.

ADAS-FLIR
El dataset 'ADAS' proporcionat per FLIR[10] a través de la seva pàgina web i de lliure ús per a propòsits educatius serà utilitzat per als entrenaments preliminars a mode de dataset de proba. Aquest dataset està compost per imatges fetes des de un cotxe simultàniament amb una càmera normal i una infraroja, fet que proporciona un escenari ideal per a la comparació entre models RGB i IR. El dataset disposa de classes etiquetades en el format JSON.

Figura 2:
Imatge IR etiquetada del dataset ADAS-FLIR

El dataset inclou etiquetes tant per persones com per animals, al igual que cotxes i altres classes relacionades amb la conducció autònoma, proporcionant etiquetes per a 15 classes diferents en total. Per als entrenaments que realitzarem aquestes classes no rellevants seran descartades ja que no aporten valor en els casos d'ús objectius d'aquest projecte.

Aquest dataset no ha estat utilitzat en els entrenaments finals degut al baix grau d'usabilitat del conjunt de dades posterior al processament mitjançant Roboflow per habilitar-lo, fent-lo un mal candidat per als entrenaments.

AAU-PD-TP
El dataset 'AAU-PD-TP' està disponible a Kaggle[12], i consta de diversos grups d'imatges de diferents característiques. Totes les imatges IR sent de càmeres aèries. Aquest dataset utilitza tècniques de Data augmentation, tot i que l'autor no especifica quines tècniques ha utilitzat.

El dataset ja ve etiquetat en format Darknet, amb cada conjunt d'imatges ja diidit entre test i traint.

Per al entrenament s'han utilitzat les imatges dels conjunts Far_viewpoint, Good_condition, Low_resolution, Occlusion, Shadow, Similar_temperature i Wind. Aquests conjunts equivalen a un total de 1200 imatges.

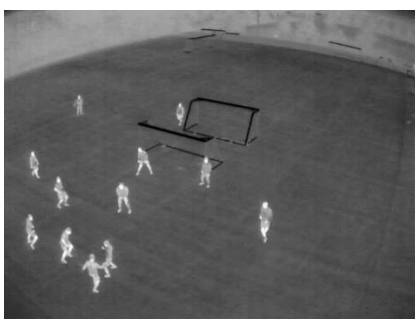


Figura 3:

Imatge IR del dataset AAU-PD-TP

Els altres conjunts del dataset han estat descartats degut a la baixa qualitat de les dades d'aquests, sent massa diferents al cas d'ús objectiu del model a entrenar.

IMEThermal
Aquest dataset creat per l'Instituto Militar de Engenharia de Rio de Janerio[11]. Consta de diferents gravacions amb una càmera tèrmica i una en escala de grisos de persones entre vegetació i fum. El dataset en el seu format original no està preparat per a ser utilitzat per a detecció d'objectes, ja que va ser creat amb l'objectiu d'estudiar la fusió d'imatges entre l'espectre visible i l'infraroig, el qual no requereix etiquetatge.

Figura 4:
Imatge IR(esquerra) i Grayscale(dreta) del dataset 'IMEThermal'

Ja que suposa un dataset pràcticament únic en especificacions ja que compleix exactament els requeriments establerts per el projecte, s'ha realitzat el procés d'etiquetatge d'una part d'aquest (take_1, take_3 i take_4) de forma manual.
El dataset resultant està composat d'unes 3800 imatges IR i les corresponents imatges en escala de grisos per a validació amb models no-IR.

El procés d'etiquetatge ha estat relativament senzill gràcies a l'eina YoloLabel, ja comentada a Metodologia. Trigant al voltant de 8 hores de feina en tenir etiquetat tot el dataset que s'ha confeccionat.

Combined Dataset
Al llarg del desenvolupament, s'han agrupat els datasets AAU-PD-TP i IMEThermal per a crear conjunts d'entrenament i test que continguin imatges tant de vegetació a nivell de terra com aèries.

Construint aquest dataset disposarem de dades properes al cas d'ús objectiu, i comprobar la versatilitat del model respecte als diferents punt de vista en els que volem que operi.

Aquest dataset també ens proporciona l'oportunitat de comprovar l'efectivitat del transfer learning per als diferents punts de vista, al poder comprovar els resultats d'aquest en vers a models més especialitzats en un punt de vista en concret.

5 RESULTATS

5.1 Paràmetres de configuració

Existeixen molts paràmetres de configuració a tenir en compte en l'entrenament de models d'apranantatge màquina. No existeix una configuració d'entrenament universal per a qualsevol model i dataset.

Per a determinar quan finalitzar un entrenament hem utilitzat un nombre d'iteracions proporcional al nombre de classes a entrenar. Seguint les indicacions dels propis desenvolupadors de Darknet, la norma general que s'utilitza es la de 2000*n^{o} classes. De la mateixa manera, els filtres de les capes prèvies a les 3 capes yolo han de ser modificats per a que siguin proporcionals

Una loss per sota de 0.005 és considerada indicativa d'un model sobreentrenat, per tant si durant l´entrenament s'observa una estabilització de la loss per sota d'aquest llindar finaliztarem l'entrenament.
Max_batches defineix el nombre màxim d'iteracions del procés d'entrenament.

5.2 Mètriques

Per tal de poder visualitzar els resultats dels diferents models, s'utilitzaran diferents mètriques, ja establertes com a referents dintre del framework de darknet. Aquestes mètriques són l'IOU (Intersection Over Union), loss i mAP (mean Average Precision). A l'apartat 1 de l'apèndix es pot trobar una explicació detallada sobre aquestes mètriques.

Al llarg de l'entrenament s'enmagatzemen els valors de Class_Loss, IOU_Loss i Total_Loss de cada iteració, al igual que es realitza el càlcul de mAP cada 4 Epochs (1000 iteracions). El valor de mAP s'obté a partir de realitzar un test de validació, per tant es equivalent a realitzar-lo en la versió final dels pesos de la xarxa. Tot i així, poder visualitzar l'evolució del mAP al llarg de l'entrenament pot ajudar a seleccionar versions intermitges de l'entrenament que siguin més robustes que pesos posteriors que son més propensos a ser overfitted. Finalment, per evaluar els resultats de les validacions realitzades s'utilitzarà la precision i el recall i així poder visualitzar la qualitat del model respecte al dataset IMEThermal que es el més proper al cas d'ús objectiu.

Aquestes mètriques ens permenten inferir sobre la qualitat de la xarxa resultant tant al llarg del procés d'entrenament (per evitar finalitzar entrenaments fundamentalment erronis) o detectar overfitting o underfitting al model resultant de manera ràpida.

Les gràfiques dels resultats han estat realitzades mitjançant la llibreria matplotlib de python a partir de l'ouptut dels diferents processos d'entrenament i test de cada model.

5.3 Primer Model: IMEOnly

S'ha realitzat un entrenenament inicial mitjançant el dataset IMEThermal a partir dels pesos coco originals.

El model ha estat entrenat durant 6000 iteracions, procés que dura al voltant de 6 hores, modificant els paràmetres del fitxer de configuració per a establir subdivisions=16 i una sola classe a detectar.

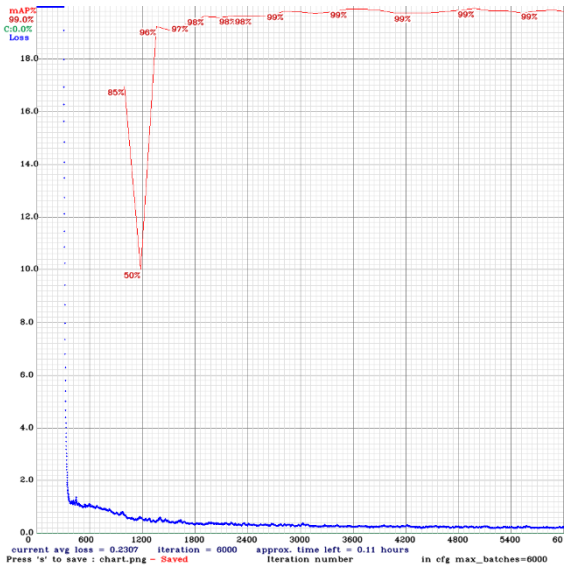


Figura 5:
Resultats de l'entrenament del segonmodel mostrant avg. loss i mAPal llarg de les iteracions.

Els resultats del model són aparentment molt bons, assolint una mAP de 99%. Tot i que una mAP tant alta pot ser indicvativa d'overfitting, també pot ser a causa de la poca diferència entre les imatges del conjunt de test i train. Més endavant s'estudiarà en major profunditat la qualitat del model tot i comparant-lo amb les altres versions.

5.4 Segon Model: AAUOnly

El segon model ha estat entrenat amb el dataset AAU-PD-TP, seguint els mateixos procediments i configuració que en el Primer Model.

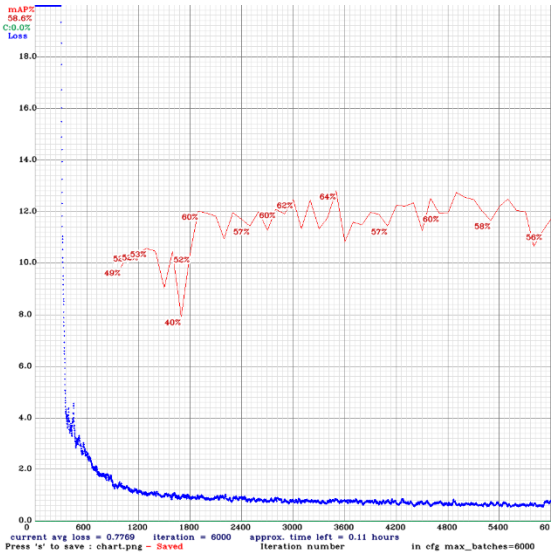


Figura 6:
Imatge de proba de Darknet amb prediccions del model COCO.

Els resultats de l'entrenament són relativament bons, amb la loss disminuint adecuadament, però amb la mAP estancant-se al voltant del 60%. Aquesta baixa accuracy s'ha atribuït al fet de que les imatges d'aquest dataset, al ser aèries i amb grups de múltiples persones, amb l'atribut de configuració de subdivisions sent 16 pot ser que no sigui capaç de detectar les features corresponents a cada individu de manera eficaç.

Per tal de comprovar si es així, sería interessant realitzat un altre entrenament augmentant les subdivisions 32. Cal tenir en compte però l'impacte de les subdivisions sobre el temps d'execució del model final, ja que l'objectiu es que sigui aplicat sobre vídeo en directe, i masses subdivisions poden augmentar ràpidament la complexitat temporal del model.

5.5 Tercer Model: AAU_IME

El tercer model es el resultat de reentrenar el model primer model (entrenat prèviament sobre AAU-PD-TP) i ara entrenat mitjançant el dataset IMEThermal.

En aquest cas s'ha realitzat un entrenament de només 2000 iteracions, ja que al tractar-se del model que realitza dues transferències de coneixement el procés de convergència hauria de ser menor. L'entrenament s'ha realitzatseguint els mateixos principis i configuració que amb el primer model.

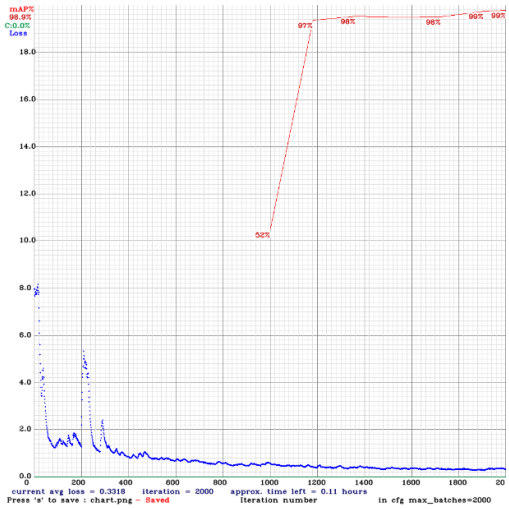


Figura 7:
Resultats de l'entrenament del tercer model mostrant avg. loss i mAPal llarg de les iteracions.

Els resultats de l'entrenament mostren una molt bona mAP, alhora que es pot apreciar la baixa loss desde l'inici de l'entrenament gràcies al transfer learning realitzat.

L'augment de la loss en un pic irregular al voltant de les 200 iteracions sembla indicar que algunes dades utilitzades són defectuoses. Després de comprovar-ho en profunditat, s'han localitzat un total de 10 imatges del dataset IMEThermal les quals no van ser etiquetades correctament, raó per la qual podria aparèixer aquesta los irregular.

5.6 Quart Model: Combined

Finalment, per tal de comprovar si el procés de Trasnfer Learning en múltiples etapes és o no és efectiu, s'ha realitzat un últim model entrenat a partir dels pesos del model base mitjançant un dataset combinat. Creat a partir dels datasets AAU-PD-TP i IMEThermal.

L'entrenament a sigut realitzat utilitzant els mateixos paràmetres de configuració que en tots els altres entrenaments.

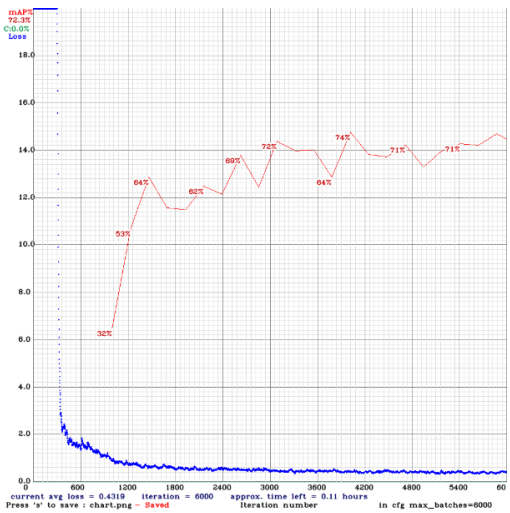


Figura 8:
Resultats de l'entrenament del segonmodel mostrant avg. loss i mAPal llarg de les iteracions.

Els resultats de l'entrenament mostren una mAP màxima del 74%. Aquest valor suposa una millora sobre la detecció respecte el primer model, però en canvi presenten un empitjorament considerable respecte els resultats dels models que han entrenat sobre IMEThermal individualment. Com que els resultats relacionats amb el dataset IMEThermal generen valors de mAP molt alts,com ja s'ha mencionat, existeix la possibilitat de que sigui a causa de overfitting.

En primera instància, aquest quart model podria ser més robust de cara a treballar en escenaris diferents, però cal realitzar més proves de detecció per a confirmar-ho.

5.7 Comparativa global

Per tal d'assolir el millor model possible dintre de l'abast del projecte s'han realitzat una sèrie d'entrenaments addicionals als establerts inicialment. Aquests entrenaments addicionals s'han basat en intentar comprovar la utilitat real del transfer learning en el cas plantejat. Per això s'han realitzat entrenaments individuals per a cada dataset, sense utilitzar els pesos resultants d'un per a l'entrenament de l'altre.

Com que el dataset Combinat està format per imatges extretes dels altres dos datasets, aquesta comparació ens serveix per inferir l'efectivitat dels models respecte a casos aèris o casos a nivell de terra, ja que aquesta es la principal diferència entre les dades dels datasets principals.

En les figures 9,10 i 11 mostrades a continuació es recullen els resultats de detecció de cada model sobre els conjunts de validació dels datasets utilitzats.

	IMEThermal		
Models	F1-score	mAP	avg IOU
IMEOnly	0.98	0.99	79.36
AUUOnly	0.04	0.009	1.45
AUU_IME	0.98	0.98	76.48
Combined	0.98	0.99	77.9

Figura 9:
Resultats del test de tots els models sobre dataset IMEThermal

	AAU-PD-T		
Models	F1-score	mAP	avg IOU
IMEOnly	0.02	0.01	21.79
AUUOnly	0.72	0.63	53.24
AUU_IME	0.01	0.08	62.44
Combined	0.68	0.59	49.56

Figura 10:
Resultats del test de tots els models sobre dataset AAU-PD-T

	Combined		
Model	F1-score	mAP	avg IOU
IMEOnly	0.50	0.35	75.91
AUUOnly	0.27	0.35	13.05
AUU_IME	0.49	0.40	76.27
Combined	0.79	0.737	59.59

Figura 11:
Resultats del test de tots els models sobre dataset Combined

Observant els resultats, tal com s'esperava, s'aprecia una diferència considerable entre els resultats d'IMEThermal i els resultats per a AUU-PD-T. Els mals resultats del Model AUUOnly per a IMEThermal són consistents amb el que s'esperava, ja que el model no ha entrenat amb cap imatge d'aquell dataset. Passa el mateix amb el IMEOnly sobre el dataset AUU-PD-T.

Analitzant els resultats de l'últim test, podem afirmar que l'entrenament utilitzant el dataset combinat ha generat un millor model resultant que el transfer learning per etapes realitzat al AUU_IME.

El Model AAU_IME té mals resultats per al dataset AUU-PD-T, mentre que el test sobre IMEThermal és equiparable als resultats dels altres models.
Aquest comportament però, no és únic per al Model AUU_IME. Els resultats de tots els models sobre el dataset AUU-PD-T són molt pitjors que amb les dades d'IMEThermal, amb la màxima mAP sent l'assolida per el Model AUUOnly (exclusivament entrenat sobre les dades).

Tot i que és possible que els resultats per AAU-PD-T siguin conseqüència de males dades fruit de les tècniques de data augmentation utilitzades en la construcció del dataset, creiem possible que es tracti d'un problema amb les subdivisions realitzades a les imatges, com ja s'ha comentat a l'apartat corresponent de l'entrenament del Model AUUOnly. Es possible que les imatges aèries amb múltiples persones requereixin de un major nombre de subdivisions per tal d'augmentar la seva percepció al detall i així poder detectar objectes més petits. Alternativament, també es podria augmentar el tamany d'entrada de les imatges, efectivament augmentant l'àrea de les subdivisions.

Finalment, podem afirmar que el Model Combinat és el model més robust dels generats, sent capaç de realitzar deteccions en imatges tant aèries com a nivell de terra amb una F1-score de 0.79. Aquest valor és realment bo, superant els valors dels pesos originals sobre el dataset COCO. Això es un bon indicatiu de que l'aplicació objectiu en vídeo és viable.

S'han realitzat visualitzacions de deteccions utilitzant tots els models per a diferents escenaris. Es pot veure una petita mostra d'aquestes deteccions a l'apartat 1 de l'apèndix. En aquestes deteccions es pot observar el que mostren els resultats numèrics. Els models entrenats tenen dificultats a detectar persones en diferents punts de vista de manera consistent, generant falsos positius quan es busquen features aèries (petites) en imatges a nivell de terra, que contenen siluetes genarlment més grans.

Tot i així cal mencionar que el model Combinat és capaç de realitzar deteccions amb un bon grau de confiança en ambòs punts de vista, aquests resultats segurament serien facilments millorables si s'utilitzessin múltiples classes per a diferents distàncies, o potencialment amb entrenaments amb major quantitat i qualitat de dades.

6 Conclusió

Al llarg d'aquest projecte s'han entrenat múltiples models de YOLOv3 orientats a la detecció de persones. El model Combinat final ha assolit una mAP del 73.7% en la detecció de persones en entorns de baixa visibilitat. Aquest model pot ser aplicat tant en imatges com en vídeo, i es capaç d'identificar persones entre vegetació amb una precisió superior a un detector de persones equivalent en imatges RGB.

El model resultant demostra la viabilitat de la detecció de persones en imatges tèrmiques en escenaris de baixa visibilitat, amb una precisió equiparable a models per a la detecció de persones sense ofuscació.

Per assolir aquest model final s'han realitzat diversos entrenaments amb 4 datasets diferents, acompanyats del corresponent estudi de resultats per a corretgir el procés d'entrenament. Considero que els objectius principals del projecte han estat tots assolits.

Respecte als objectius secundaris, han estat complerts en la seva majoria, tot i que no he pogut aprofundir tot el que hauria volgut en la integració del model en un cas d'ús real. Tanmateix, queda molt espai de millora de cara a millorar el model, gràcies principalment als coneixements obtinguts al llarg de la resolució d'aquest projecte.

El desenvolupament ha estat una molt dinàmic, ja que en certes etapes he aprés coses que m'han fet veure millores possibles en passos anteriors, cosa que m'ha fet reorganitzar l'enfoc del desenvolupament uns quants cops. Un clar exemple es l'ús del transfer learning. Malgrat la seva gran utilitat en entrenaments, buscar crear un model centrat en detectar una sola classe a partir de models amb gran quantitat de classes resulta no ser tant efectiu com creia. Tot i així, estic satisfet amb el model resultant i els coneixements adquirits al llarg de l'elaboració del projecte.

De cara a continuar el desenvolupament, en un futur m'agradaría explorar la utilització d'una xarxa més petita, com podria ser la pròpia tiny-yolov3, ja que crec que es podrien obtindre resultats semblants amb un cost temporal i

espacial encara menor.
Un altre camí per a millorar les deteccions entre punt de vista aèri i punt de vista a nivell de terra sería la creació de diferents classes per a persones properes i llunyanes, permetent al model realitzar una diferenciació més clara entre features.

Un punt a ampliar que portaría un gran benefici al model de detecció de persones sería ampliar en quantitat i qualitat de datasets més diversos i així millorar l'accuracy del model en altres escenaris. Especialment en el cas d'imatges amb punts calents causats per foc.

Finalment, sería molt interessant explorar l'objectiu original del projecte en la seva concepció, en el qual planejava explorar la implementació del desplegament del model en una càmera tèrmica real.

AGRAÏMENTS

Primer de tot m'agradaria agrair el suport i confiança que els meus pares i la meva germana han tingut en mi al llarg del desenvolupament d'aquest projecte, però també a través de tots els anys de carrera. Gràcies per donar-me ànims quan més ho necessitava, no ho hauria pogut fer sense vosaltres.

També voldria agrair al tutor del projecte, en Robert Benavente, pel seu mentoratge i suport, sense el qual el resultat final d'aquest article seria indubtablement pitjor. Moltes gràcies per totes les reunions i feedback proporcionat en tots els punts del desenvolupament i redacció de l'informe.

BIBLIOGRAFIA

[1] Boesch, Gaudenz, "Object Detection in 2022: The Definitive Guide", viso.ai , 2022, [Online]. Avaliable: https://viso.ai/deep-learning/object-detection/

[2] Sayantini. "Keras vs TensorFlow vs PyTorch : Comparison of the Deep Learning Frameworks", Edureka!, 2021, [Online]. Avaliable: https://www.edureka.co/blog/keras-vs-tensorflow-vs-pytorch/

[3] Adam Van Etten, "You Only Look Twice: Rapid Multi-Scale Object Detection in Satellite Imagery" CosmiQ Works, 2020, [Online]. Avaliable: https://www.arxiv-vanity.com/papers/1805.09512/

[4] Alex Kirzhevsky, Ily a Sutskever, Georey E Hinton, "ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information PRocessing Systems", NIPS, 2012, [Online]. Avaliable: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[5] Cheng, Lin, Zijiang Yang, "GRCNN: Graph Recognition Convolutional Neural Network for Synthesizing Programs from Flow Charts.", 2020, https://arxiv.org/abs/2011.05980

[6] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement.", 2018, [Online]. Avaliable: https://arxiv.org/abs/1804.02767

[7] "Darnket53", The Mathworks [Online]. Avaliable: https://es.mathworks.com/help/deeplearning/ref/darknet53.html

[8] Afroz Chakure, "All you need to know about YOLOv3", dev.to, 2020, [Online]. Avaliable: https://dev.to/afrozchakure/all-you-need-to-know-about-yolo-v3-you-only-look-once-e4m

[9] FLIR, "ADAS dataset", Teledyne FLIR, 2013, [Online]. Avaliable: https://www.flir.com/oem/adas/adas-dataset-form/

[10] SMT/COPPE/Poli/UFRJ, IME, "Visible-Infrared Database", CAPES/Pró-Defesa Program, 2021, [Online]. Avaliable: http://www02.smt.ufrj.br/~fusion/

[11] Huda, "AAU-PD-T", Kaggle Dataset, 2020, [Online]. Avaliable: https://www.kaggle.com/datasets/noorulhuda90/aaupdt?resource=download

[12] A.Bañuls, A.Mandow, R.Vàzquez-Martin, J.Morales, A.Garcia-Cerezo, "Object Detection from Thermal Infrared and Visible Light Cameras in Search and Rescue Scenes", IEEE International Symposium on Safety, Security and Rescue Robotics, 2020, [Online]. Avaliable: https://www.uma.es/media/files/SSRR_2020_Object_detection_TIR_and_Visible_light_in_Search_And_Rescue_Accepted_version.pdf

[13] Nitin Tiwari, "YOLOv3 Custom Object Detection with Transfer Learning", TFUG Mumbai Weekly , 2020, [Online]. Avaliable: https://tiwarinitin1999.medium.com/yolov3-custom-object-detection-with-transfer-learning-47186c8f166d

[14] Joseph Redmon, "Darknet: Open Source Neural Networks in C", 2022 , [Online]. Avaliable: https://pjreddie.com/darknet/

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 2015, https://arxiv.org/abs/1506.02640

[16] Stephen Oni, "Understanding Anchors(backbone of object detection) using YOLO" , Becoming Human, 2020, [Online]. Avaliable: https://becominghuman.ai/understanding-anchors-backbone-of-object-detection-using-yolo-54962f00fbbb

[17] Palmero C., Clapés A., Bahnsen C., Møgelmose A., Moeslund T., Escalera S, "Multi-modal RGB–Depth–Thermal Human Body Segmentation.", International Journal of Computer Vision, 2016, [Online]. Avaliable: https://www.researchgate.net/publication/301313195_Multi-modal_RGB-Depth-Thermal_Human_Body_Segmentation

[18] V. Roman, "CNN Transfer Learning & Fine Tuning", Medium, 2020, [Online]. Available: https://towardsdatascience.com/cnn-transfer-learning-fine-tuning-9f3e7c5806b2

APÈNDIX

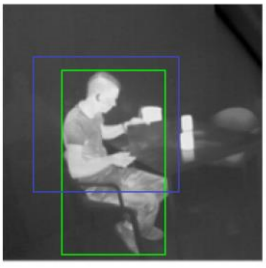
A1. MÈTRIQUES

A continuació es recullen els resultats de diferents deteccions realitzades per cadascun dels quatre models entrenats.

IoU (Intersection over Union)

L'IoU mesura el solapament entre dos bounding boxes, S'utilitza per a mesurar com de bona es la bounding box predita.

$$\text{IoU} = \frac{\text{Àrea solapada}}{\text{Àrea Unió}}$$



mAP (mean Average Precision)

Aquesta mètrica correspon a l'area sota la corba de precision-recall obtinguda de realitzar diverses deteccions sobre un conjunt de dades no vist en l'entrenament del model. La fórmula per a calcular l'Average Precision es la següent:

$$\text{AP} = \int_0^1 p(r)dr$$

Per tant, mAP es tracta de computar la mitjana d'APs per a cada imatge del conjunt de validació.
Com que precision i recall sempre es troben entre 0 i 1, l'AP i l'mAP també seran valors entre 0 i 1.

Loss

Loss és la mètrica utilitzada per al model per a penalitzar les males prediccions. És un nombre indicatiu de com de malament està funcionant el model, i és el valor que s'intenta minimitzar en el procés d'entrenament. Existeixen moltes funcions diferents per tal de calcular la loss d'un model, amb la MSE sent l'exemple més bàsic.

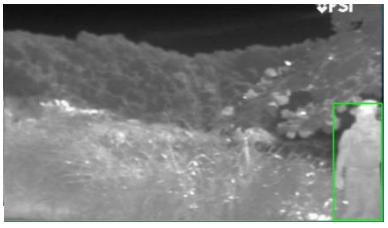
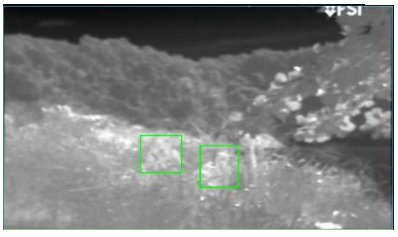
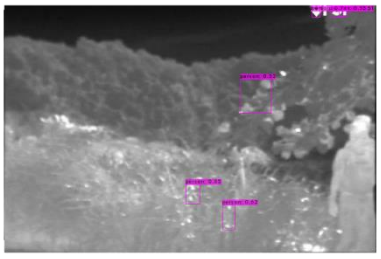
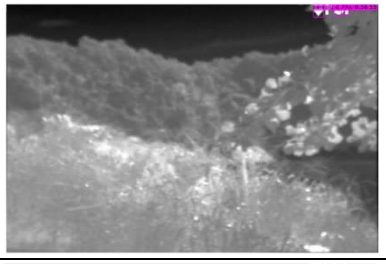
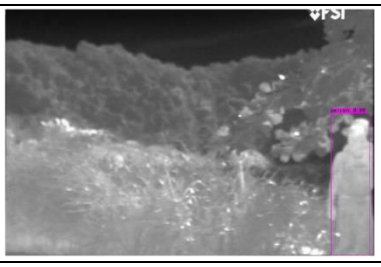
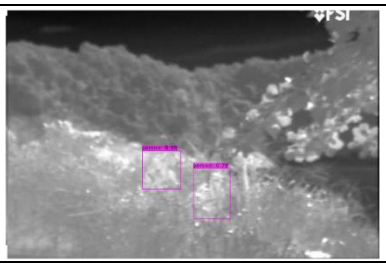
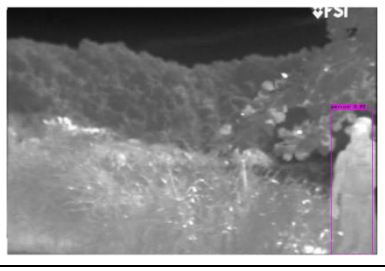
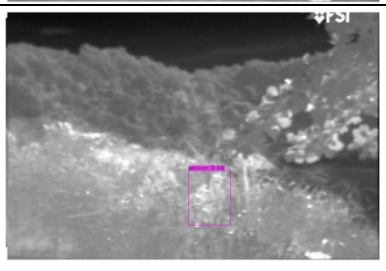
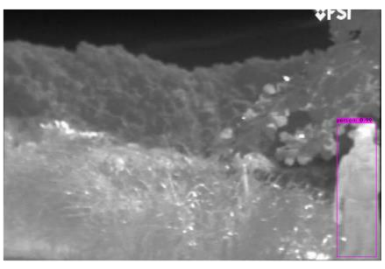
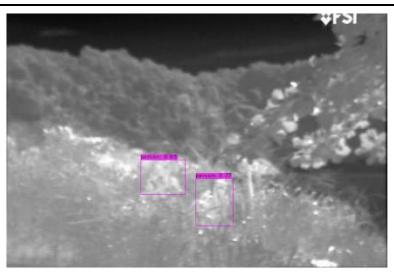
En el cas de YOLOv3, s'utilitza Binary Cross Entropy, la qual computa tres tipus de loss diferents per a analitzar diferents aspectes del model:

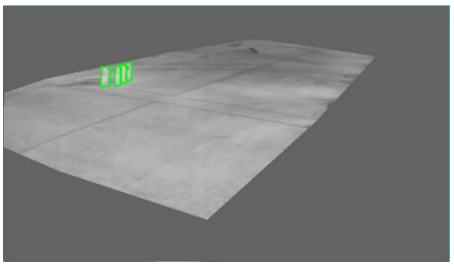
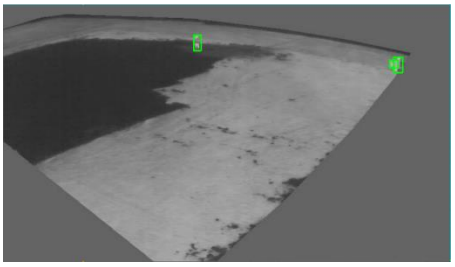
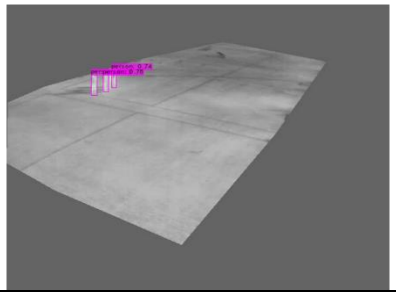
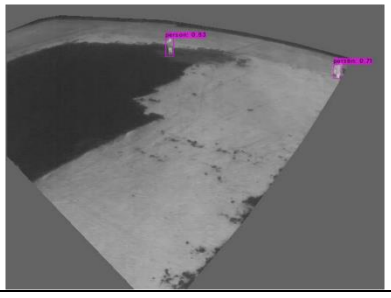
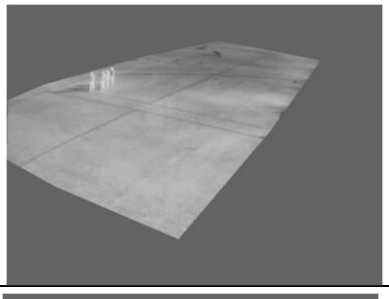
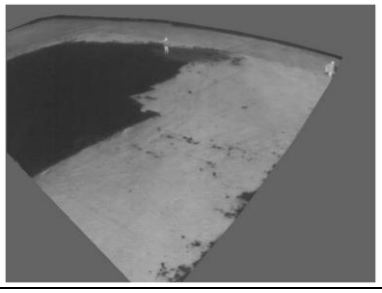
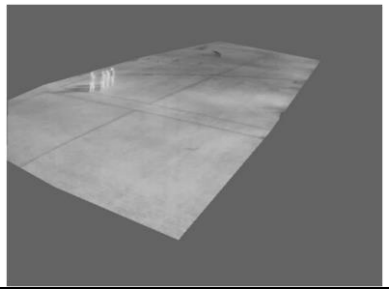
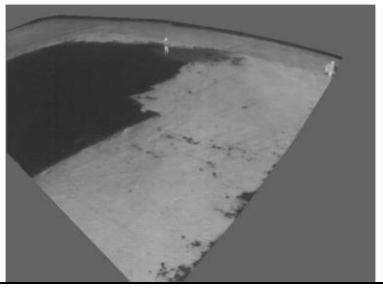
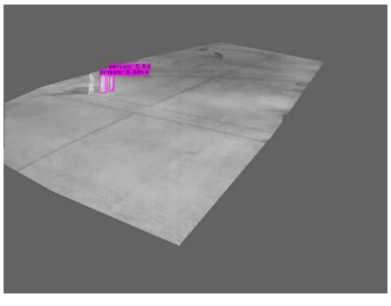
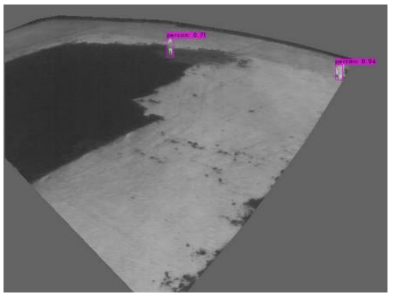
-Classification loss (classe correcta)
-Localization loss (errors de bounding box)
-Confidence loss (error en la identificació de les features)

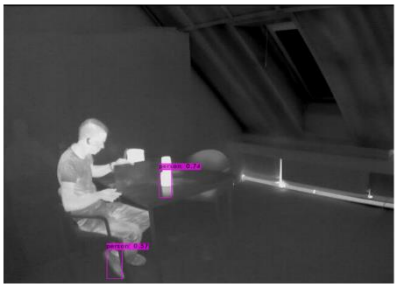
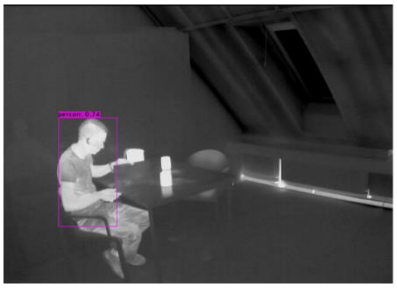
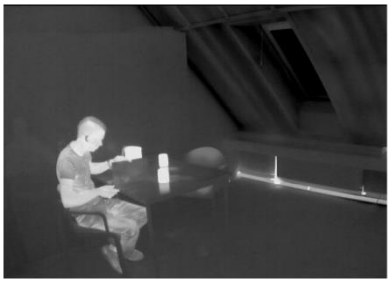
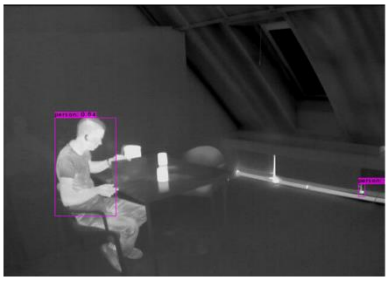
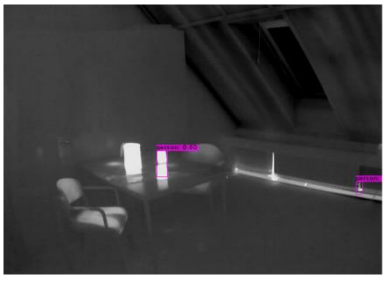
Aquestes tres mètriques son ajuntades per a obtenir la loss total del model.

A2. MOSTRA DE RESULTATS

A continuació es recullen els resultats de diferents deteccions realitzades per cadascun dels quatre models entrenats.

	IMEThermal test	
	Take4_IR_1406	**Take4_IR_1283**
Groundtruth		
AAUOnly		
IMEOnly		
AAU_IME		
Combined		

	AAU-PD-T test	
	Image_979	**Image_777**
Groundtruth		
AAUOnly		
IMEOnly		
AAU_IME		
Combined		

	Alternative test: AAU VAP (Kaggle)	
	Image_979	**Image_777**
Groundtruth		
AAUOnly		
IMEOnly		
AAU_IME		
Combined		

A3. PLANIFICACIÓ

Diagrama de Gantt utilitzat per a la planificació del desenvolupament del projecte.