
This is the **published version** of the bachelor thesis:

Arenas Pérez, Joaquín. *Predicting human eye fixations and aesthetic preferences from individualized DNNs*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Intel·ligència Artificial)

This version is available at <https://ddd.uab.cat/record/336315>

under the terms of the  license.

Predicting Human Eye Fixations and Aesthetic Preferences from Individualized DNNs

Joaquín Arenas Pérez

June 28, 2026

ABSTRACT

This project investigates whether individualized Alignment Importance Score (AIS) heatmaps predict human ocular fixations during aesthetic evaluation better than Grad-CAM and Saliency Maps. A VGG-16 network fine-tuned on AVA (74.98% accuracy, $AUC = 0.755$) generated heatmaps for 200 WikiArt paintings which are compared against eye-tracking data from four participants. Contrary to what was initially thought, AIS did not outperform Grad-CAM at predicting fixation location: Grad-CAM achieved the highest correlation ($\rho = 0.089$), followed by saliency ($\rho = 0.063$) and AIS ($\rho = 0.057$). This pattern held consistently across all participants, suggesting that Grad-CAM's intrinsic, image-specific salience aligns more closely with gaze behavior than AIS's relational importance in the cohort of subjects. Results are interpreted with caution given the small cohort ($n = 4$).

Keywords: Aesthetic perception, Eye-tracking, CNNs, Explainable AI (XAI), Alignment Importance Score (AIS), Grad-CAM, Saliency maps, Representational Similarity Analysis (RSA), VGG-16, Computational aesthetics



1 INTRODUCTION AND CONTEXT

Aesthetic perception is one of the most distinctive human capacities, consisting basically of our ability to judge whether an object is beautiful or artistically valuable. It involves complex cognitive and emotional processes that go beyond normal object recognition and includes many subjective and even unconscious structures that cannot be easily explained or described. In this work, I focus on aesthetic perception regarding images, specifically paintings. Understanding which visual features are decisive in this judgment is an open question in aesthetics studies, computer vision, and perceptual psychology. To have a wide vision, we can classify those features as low level (color, texture, contrast, composition) and high level (semantics, context, emotion).

Is it possible for modern computer science to approximate a prediction of those aesthetic judgments? In recent years, deep learning has demonstrated a remarkable ability to learn hierarchical image representations, from edges and textures in the early layers to semantic concepts in the deeper layers. This representational hierarchy offers

a unique opportunity for trying to study which layers and which features are most relevant for predicting human aesthetic preference.

This project sits at the intersection of image aesthetic quality assessment, the explainability of deep learning models, and the analysis of visual features. Specifically, the project aims to study the role played by low-level features (extracted from the early convolutional layers) and high-level features (deeper layers) in predicting human aesthetic judgment, and compares three types of explainability heatmaps against actual ocular fixation data collected from human participants in real-life experiments. Those three heatmaps are Saliency Heatmaps, Grad-CAM Heatmaps, and individualized Alignment Importance Score (AIS) heatmaps.

2 STATE OF THE ART

2.1 Aesthetic Preference for Art can be Predicted from a Mixture of Low- and High-Level Visual Features

Iigaya et al. (2021) [2] develop the Linear Feature Summation (LFS) model, which represents the subjective value of an artwork as a weighted sum of low-level features (color, saturation, brightness, texture, and composition) and high-

- Contact E-mail: 1673338@uab.cat
- Supervised by: Daniel Pacheco Estefa
- Academic Year 2025/26

level features (concreteness, dynamism, chromatic temperature, and emotional valence). Using data from both in-lab participants and a large-scale Amazon Mechanical Turk sample with over 1,300 participants, they show that the LFS model successfully predicts aesthetic ratings both within and across individuals, even for previously unseen stimuli.

One of the article’s most important contributions is the analysis of VGG-16 pretrained on ImageNet, which is adopted for this work. VGG-16 is a model for image classification that is not explicitly trained on aesthetic judgment. Nevertheless, the paper demonstrates that VGG-16 can predict human ratings, and that its internal layers spontaneously encode the LFS features in a hierarchical manner. It is shown that low-level features emerge in the early layers, while high-level features such as concreteness and dynamism are represented more robustly in the deeper layers.

2.2 Explaining Human Comparisons Using Alignment Importance Heatmaps

Truong et al. (2025) [3] introduce the Alignment Importance Score (AIS), a metric that quantifies each feature map’s unique contribution to the alignment between the network’s representational geometry and that of humans. The idea is that the more the alignment decreases upon removing a given feature map, the higher its AIS.

This method is validated along two dimensions. First, selecting only the highest-AIS feature maps improves out-of-sample prediction of human similarity judgments. Second, the authors generate image-specific heatmaps by weighting feature maps according to their AIS and projecting them back onto image space. These visualizations indicate which regions of an image are most important for comparison within a given set.

Crucially, the heatmap is context-dependent: it changes depending on which other images are included in the comparison set, because AIS is computed relationally rather than intrinsically. A given region of a painting may be highly salient when compared against one set of images, yet contribute little when the comparison set changes. This relational property stands in contrast to standard saliency methods (e.g., Grad-CAM), which typically treat an image in isolation and produce a fixed map regardless of comparative context.

3 OBJECTIVES OF THIS PROJECT

- **Main Objective:** To determine whether ocular fixations during aesthetic assessments are better predicted by individualized AIS heatmaps than by visual saliency maps or Grad-CAM heatmaps.
- **Sub-objective 1:** Train a VGG-16 network to predict the likability of art images.
- **Sub-objective 2:** Generate three types of heatmaps for each of the paintings: Saliency Heatmaps, GradCam Heatmaps and AIS individualized heatmaps
- **Sub-objective 3:** Determine whether individualizing the AIS heatmap to each participant’s own rating structure yields a closer match to that participant’s aesthetic

processing than a single, non-individualized explainability method (Grad-CAM or saliency) applied uniformly across participants.

4 METHODOLOGY

4.1 Dataset for Training VGG-16: AVA (Aesthetic Visual Analysis)

The first part of the project is to train VGG-16 model on aesthetic rating. For that, the AVA dataset is used. The AVA dataset [4] consists of 255,530 images collected from an online photography community, each annotated with aesthetic quality ratings on a 1–10 scale by multiple users. The mean number of ratings is 210 per image. The weighted average is computed as follows:

$$s = \frac{\sum_{i=1}^{10} i \cdot v_i}{\sum_{i=1}^{10} v_i} \quad (1)$$

where v_i is the number of votes at rating i . A binary label is derived as: high quality if $s \geq 5.0$, low quality otherwise. Dataset statistics after preprocessing:

- Total images: 255,530 (all with ≥ 10 votes)
- Mean aesthetic score: 5.38
- Class distribution: 71.0% high quality, 29.0% low quality
- Train / Val / Test split: 178,855 / 38,326 / 38,327

4.2 Domain Transfer

AVA is formed of photographs, not paintings, which is the focus of this work, and no equivalent paintings-only dataset of comparable size exists. Nevertheless, when training VGG-16 on photographic aesthetic ratings, the mid-level visual features it learns (compositional structure, contrast, salience, spatial distribution of visual complexity etc) are not exclusive to photography but reflect general properties of visual organization. We therefore treat the network’s internal representations as the primary object of analysis. What matters here regarding VGG-16 is not only the classification it makes, but which spatial regions of the image drove the network’s internal representation. Nevertheless, it is important to consider that this does not eliminate the differences and discrepancies that can arise when training and testing across different visual domains.

4.3 VGG-16 Model Architecture

VGG-16 is a deep convolutional architecture with 16 weighted layers organized into five convolutional blocks followed by a dense classifier. In this project two versions of training are built for a comparative study, to assess whether the results obtained are meaningful:

- **Pretrained model:** weights initialized from ImageNet, fine-tuned on AVA with the convolutional layers frozen (classifier-only mode). Trainable parameters: 120,594,945 (89.1%). This way, the layers already “know” how to detect general visual features (edges, textures, shapes, object parts).

- **Random baseline (v2):** weights are initialized randomly.

The original VGG-16 classifier head (`Linear(4096, 1000)`) is replaced in order to allow the fine-tuning on aesthetic preferences. It is replaced by:

$$\text{Classifier} = \text{Dropout}(0.5) \rightarrow \text{Linear}(4096, 256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.5) \rightarrow \text{Linear}(256, 1) \quad (2)$$

Original VGG-16 for image-classification ends with `Linear(4096, 1000)`, a fully-connected layer that takes the 4096-dimensional feature vector produced by the earlier layers and maps it to 1000 output values, one score per ImageNet category, typically followed by a softmax to turn those into class probabilities. This is useless for this project, so it was replaced. The new head takes the 4096-dimensional feature vector produced by VGG-16’s convolutional backbone and compresses it through an intermediate 256-unit layer before collapsing it to a single output value, the predicted aesthetic score. The two dropout layers (50% each) regularize the head against overfitting, given the relatively small size of the head compared to the frozen backbone, while the ReLU activation introduces the non-linearity needed to model the relationship between visual features and aesthetic judgment, rather than a purely linear mapping.

Other training technical details:

- Loss: Binary Cross-Entropy with Logits
- Optimizer: Adam ($\text{lr} = 10^{-4}$, weight decay = 10^{-4})
- Scheduler: `ReduceLROnPlateau` (patience= 3, factor= 0.5)
- Early stopping: patience= 5 epochs
- Max epochs: 20
- Batch size: 32
- Hardware: NVIDIA Titan X (Pascal), 12.8 GB VRAM

4.4 Heatmap Methods

In this project, three types of heatmaps are compared. Those different heatmaps are used basically to evaluate which parts of a specific painting mattered most to the trained network.

Saliency Maps: Computed as the gradient of the model output with respect to the input pixels:

$$S(x) = \left\| \frac{\partial f(x)}{\partial x} \right\|_{\text{max-channel}} \quad (3)$$

These maps reflect which input pixels most affect the model’s output, independently of semantic content. Because they operate at the pixel level, saliency maps tend to be noisy and spatially diffuse, highlighting individual pixels rather than coherent regions: they have no notion of semantic content; the map only “knows” that those pixels numerically influence the output the most.

Grad-CAM: Grad-CAM [5] uses the gradients flowing into the last convolutional layer (`features[28]` in VGG-16) to produce a class activation map:

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right), \quad \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (4)$$

For each one of the paintings, a forward pass through the network produced the model’s output score, after which the gradient of that score with respect to each of the 512 feature maps at `features[28]` is computed via backpropagation. Each feature map’s gradient is spatially averaged to obtain a single importance weight, and the 512 feature maps are then combined as a weighted sum, followed by a ReLU activation to retain only positive contributions, producing a single coarse activation map at the feature layer’s native spatial resolution. This coarse map, considerably lower resolution than the original input due to the convolutional downsampling that precedes it, is then upsampled to match the painting’s original dimensions.

Individualized AIS Heatmaps: The Alignment Importance Score (AIS) [3] scores each of VGG-16’s 512 feature maps by how much they help the network’s representational geometry align with a specific participant’s aesthetic judgments. For AIS then, the fixation maps explained in the following sections are used. Unlike Grad-CAM and saliency, which evaluate a single image in isolation, AIS is relational and individualized meaning that a feature map’s score depends on the full 200-painting comparison set, and differs from one participant to the next.

To compute it, each painting’s `features[28]` activations are pooled into a 512-dimensional vector, and a 200×200 model similarity matrix is built via cosine similarity. A second 200×200 matrix captures how differently the participant rated every pair of paintings. The Spearman correlation between them gives ρ_{baseline} . Each feature map is then removed in turn, the model matrix is rebuilt without it, and the correlation is recomputed as $\rho_{\text{without } k}$; the drop,

$$\text{AIS}_k = \rho_{\text{baseline}} - \rho_{\text{without } k}, \quad (5)$$

is that feature map’s individual contribution to alignment for this participant. Negative scores are zeroed out, the remaining positive scores are normalized to sum to one, and these weights are used to combine a painting’s 512 feature maps into a single spatial map, which is upsampled and colorized to produce the final, individualized AIS heatmap.

4.5 How the heatmaps are laid onto images for comparisons

After each heatmap type is generated, we need to overlay them onto the original paintings to produce the visualizations presented throughout this report in order to allow visual comparisons. Regardless of the method, the resulting heatmap array is first resized to match the painting’s original pixel dimensions via bilinear interpolation, since the three heatmaps are each computed at a resolution that does not coincide with the painting’s native size: saliency is produced at the network’s input resolution (224×224), while Grad-CAM and AIS are both derived from the coarse

14×14 spatial grid at `features[28]` before being up-sampled to 224×224 as part of their own computation. Once resized to the painting’s dimensions, the heatmap’s normalized values are mapped to colors using a colormap, converting low activation into one color and high activation into another, and the resulting colorized heatmap is drawn on top of the original painting at 50 percent transparency, allowing the painting to remain visible beneath the heatmap.

4.6 Eye-Fixation Data — Ground Truth

Four participants were tested on the 200-painting stimulus set. For each painting, participants rated their aesthetic preference on a scale of 1–10 while their eye movements were recorded with a Tobii Pro Spark eye-tracker, producing one fixation density map per image and per participant. These fixation maps serve as the ground-truth signal against which all three heatmap types are evaluated.

4.7 Stimulus Set: 200 WikiArt Paintings

The stimulus set for the eye-tracking experiment consists of 200 paintings of different styles downloaded from WikiArt. There are 25 paintings for each style, to ensure diversity.

Each image was prepared at a resolution of 1920×1080 pixels with black padding, preserving the original aspect ratio of each painting without distortion. The per-trial bounding box of the actual painting (excluding padding) was recorded for every trial, allowing fixation coordinates to be normalized to image space during analysis.

The 200 paintings used as stimuli were collected from WikiArt using a Python script that queries the platform’s public JSON API. For each of the styles, the implemented script retrieves painting metadata (title, artist, year, image URL) by going through the WikiArt style endpoint and collecting the paintings. A simplified stimulus list was also created in a CSV file, containing only the fields required by the eye-tracking software (painting ID, filename, style, title, artist, and year).

4.8 Eye-Tracker Experiment

None of four participants in the experiment held a formal degree in fine arts or art history, which helps ensure that aesthetic responses reflect general rather than expert preferences.

All images were centered within the visual frame of a laptop. For each trial, the exact bounding box of the painting within the display was recorded by the eye-tracker into a CSV, enabling post-hoc normalization of fixation coordinates to the painting’s own coordinate space during analysis.

Eye movements were recorded using a Tobii Pro Spark eye tracker, operating at a sampling rate of 60 Hz. The device was mounted below the laptop display and calibrated individually for each participant using the Tobii Pro Eye Tracker Manager software.

The experimental software was implemented using PsychoPy, interfacing with the eye tracker via the `tobii-research` SDK. An important aspect of this implementation worth highlighting concerns temporal synchronization: the Tobii Pro Spark reports gaze

data with device-native timestamps in microseconds (an absolute hardware clock), while PsychoPy’s internal timer (`core.getTime()`) uses a separate software reference frame. To allow reliable post-hoc alignment of fixation samples to individual trials, the device-time onset and offset of each trial were explicitly captured at runtime (`tobii_onset_device_time`, `tobii_offset_device_time`) and stored alongside the behavioral data. This strategy allows each fixation to be assigned unambiguously to its corresponding trial during analysis.

The experiment consisted of 200 trials presented with scheduled rest breaks after trials 50, 100, and 150. On each trial, a single painting was displayed in full screen on a black background. Participants were instructed to view the painting freely and naturally, as they would when looking at a piece of art in a museum, for example. After a viewing period of 5 seconds, participants provided an aesthetic rating on a scale from 1 to 10 by clicking one of ten numbered buttons displayed at the bottom of the screen. No time limit was imposed for the rating response, although it was observed that most participants tended to answer quite quickly. Eye-tracking data (gaze coordinates and timestamps) were recorded continuously throughout the viewing period of each trial and stored in a raw gaze file alongside the behavioral ratings.

Raw gaze samples were filtered to retain only those falling within the temporal window defined by `tobii_onset_device_time` and `tobii_offset_device_time` for each trial. Fixation coordinates were then normalized from display space to painting space using the per-trial bounding box recorded during stimulus presentation. This normalization step is essential for comparing fixation distributions to heatmaps, which are defined in the painting’s own coordinate system. Fixation density maps were computed by placing a Gaussian kernel at each fixation location and summing the resulting distributions, following standard practice in eye-tracking research.

4.9 Fixation Map Generation

After data collection, raw gaze data were processed into fixation density maps (FDMs). For each participant–image pair, the pipeline proceeded as follows:

Gaze samples were loaded from the participant’s fixation CSV, which contained one row per sample with normalized gaze coordinates in the range $[0, 1]$ for both axes. These were converted to pixel coordinates in display space by multiplying by the screen resolution (1920×1080). The bounding box of the painting within the display which was stored per trial in the trials CSV, was then used to clip the gaze coordinates to the painting’s own spatial extent, discarding any samples falling on the black padding.

Valid fixation hits were accumulated on a zero-initialized canvas of size `paint-height` $\times$ `paint-width`. A Gaussian smoothing kernel was then applied. The resulting map was normalized to the range $[0, 1]$ by dividing by its maximum value.

Figure 1 (in the following page) illustrates this process with a real example: the four fixation density maps obtained for the same Color Field painting, one per participant.

TABLE 1: Stimulus set composition by artistic style

Style	N	Representative Artists
Abstract	25	Kandinsky, Hilma af Klint, Arthur Dove
Impressionism	25	Monet, Renoir, Pissarro
Realism	25	Courbet, Millet, Corot
Cubism	25	Picasso, Braque, Léger
Color Field	25	Rothko, Newman, Frankenthaler
Expressionism	25	Kirchner, Nolde, Kokoschka
Renaissance	25	Botticelli, da Vinci, Raphael
Modern Art	25	Klimt, Mucha, Matisse
Total	200	

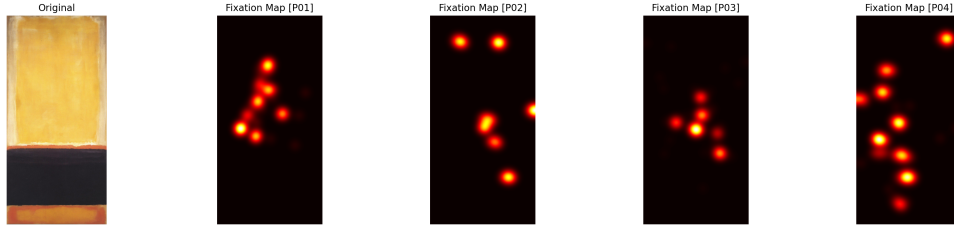


Fig. 1: Fixation density maps for all four participants on the same Color Field painting.

5 RESULTS AND DISCUSSION

5.1 Dataset Exploration

It is important to notice that the AVA dataset presents a moderately imbalanced class distribution (71% high quality, 29% low quality), consistent with the literature [4]. The mean aesthetic score is 5.38 (SD= 0.73), and the mean number of votes per image is 210. This will imply that the models tend to predict “high quality” disproportionately, which limits its discriminative power on aesthetically weak images. However, for the purpose of this project this bias is not a big problem as the models will learn robust aesthetic features from the AVA distribution, which is the important thing.

TABLE 2: AVA Dataset Statistics

Metric	Value	Notes
Total images	255,530	All with ≥ 10 votes
Mean score	5.38	Scale 1–10
Score SD	0.73	
High quality (label= 1)	181,469	71.0%
Low quality (label= 0)	74,061	29.0%
Train set	178,855	70%
Validation set	38,326	15%
Test set	38,327	15%

5.2 VGG-16 Training Results

As said before, two different models were trained.

5.2.1 Pretrained VGG-16 (Transfer Learning)

As explained before, the first model was initialized from VGG-16 weights pretrained on ImageNet and fine-tuned on the AVA aesthetic quality dataset. The pretrained

convolutional backbone was kept frozen during training, while a custom regression head — consisting of two fully connected layers ($\text{Linear}(4096 \rightarrow 256)$ and $\text{Linear}(256 \rightarrow 1)$) with ReLU activations and Dropout regularization — was trained from scratch to predict a continuous aesthetic quality score in the range $[0, 1]$.

The model converged steadily over 11 epochs before early stopping was triggered. As shown in Table 3, training loss decreased consistently across all epochs, while validation loss reached its minimum at epoch 6 (val loss = 0.5130, val accuracy = 74.98%, AUC = 0.755, Spearman ρ = 0.494). The best checkpoint was saved at this epoch and used for all subsequent heatmap generation.

TABLE 3: Pretrained VGG-16 — Training History

Epoch	Train Loss	Val Loss	Train Acc	Val Acc
1	0.5498	0.5251	72.75%	74.11%
2	0.5292	0.5202	74.15%	74.29%
3	0.5197	0.5272	74.80%	74.44%
4	0.5146	0.5220	75.09%	74.47%
5	0.5092	0.5221	75.44%	73.79%
6	0.5036	0.5130	75.79%	74.98%
7	0.4974	0.5188	76.22%	74.88%
8	0.4923	0.5202	76.53%	74.95%
9	0.4860	0.5201	76.87%	74.51%
10	0.4804	0.5162	77.21%	74.45%
11	0.4523	0.5310	79.01%	73.24%

The Spearman ρ between predicted scores and ground-truth mean opinion scores indicates a moderate but consistent ranking ability: the model correctly orders paintings by perceived aesthetic quality significantly above chance. Also, the AUC of 0.755 further confirms that the model has learned a meaningful decision boundary between high- and low-quality images, even in the presence of class imbalance.

These results validate the use of this model as a feature extractor for the downstream AIS heatmap generation.

5.2.2 Random Baseline (v2)

After the initial full-scratch random model collapsed to a constant output, a second version was trained with a lower learning rate and partial fine-tuning (`freeze_until=17`). This resolved the collapse: prediction standard deviation increased steadily across training, reaching ≈ 0.52 by epoch 18 (val loss = 0.5758). Accuracy remained close to the 71% majority baseline throughout, confirming that random v2 has not learned aesthetically meaningful features, making it an appropriate negative control against which the pretrained model’s heatmaps and representational geometry can be compared.

5.2.3 Comparative Summary

TABLE 4: Summary Comparison: Pretrained vs. Random v2 VGG-16

Metric	Pretrained	Random v2
Best Val Loss	0.5130	0.5758
Best Val Accuracy	74.98%	$\approx 71\%$
Best Epoch	6	18
Prediction SD	—	0.52
Majority Baseline	71.00%	

These results demonstrate a clear advantage of transfer learning: the pretrained model achieves 74.98% validation accuracy, outperforming the majority-class baseline by ≈ 4 percentage points, while random v2 tracks the majority baseline without exceeding it. This finding is consistent with Iigaya et al. [2], who showed that VGG-16 pretrained on ImageNet encodes aesthetic-relevant features without explicit training on aesthetic data.

5.3 Heatmap Generation

Grad-CAM and saliency heatmaps were generated for all 200 paintings using the pretrained VGG-16, yielding 400 heatmap arrays stored as 224×224 .npz files together with visualization PNGs.

For the qualitative comparison figures presented throughout Sections 5 and 6, saliency maps required an additional visualization step beyond their original computation. As described before, saliency heatmaps are computed using the pixel-wise gradient of the model’s output with respect to the input image, producing a map at the full input resolution with no inherent spatial smoothing. This gives saliency maps a visibly speckled, high-frequency appearance that contrasts sharply with the smooth, blob-like character of AIS and Grad-CAM heatmaps. To allow direct visual comparison across the three methods, a Gaussian blur was applied to each saliency map purely for figure-generation purposes, smoothing the noisy pixel-level pattern into a comparably continuous spatial distribution without altering its underlying ranking of image regions.

Above we can see two examples of saliency heatmaps generated from an Expressionist painting.

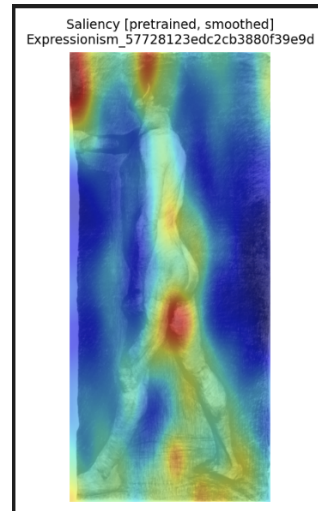


Fig. 2: Smoothed saliency map (pretrained model, $\sigma = 8$) overlaid on the Expressionist figure study.

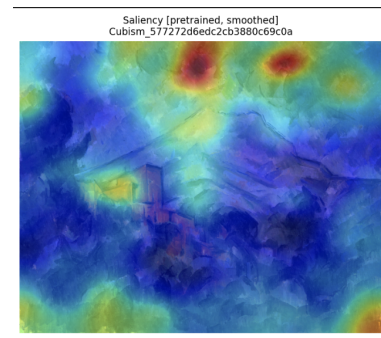


Fig. 3: Smoothed saliency map (pretrained model, $\sigma = 8$) overlaid on the Cubist figure study.

In the first saliency heatmap, we can see that the most important zones are near the head of the figure and in the hips. It is not a bad representation, although if we compare it with the gradcam heatmap showed on fig 5 below, we can see that gradcam is more precise, showing that the more important areas are near the head and specifically on the prominent backside, which is what will most draw the eye at first glance.

Grad-CAM heatmaps were generated as described before, using the gradients flowing into the last convolutional layer of the pretrained VGG-16 model. As we will see, GradCam has been the best performing heatmap, so the results will be explored in the following section (see fig. 5).

Individualized AIS heatmaps were generated for the four participants described above, 200 paintings each. The first step of AIS is to build the baseline alignment, which is also called the Representational Similarity Alignment (RSA). It is computed by building two matrices for each participant and correlating them: On one side, the model Representational Dissimilarity Matrix (RDM) and on the other side the human RDM. The model RDM is computed by passing each of the 200 paintings through VGG-16 up to `features[28]`, spatially mean-pooling the resulting activations into a single 512 dimensional vector per painting, and computing the cosine dissimilarity between every pair of those vectors. On the other side, the Human RDM is constructed from the ratings of participants of the experi-

ment, as the absolute difference between every pair of the ratings. The human RDM basically tells how differently one participant rated every pair of paintings. Both matrices are then reduced to their upper-triangular vectors, and the Spearman correlation ρ between them is computed. It indicates whether the model’s overall sense of which paintings are alike matches that participant’s overall sense of which paintings they ranked alike. This ρ is purely relational meaning that it says nothing about any single feature map’s individual contribution, only about the alignment of the full, unweighted 512-dimensional representation as a whole.

The following Table 5 reports this baseline alignment obtained for each participant.

TABLE 5: AIS baseline alignment by participant (pre-trained model)

Participant	ρ_{baseline}
P01	0.0978
P02	0.0418
P03	0.0850
P04	0.0648

After this, AIS heatmaps can be computed. Using the value obtained above as ρ_{baseline} , each of the 512 feature maps in the last convolutional layer of VGG-16 are, in turn, removed from the representation. Then, the model RDM is recomputed from the remaining 511 feature maps, and a new correlation $\rho_{\text{without } k}$ is obtained against the same human RDM. The drop in correlation caused by removing feature map k , $\text{AIS}_k = \rho_{\text{baseline}} - \rho_{\text{without } k}$, quantifies that feature map’s individual contribution to the participant-specific alignment captured by RSA. A positive AIS_k indicates that removing that feature map weakened the model’s alignment with the participant’s judgments, meaning the feature map was contributing positively; a negative AIS_k indicates the opposite, that removing it actually improved alignment. This masking-and-recomputation procedure is repeated independently for all 512 feature maps, producing one AIS score per feature map, per participant. Feature maps with negative AIS scores are then zeroed out, and the remaining positive scores are normalized to sum to one, so that only feature maps which genuinely contribute to alignment are retained as weights. For a given painting, these normalized weights are used to compute a weighted sum of that painting’s 512 feature maps, producing a single coarse activation map at the feature layer’s native spatial resolution, which is finally upsampled to the painting’s original dimensions and rendered with a colormap to produce the AIS heatmap.

Now we can see an example of an AIS heatmap.

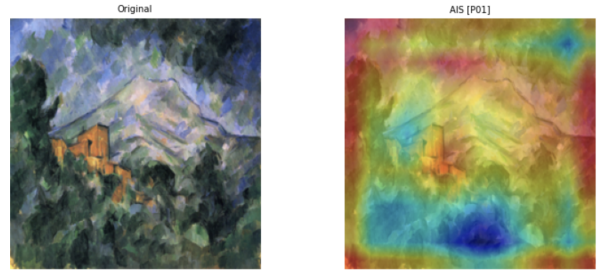


Fig. 4: Individualized AIS heatmap for participant P01 on the Cubist landscape shown in Figures 5 and 6.

Unlike Grad-CAM, which we can see on the following section (fig 5) and concentrates activation on the painting’s central buildings, AIS assigns its highest importance to the image’s edges and corners, with the central compositional focus receiving only mid-range activation. The castle in the center and the area on the top of the mountain are also activated, showing that the AIS heatmap has some important information about relevant areas. The noise on the borders is probably due to the main limitation on this project, the small cohort of participants that took the experiment. With more participants, we could have obtained less noise probably.

5.4 Heatmap–Fixation Comparison

Fixation density maps were generated for all four participants by converting raw gaze samples to painting-relative pixel coordinates using the per-trial bounding box, accumulating them into a 2D histogram, and smoothing with a Gaussian kernel ($\sigma = 4\%$ of painting width), yielding 800 fixation maps normalized to $[0, 1]$.

Each of the three heatmap types was compared against the corresponding fixation map using pixel-wise Spearman correlation, after resizing the heatmap to the fixation map’s resolution, producing $4 \times 200 \times 5 = 4,000$ correlation values.

TABLE 6: Spearman ρ : heatmap vs. fixation map, by type and model (P01–P04, $n = 800$ per row)

Heatmap	Model	Mean ρ	SD
Grad-CAM	Pretrained	0.089	0.2354
Saliency	Pretrained	0.063	0.1329
AIS	Pretrained	0.057	0.2512

Table 6 summarizes the pixel-wise Spearman correlation between each heatmap type and the corresponding human fixation density map across all participants and paintings. Among the three visualization methods, Grad-CAM achieved the highest mean correlation ($\rho = 0.089$), indicating that its highlighted regions exhibit the strongest agreement with human gaze patterns. In contrast, both Saliency ($\rho = 0.063$) and AIS ($\rho = 0.057$) produced substantially lower mean correlations, with AIS showing the weakest correspondence to fixation behavior.

Overall, the correlation values are low for all three methods, suggesting that the model-derived heatmaps capture only a limited portion of the spatial patterns reflected in human visual attention. While Grad-CAM demonstrates a

modest advantage over Saliency and AIS, the differences between the latter two are minimal, indicating that neither approach closely predicts fixation locations. Furthermore, the relatively large standard deviations compared to the mean correlations reveal considerable variability across participants and paintings, implying that the agreement between the generated heatmaps and human fixation patterns is highly dependent on the specific image and observer. Notice that a strong limitation here is the low number of participants, and very probably a higher number of participants in the experiments would have made different results. While Table 6 reports Grad-CAM’s correlation pooled across all four participants, Table 7 breaks this same result down per participant.

TABLE 7: Grad-CAM (pretrained) vs. fixation map, by participant ($n = 200$ per row)

Participant	Mean ρ	SD	n
P01	0.102	0.2458	200
P02	0.0861	0.2334	200
P03	0.0758	0.2295	200
P04	0.0918	0.2300	200

All four participants show a positive Grad-CAM–fixation correlation individually, consistent in sign with the pooled result, indicating that the pooled mean in Table 6 is not masking participants who pull in opposite directions. All four participants individually show a positive Grad-CAM–fixation correlation. None show a negative or null relationship that the pooled average could be masking. This means the Grad-CAM advantage isn’t an artifact of one participant with an unusually strong correlation dragging the group mean up — it’s a directionally consistent pattern across the whole (small) cohort.

In the following images we can see what GradCam predicts and the actual fixation map for those paintings for participant P01.

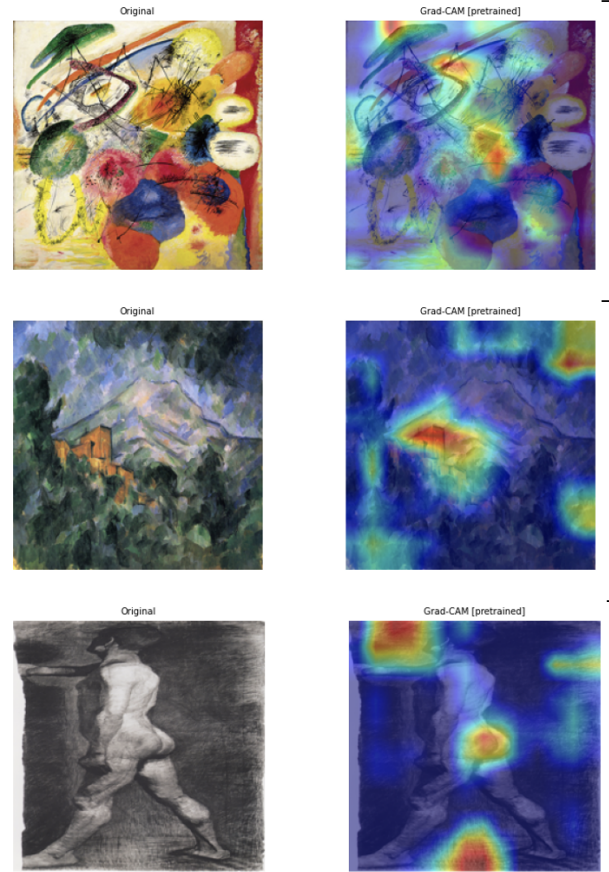


Fig. 5: Grad-CAM heatmaps (pretrained model) for three paintings: an Abstract composition, a Cubist landscape, and an Expressionist figure study. In each case, Grad-CAM concentrates activation on regions of high local contrast or structural complexity within the painting.



Fig. 6: Participant P01’s fixation density map overlaid on the same three paintings shown in Figure 5. Abstract-Cubism-Expressionism

Across the three paintings shown here, the Cubist landscape shows the clearest qualitative agreement between the GradCam heatmap and the fixations map: both Grad-CAM and P01’s fixations converge tightly on the same central region, the warm-toned buildings nestled against the mountain’s base, probably the area of highest local contrast in the scene. The Abstract painting shows partial overlap: Grad-CAM spreads its activation broadly across nearly the entire canvas, picking out several distinct high-contrast shapes and lines, while P01’s fixations cluster more narrowly around two or three specific regions, one of which coincides with a Grad-CAM hotspot near the upper-left while the others fall on different sub-regions Grad-CAM treats as comparatively secondary. The Expressionist case shows the weak-

est spatial agreement of the three: Grad-CAM’s activation concentrates heavily on the dark pillar-like form on the left and a smaller hotspot near the figure’s torso, whereas P01’s fixations are distributed along the length of the standing figure itself, hips, torso, and lower legs, largely bypassing the pillar that Grad-CAM treats as most salient. These three examples illustrate that Grad-CAM and human fixation align better when a painting has one dominant, unambiguous focal region (as in the Cubist landscape), but the two diverge more once a painting offers multiple plausible regions of interest, with Grad-CAM’s edge/contrast-driven activation pulling toward different content than where a human viewer’s attention is actually drawn.

5.5 Grad-CAM, not AIS, Best Predicts Raw Fixation Location

As shown in Table 6, Grad-CAM is the method that shows the strongest alignment with real fixations.

This result is noteworthy because AIS and Grad-CAM are both computed from the *same* final convolutional layer (`features[28]`, 512 feature maps encoding mid-to-high-level semantic structure), so the gap between them cannot be attributed to the depth of the representation used. The key difference lies in *how* that information is weighted: Grad-CAM weights feature maps according to their contribution to the model’s output for a *single image* considered in isolation, whereas AIS weights them according to their contribution to the relational alignment with the similarity structure of the entire 200-painting set. This suggests that AIS’s weaker performance in this comparison is not due to using less informative features, but specifically to its relational nature: an importance score defined in comparative terms need not coincide with the intrinsic salience of a single image, which is what seems to drive gaze location during a free-viewing task. Nevertheless, this conclusion is limited by the fact that only four participants took the experiment.

5.5.1 Comparison by Artistic Style

TABLE 8: AIS (pretrained) vs. fixations by style (P01–P04, $n = 100$ per style)

Style	Mean ρ
Color Field	0.1875
Renaissance	0.0930
Cubism	0.0620
Expressionism	0.0275
Impressionism	0.0260
Abstract	0.0245
Modern Art	0.0219
Realism	0.0132

Color Field paintings stand out as the style with by far the strongest AIS–fixation alignment. All other styles show lower correlations, with Realism showing the weakest alignment.

Here an interesting thing to consider was observed. Color Field paintings are the ones with the strongest AIS–fixation

alignment, but if we consider the RSA results (baseline computed before masking the feature maps), Color Field instead shows the *weakest* alignment of any style, the exact opposite ranking from the pixel-level comparison. RSA and pixel-level AIS–fixation alignment test fundamentally different properties of model–human agreement. AIS–fixation alignment evaluates a spatial question within a single image: whether the heatmap’s predicted regions of importance coincide with where a participant’s gaze actually landed on that specific painting. By contrast, RSA, evaluates a relational question across many images, whether the model’s overall pattern of similarity between paintings matches a participant’s overall pattern of rating similarity, independent of any spatial location. This observation can be attributed to the visual homogeneity of Color Field paintings which implies that the model struggles to differentiate between the 25 color field paintings. On the human part, differently, the human participant can differentiate between paintings. For example they can like a blue one and not a yellow one, although for the model, they are both very similar. This election can be done for subtle reason that the model can never tell, for example the mood, the color warmth etc.

5.6 Different exploratory options done to try to improve AIS that didn’t improve the results

Three further exploratory checks were conducted in order to improve the AIS results. None of the three improved the result which makes us conclude that the main problem regards the small amount of participants that took the behavioral experiment. No more experiments were performed due to time constraints.

Pooled ratings. Instead of doing for each one of the participants, AIS was recomputed from a single human similarity matrix built by averaging all four participants’ ratings together before the AIS procedure, rather than computing AIS separately per participant. This yielded $\rho_{\text{baseline}} = 0.0751$, a value comparable to but not exceeding the strongest individual participant (P01: $\rho = 0.0978$).

Earlier convolutional layer. AIS was recomputed from `features[14]` (256 channels, lower-level visual features) rather than the main pipeline’s `features[28]`, for all the four participants. This produced a *more negative* fixation correlation for every participant (overall mean $\rho = -0.0835$ vs. -0.0312 ; range -0.1169 to -0.0668 vs. -0.0349 to 0.0545), arguing against layer depth as the primary explanation for AIS’s weak performance.

Smaller comparison set. AIS was recomputed using only the first 50 paintings in the stimulus metadata (25 Abstract, 25 Impressionism) instead of the full set of 200, using the same `features[28]` layer. This also produced a worst result.

6 CONCLUSIONS

Having trained a VGG-16 aesthetic quality model on AVA, generated all three heatmap types for 200 WikiArt paintings, collected eye-tracking and rating data from four participants, and compared the resulting heatmaps against both

raw fixation density maps the following conclusions can be drawn.

6.1 VGG-16 fine-tuning

The VGG-16 model fine-tuned on AVA achieved satisfactory performance (74.98% validation accuracy, $AUC = 0.755$, Spearman $\rho = 0.494$), demonstrating that transfer learning from ImageNet to aesthetic quality assessment is a good option even when training on photographs and testing on paintings. The random baseline (v2) performed worst, confirming that the pretrained model’s representations are genuinely aesthetic-relevant rather than artifacts of the training procedure.

The model’s performance is particularly noteworthy given the domain shift between training (photographs) and evaluation (paintings). The fact that a network trained exclusively on photographic aesthetic ratings can achieve 74.98% accuracy on the validation set, and shows a Spearman correlation of 0.494 with ground-truth mean opinion scores indicates that the visual features underlying aesthetic judgment are at least partially domain-general. Compositional principles such as for example balance, contrast, and the distribution of visual weight, which the network learned from photographs, appear to transfer to paintings despite the substantial differences in texture, brushwork, and subject matter between the two domains.

The comparison with the random baseline (v2) reinforces this interpretation. The random-initialized model never exceeded the 71% majority baseline accuracy, and its predictions remained essentially constant throughout training, confirming that the pretrained model’s performance is genuinely driven by the visual features it inherited from ImageNet and refined on AVA, rather than by the architecture alone.

6.2 Grad-CAM Outperforms AIS and Saliency in Fixation Prediction

As explored in the preceding sections, individualized AIS heatmaps did not outperform Grad-CAM at predicting human fixation locations during aesthetic evaluation. Table 6 summarizes the pixel-wise Spearman correlation between each heatmap type and the corresponding human fixation density map, pooled across all participants and paintings. Grad-CAM achieved the highest mean correlation ($\rho = 0.089$, $SD = 0.235$), followed by saliency ($\rho = 0.063$, $SD = 0.133$) and AIS ($\rho = 0.057$, $SD = 0.251$).

The magnitude of the difference between Grad-CAM and AIS, while modest in absolute terms is noteworthy given that both methods are computed from the *same* final convolutional layer of VGG-16 (*features* [28], consisting of 512 feature maps encoding mid-to-high-level semantic structure). The key distinction between the two lies not in the depth of the representation used, but in *how* the information contained in those 512 feature maps is weighted and aggregated. Grad-CAM assigns importance weights to each feature map based on its contribution to the model’s output for a *single image considered in isolation*, computing the gradient of the output score with respect to each feature map and spatially averaging the result to produce a single

scalar per feature map. This procedure yields a weight vector that reflects the intrinsic salience of each feature map for the specific painting under consideration, independent of any external reference set. In essence, Grad-CAM asks: *Which regions of this particular image most strongly drive the model’s aesthetic judgment?* AIS, by contrast, assigns importance weights based on each feature map’s contribution to the *relational alignment* between the model’s representational geometry and the participant’s rating structure across the *entire 200-painting stimulus set*. The AIS computation is inherently comparative: it measures the drop in RSA correlation ($\rho_{\text{baseline}} - \rho_{\text{without } k}$) that occurs when each feature map is removed from the 512-dimensional global representation, and this drop is defined relative to the participant’s full set of 200 ratings, not relative to any single image. A feature map that receives a high AIS score is one that plays a crucial role in distinguishing between paintings in the comparison set—that is, a feature map whose removal substantially weakens the alignment between the model’s pairwise similarity structure and the participant’s pairwise rating differences across the entire collection.

This distinction offers a possible explanation for why AIS underperforms Grad-CAM in predicting fixation location. During a free-viewing aesthetic evaluation task, a participant’s gaze is drawn to regions of an image that are intrinsically salient or visually compelling—areas of high contrast, compositional focus, semantic interest, or formal complexity. These are precisely the kinds of regions that Grad-CAM, with its image-specific, output-gradient-based weighting, is designed to highlight. AIS, by contrast, highlights regions that are important for *comparative* purposes—regions that help distinguish this painting from others in the set in terms of the model’s representational geometry. A region that is crucial for distinguishing between paintings in the set need not be the same region that attracts a viewer’s first glance when viewing a single painting in isolation. The two notions of importance—intrinsic salience and relational distinctiveness—are conceptually distinct and, as the results suggest, partially empirically dissociable.

Finally, it is worth noting that the correlation magnitudes for all three methods are modest in absolute terms (none exceeding $\rho = 0.10$). This indicates that even the best-performing method—Grad-CAM—captures only a small fraction of the variance in human fixation patterns. This is not entirely surprising because fixation during aesthetic viewing is influenced by multiple interacting factors, including low-level visual salience, semantic content, personal interest, task demands, and individual differences in viewing strategy, none of which are fully captured by a single model trained on average aesthetic ratings. Future work could explore hybrid approaches that combine multiple sources of information—for example, by integrating Grad-CAM with semantic segmentation or object detection—or that train models directly on fixation data rather than on aesthetic ratings. Such approaches may yield stronger predictive performance and a more complete understanding of the relationship between computational aesthetics and human visual attention.

6.3 Limitations: A Cohort of Four Participants

All of the conclusions above must be read against the central limitation of this study which is that the experimental cohort consists of only four participants. With $n = 4$, every per-participant and per-style estimate reported in this work is a comparatively imprecise estimate of the corresponding population-level quantity. Differences between heatmap types or styles that are individually statistically significant within this sample do not guarantee totally that the same ranking of methods would emerge in a larger and more diverse cohort, and small differences between closely-spaced mean ρ values should not be over-interpreted. The conclusions drawn here should therefore be understood as evidence of a coherent and interpretable pattern within this specific cohort, rather than as a definitive, generalizable characterization of how AIS, Grad-CAM, and saliency maps compare as predictors of human aesthetic attention. Nevertheless, the implemented algorithms and pipeline will be perfectly applicable if the experiments have to be made for a bigger n number of participants.

6.4 Philosophy of mind considerations

The methodology pursued in this project rests on a number of philosophical commitments that are worth making explicit, since they help us to interpret the results. Throughout this work, aesthetic experience has been studied entirely through third-person, measurable proxies: gaze coordinates, numerical ratings, and the geometric structure of internal model activations. This approach is standard in computational and cognitive neuroscience, and it is the only tractable way to study aesthetic judgment empirically. It is nonetheless worth noting that it leaves untouched a much deeper question that philosophers of mind name the explanatory gap: even a complete account of the causal and computational processes that produce a behavior (here, a fixation pattern or a numerical rating), does not by itself explain why, or whether, that process is accompanied by any subjective, first-person experience of finding a painting beautiful. The correlations reported in this work, however informative about the structure of behavior and representation, are agnostic on this question; a model could in principle reproduce every measured correlation in this study while having no aesthetic experience whatsoever, just as, conversely, two human participants could share an identical felt sense of beauty while producing very different gaze patterns and ratings due to unrelated factors such as attention, fatigue, or response style. The gap between AIS and Grad-CAM observed in Section 5.5 is therefore best understood as a gap between two behavioral and representational signatures of aesthetic processing, not as evidence bearing directly on the subjective experience that may or may not accompany either signature.

This project, like most work at the intersection of deep learning and human cognition, operates productively within a functionalist framework: it compares systems by what their internal states do, how feature maps contribute to alignment or output, rather than by how those states are physically realized. This is a reasonable and common working assumption, but it is an assumption nonetheless, and the

conclusions of this work should be read as claims about behavioral and representational alignment between model and human, rather than as claims about the presence, absence, or nature of subjective aesthetic experience itself.

Explaining the nature of the aesthetic experience as a whole is not the objective of this project, and actually, facing back to German Romanticism poets, we can say that that is probably an impossible task. Art and aesthetic experience has a human, private and individual component that probably disappears when we tried to study it in detail. In some sense, what makes art valuable is some kind of mysterious and unexplainable experience, that disappears when we try to analyze it in parts, because it just works as that, as an experience. To illustrate that and end the report, I find convenient to pay attention to Goethe's poem *The Joy*. The poem describes a dragonfly whose colors shift continuously and beautifully in flight, "now dark... now bright... now blue, now red, and now of greenish hue", until the human finally captures it and examines it closely, only to find it "of a sad and dingy blue." Goethe addresses the poem to the "Joy-Dissector," suggesting that the very act of analytical capture can drain the phenomenon under study of the quality that made it worth studying in the first place. Aesthetic experience, like the beauty of the dragonfly in flight, may lose something essential when decomposed into pixel-wise correlations, feature-map activations, and representational dissimilarity matrices. The heatmap and RSA analyses in this project should therefore be read as partial, necessarily reductive proxies for a phenomenon that, much like the dragonfly's color, may not survive being held still for inspection.

Joy

A Dragon-fly with beauteous wing
Is hov'ring o'er a silv'ry spring;
I watch its motions with delight,
Now dark its colours seem, now bright;
Chameleon-like appear, now blue,
Now red, and now of greenish hue.
Would it would come still nearer me,
That I its tints might better see
It hovers, flutters, resting ne'er!
But hush! it settles on the mead.
I have it safe now, I declare!
And when its form I closely view,
'Tis of a sad and dingy blue —
Such, Joy-Dissector, is thy case indeed.

— Johann Wolfgang von Goethe

REFERENCES

- [1] J. W. von Goethe, "Joy." [Poem]. Available: https://www.yourdaily poem.com/listpoem.jsp?poem_id=798. Accessed : Jun.23, 2026. *Publicdomain*.
- [2] Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2021). Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nature Human Behaviour*, 5(6), 743–755. <https://doi.org/10.1038/s41562-021-01124-6>

- [3] Truong, N., Pesenti, D., & Hasson, U. (2025). Explaining human comparisons using Alignment-Importance Heatmaps. *Computational Brain & Behavior*, 8, 421–441. <https://doi.org/10.1007/s42113-025-00235-x>
- [4] Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. *Proceedings of CVPR 2012*, 2408–2415.
- [5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of ICCV 2017*, 618–626.
- [6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [7] D. Kahneman, *Pensar rápido, pensar despacio*, trad. J. Chamorro Mielke. Barcelona, España: DeBolsillo, 2012.
- [8] Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4.