IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 September 2025, accepted 15 September 2025, date of publication 29 September 2025,
date of current version 9 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3615404

RESEARCH ARTICLE

Comparative Perspective of Visual Attention: From Human Focus to Visual Transformers–An In-Depth Review

LUIS GUILLERMO OLIVEROS PIÑERO[1], (Student Member, IEEE), MIGUEL CARRASCO[2], JOSÉ ARANDA[1], (Member, IEEE), AND CÉSAR GONZÁLEZ-MARTÍN[3]

[1]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibañez, Santiago de Chile 7941169, Chile
[2]Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Santiago de Chile 8370109, Chile
[3]Faculty of Education Sciences and Psychology, University of Cordoba, 14071 Cordoba, Spain

Corresponding author: Luis Guillermo Oliveros Piñero (loliveros@alumnos.uai.cl)

This work was supported by the Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez.

ABSTRACT Although neuroscience has made considerable progress in recent decades by proposing robust models that explain the mechanisms of attention and perception in humans, emulating this capability using computational techniques remains complex. It was not until the development of models such as Visual Transformers (ViT) that it became possible to partially replicate this uniquely human trait. The main objective of this study was to explore the extent to which attention models, such as ViT, can reproduce the manner in which people distribute their visual attention when exposed to various stimuli, particularly in the context of handcrafted objects. Human fixations (i.e., attention) were recorded using an eye tracker, while the ViT model processed the same images to generate attention maps to evaluate the degree of similarity between the two patterns. For this purpose, heatmaps were constructed, and quantitative metrics were applied to assess their similarity. The results revealed areas of convergence and significant differences, highlighting the current limitations of computational models in capturing the more subtle aspects of human perception. This comparison not only helps us better understand the capabilities of ViT but also provides a foundation for reflecting on future improvements in automated attention models and their potential applications in contexts where visual interpretation is crucial.

INDEX TERMS Attention, eye-tracker, experiments, comparison, human attention, multihead attention, transformer, visual transformers, vision computer, vision transformers.

I. INTRODUCTION

Visual attention is an essential component of human perception and plays a crucial role in our ability to process and understand our surrounding environment. This ability enables us to focus on the most salient features of our complex surroundings rather than compressing everything into a static representation [51]. Attention is not limited to visible eye movements, that is, shifting the gaze toward an object or location in the visual field (a process known as overt attention [16]), but also manifests as a mental mechanism that shifts the focus of attention without moving the eyes (covert attention). These mental shifts guide our focus without visible changes in gaze direction [39], influencing how we interpret and process information.

Since the emergence of the first theories of attention in the 1980s ([2], [25], [26], [38], [40], [49]), our understanding of how attention manifests in various contexts has significantly advanced. Once a purely theoretical idea, attention is now understood as a physiological and cognitive reality that has drawn the attention of the scientific community for decades. This has inspired the development of computer vision models that emulate the characteristics of the human visual system, providing a valuable perspective for both enhancing artificial intelligence and deepening our understanding of human cognition [20].

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

Attention mechanisms, proposed for machine translation, dynamically assign importance to various parts of a sequence based on context [3]. This approach adjusts the weights according to the features of the input information. In the domain of digital images, attention mechanisms focus on the most relevant regions and discard less important areas [20]. Inspired by the human cognitive system, these mechanisms emulate cognitive awareness by amplifying critical information and emphasizing essential data [22].

In deep learning, attention mechanisms were introduced in 2016 to address the inherent challenge of retaining and recalling relevant information in encoder-decoder architectures while processing data sequences [3]. This challenge, known as the forgetting problem, manifests as a progressive loss of critical information as the network advances in sequence analysis. A year later, in 2017, the transformer architecture marked a turning point in the implementation of attention mechanisms [48]. By eliminating the need for recurrences and convolutions, transformers efficiently capture the relative importance of each element in an input sequence. This breakthrough revolutionized sequential data processing in deep learning models, significantly improving the modeling capacity and performance for complex sequence interpretation tasks.

A distinctive aspect of the transformer architecture is the implementation of an advanced form of attention, known as multi-head attention [48]. This mechanism combines multiple attention layers operating in parallel, each of which applies different linear transformations to the same input. Instead of computing attention only once, multi-head attention splits the input into fixed-size segments and independently computes the scaled dot product attention for each segment [11]. This technique allows the model to simultaneously focus on different representation subspaces and positions within the data, thereby enhancing information processing and interaction. Compared with simpler attention approaches, the multi-head mechanism improves the model's ability to understand and process complex data effectively and efficiently [48].

In recent years, several studies have significantly contributed to bridging the gap between human attention patterns and those generated by Visual Transformers (ViT). Cadoni et al. investigated the correlation between both types of attention by developing a dataset based on human fixations on facial images, demonstrating that ViT can replicate human patterns by highlighting discriminative regions [9]. Complementarily, Park and Kim analyzed self-attention mechanisms in ViT and concluded that they improve both accuracy and generalization by flattening loss landscapes and act as low-pass filters, unlike CNNs, which function as high-pass filters [36].

This study proposes a comparative analysis between the attention mechanism of the visual transformer (ViT) architecture—introduced by Dosovitskiy et al. [14] and the attention patterns observed in a group of participants. In this experiment, participants observed a set of images containing handcrafted objects (described in detail in the Methodology section), and heat maps were generated to reflect their attention patterns. For comparison, the images used in the experiment were processed using a pretrained ViT model to produce equivalent heatmaps. This study aims to quantitatively and qualitatively evaluate the similarities and differences between human and ViT-generated attention patterns.

This work contributes to the literature in two main ways. The first is a quantitative comparative analysis of human attention patterns and those of the ViT model, focusing on human fixation on specific images. Various metrics will be used, such as Wasserstein distance and its variants, maximum mean discrepancy, Hellinger distance, total variation distance, and Kullback-Leibler divergence, among others. The second contribution is the creation of a dataset documenting eye movements and fixations collected through an experiment in which participants observed images of handcrafted objects.

II. STATE OF ART

A. ATTENTION TO THE VISION OF THE HUMAN

Visual attention in humans is a key mechanism of the nervous system that enables us to perceive, process, and highlight specific locations, objects, or features within our visual field. Although we are not always aware of its operation, visual attention plays a fundamental role in visual perception and in various aspects of our interaction with the visual environment, such as learning and memory.

Research on visual attention has undergone remarkable evolution since Treisman and Gelade laid the theoretical foundations by proposing the feature integration theory of attention [46]. This foundational theory posits the necessity of a sequential process to integrate multiple visual features, introducing the critical distinction between focal attention and top-down processing.

The distinction between covert and overt attention, initially proposed by Posner et al., has been a central theme in the research [40]. Ward expanded on this line of work by demonstrating the brain's capacity to enhance sensory processing without eye movements [49]. Bisley emphasized the importance of visual attention for perception and for all the ways in which we use perception, including learning, memory, and interaction with the visual environment [6]. Typically, we think of focusing attention on objects or features in terms of making a quick eye movement (a saccade) to bring the object of attention to the center of the gaze. However, our visual system can also process information from selected peripheral regions of the retina. When done consciously, this is often described as "looking out of the corner of the eye" (see the example in 1).

Belyusar et al. later refined these definitions, characterizing overt attention as the explicit engagement of the motor system and covert attention as stealthy silent deployment of the attentional spotlight [4]. The relationship between eye

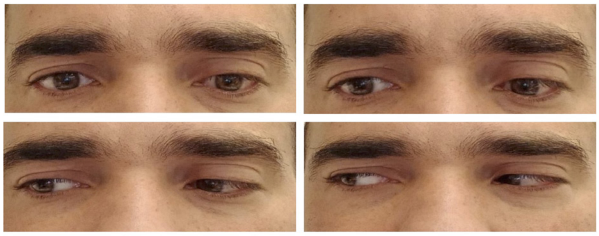

FIGURE 1. Process of a glance through the corner of the eye (own source).

movements and attention has been extensively investigated. Groner and Groner laid the groundwork by exploring this interconnection [18], followed by Crawford and Müller, who investigated saccadic movements [13]. Engelke et al. further deepened this distinction, showing that both mechanisms operate with similar efficiency in terms of response times, particularly in everyday scenarios, such as driving [16].

In the context of specific visual processing, Inhoff et al. [26] and Henderson et al. [25] examined the role of attention during reading and the processing of extrafoveal information in detail. Kulke et al. expanded this understanding by demonstrating how attention enhances visual responses to stimuli, while inhibitory mechanisms suppress responses to distractors. They observed that visual response amplitudes decreased as the number of distracting stimuli increased, while attentional responses increased [29].

Advances in computational modeling have been made. Itti and Koch developed a pioneering pre-attentive visual attention model [27], whereas Satoh and Miyake proposed a model based on scale-space theory [43]. Frielink-Loing et al. found that attention distribution is independent of eye movements, with covert attention consistently showing an anisotropy in object tracking their experiments revealed that covert attention always considers motion information when tracking objects, whereas overt attention is more flexible, and its anticipatory nature depends on the task [17].

The integration of visual attention with other cognitive processes is another crucial aspect. Cowan established links between short-term memory and attention [12], whereas Driver and Spence broadened our understanding by examining spatial attention across different sensory modalities [15]. Blair and Ristic added another dimension by arguing that covert attention reflects mental readiness, whereas overt attention incorporates oculomotor resources, they found a high degree of similarity between automated, voluntary, and combined covert and overt attention when tasks and stimuli were matched across both response conditions [7].

More recent research by Parr and Friston proposed a theoretical framework integrating covert sensory selection with the active manipulation of sensory structures, suggesting two fundamental interpretations of attention [37]; the first involves covert selection among multiple sensory channels, attributing greater importance to sensory streams that convey the most reliable information about the states of the world and the second interpretation requires a more active approach, involving the overt manipulation of sensory structures to deliberately select the data perceived, thus demonstrating the complementarity and flexibility of both types of attention in our interaction with the environment.

B. ATTENTION MECHANISMS IN COMPUTER VISION

In 2016, Xu et al. presented a visual attention-based approach for automatic image caption generation [51]. They proposed a model that uses a deep output layer to calculate the probability of the output word, considering the Long Short Terms Memory (LSTM) state, context vector, and previous work. They explored two attention mechanisms: stochastic (“hard”) and deterministic (“soft”), achieving better results than previous methods in metrics such as BLEU and METEOR. In addition, they demonstrated that the model intuitively aligns with visual attention.

In 2018, Milanova conducted a review of visual attention mechanisms [34], highlighting their relevance in human visual perception, where they help locate regions of interest and process subsets of visual input. They pointed out that visual attention is key in computer vision, neuroscience, and deep learning, with applications such as object segmentation and recognition, image captioning, and visual question answering (VQA). They classified attention models into bottom-up models based on scene features and top-down models guided by the observer’s prior knowledge.

In 2021, Chaudhari et al. provided a detailed review of attention models in neural networks, proposed a taxonomy of attention techniques, and explored architectures and applications [11]. They addressed how attention improves the interpretability of neural networks and discussed the co-attention and self-attention models. They analyzed various levels of abstraction and positions in attention models, differentiating between soft, hard, and local attentions.

In 2022, Guo et al. provided a comprehensive overview of attention mechanisms in computer vision, categorizing them into channel, spatial, temporal, and branch attention. They highlighted their applications in tasks such as image classification, object detection and semantic segmentation [20]. They analyzed the advantages and limitations of these mechanisms and suggested future research directions, concluding that attention-based models could eventually replace convolutional networks as a more powerful and general architecture for image classification.

In 2022, Hassanin et al. categorized fifty attention techniques in deep learning [22]. They discussed the strengths, limitations, and applications of these networks, including spatial, spectral, pixel-wise contextual, pyramidal, and regional attention, as well as self-attention and non-local networks. They explored multimodal attention and proposed techniques to improve it, such as reinforcement and transfer learning, addressing challenges and open questions.

In 2022, Guo et al. directly addressed the limitations of self-attention mechanisms applied to computer vision, such as the difficulty in preserving the spatial structure of images, high computational cost, and low efficiency in capturing complex visual patterns [19]. To address these issues, they proposed the Large Kernel Attention (LKA) mechanism, which combines the efficiency of convolution with the ability of self-attention to model global dependencies. Building on this foundation, they developed the Visual Attention Network (VAN) architecture, which achieved superior results compared to transformers and traditional convolutional networks in tasks such as image classification, object detection, and semantic segmentation. Through ablation studies and experiments on the ImageNet-1K benchmark, the authors demonstrated that the LKA provides an efficient and effective alternative for visual attention modeling.

C. VISUAL ATTENTION FROM VISUAL TRANSFORMERS

In 2017, Vaswani et al. introduced the "Transformer" architecture, based on multi-head attention, eliminating the need for convolutional and recurrent networks in text processing [48]. This architecture achieved state-of-the-art results in machine translation, surpassing previous models in terms of quality, parallelization, and training efficiency. In 2021, Dosovitskiy et al. introduced the Vision Transformers (ViT) model, which processes images as sequences of patches [14]. ViT has demonstrated outstanding performance in image classification when trained on large datasets, surpassing traditional convolutional networks. In the same year, Han et al. provided an overview of the state of the art of Transformers in computer vision, highlighting their application in tasks such as object detection, segmentation, image generation, video processing, and pose estimation [21]]; they also proposed a Transformers in Transformers (TNT) model, which introduces patch subdivisions to improve feature representation through local and global positional encodings. TNT have higher accuracy on datasets such as ImageNet with lower computational costs than standard transformers.

In 2021, Tuli et al. analyzed the error consistency between humans, CNNs, and ViTs, and found that ViTs exhibited greater agreement with human decisions owing to their bias toward shape over texture [47]. This positions them as models that are more like human vision than CNNs. Raghu et al. explored the differences in internal representations between ViTs and CNNs and demonstrated that ViTs generate more uniform and global representations [41]. They also observed that residual connections in ViTs are key to propagating features between layers.

In 2022, Xu et al. published a detailed analysis of visual Transformers, addressing their structural design and application in both high- and low-level vision, content generation, and multimodal learning [52]. Models such as DETR, DALL-E, and TransGAN stood out in tasks like image restoration, generation, and text-image fusion. That same year, Yang et al. reviewed the use of Transformers in visual learning, demonstrating their superiority over convolutional networks in tasks such as segmentation, detection, and image synthesis [53]. They highlighted variants such as X-Transformers and Segmented, which leverage global context in segmentation. James Wensel et al. investigated Transformers for human activity recognition, proposing Recurrent Transformers (ReT) and Vision Transformers (ViT), which improve speed and scalability compared to traditional CNNs and RNNs, highlighting the need for lightweight models for resource-limited devices [50].

In another 2022 study, Cadoni et al. investigated the correlation between human attention and ViT attention [9]. They developed a dataset of human fixations on facial images and compared them with attention maps generated by ViTs, showing that these models can mimic human attention by highlighting discriminative regions of the images. Park and Kim analyzed the self-attention mechanisms in ViTs, highlighting that they improve both accuracy and generalization by flattening loss landscapes [36]. They also demonstrated that ViTs act as low-pass filters, in contrast to CNNs, which operate as high-pass filters, suggesting a complementarity between the two architectures.

In 2023, Mehrani and Tsotsos concluded that attention mechanisms in ViTs do not replicate human attention but instead perform perceptual grouping based on similarities [33]. They also pointed out limitations in tasks such as detecting unique elements, where ViTs do not outperform CNNs. Moutik et al. conducted a comparative analysis between CNNs and ViTs in action recognition tasks, highlighting that ViTs perform better on large datasets thanks to their ability to model long-range relationships, while CNNs are preferable in scenarios with limited data [35].

III. METHODOLOGY

The proposed method consists of three stages: data understanding, heatmap generation, and model evaluation and validation (see Figure 2). The first stage involves collecting, understanding, and preparing the data obtained from the experiments. The second stage consists of creating heatmaps based on the participants' visual attention patterns and, in parallel, applying the Visual Transformers architecture to the data to generate heatmaps, allowing for analysis of the attention mechanism. The third stage involves evaluating and comparing the results using various metrics. Finally, we present a discussion of the results obtained.

A. DATA UNDERSTANDING AND PREPARATION

The collection of data on participants' fixations was carried out using the following equipment: a 52-inch TV with a projection speed of 60 MHz, a pupil-Labs Core model eye-tracker lens connected to a desktop computer with a 64x AMD Ryzen 5 5600X 6-Core 3.70 GHz processor, 48 GB of installed RAM, and an NVIDIA GeForce RTX 3060 graphics card (GPU), running Windows 11 Pro. The following software was installed: pupil capture version 3.5-1 for Windows 11 and its add-ons pupil player and pupil service. This set of applications was used for calibrating the eye tracker, collecting fixation data from each participant, and analyzing the data to separate each image and the corresponding participant fixations. OBS Studio version

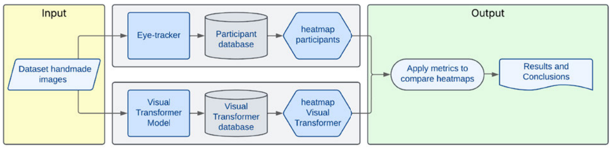

FIGURE 2. Methodological diagram illustrating the flow of an experiment whose objective is to compare human visual attention with the attention generated by a computational model of the Visual Transformer type [14]. The process begins with a set of images of basketry and jars that are used as visual stimuli. These images are presented to different human participants, whose attention is recorded through the use of an eye-tracker, generating a database that is then used to construct heat maps representative of human visual attention. In parallel, the same images are processed by a Visual Transformer model, which generates a second database and produces heat maps reflecting the model's attention. Both heat maps, that of the participants and that of the model, are then compared using several specific metrics that allow to analyze the degree of similarity between them. Finally, this comparison allows obtaining results and conclusions about human visual attention behavior versus the computational model.

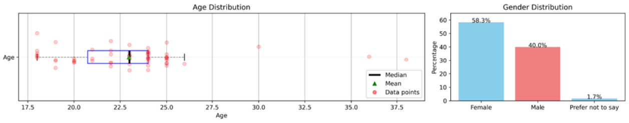

FIGURE 3. Distribution of participants in the experiment according to gender and age group.

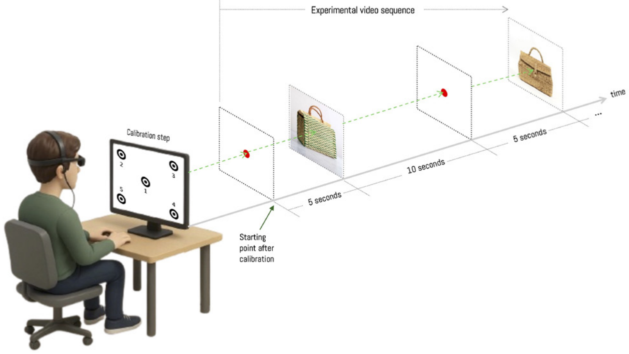

FIGURE 4. Experimental procedure for object visualization. Before starting the experimental phase, a calibration procedure is performed by recording a sequence of points on the screen for each user. Once this process is completed, the experimental phase begins through the projection of an image with a white background and red dot which is displayed for 5 seconds. Then one of the 20 objects is displayed for 10 seconds. This procedure repeats until all objects have been displayed.

29.1 was installed for marker visualization on the screen, and Python version 3.12.5 was used to create scripts to manage transitions and image presentations during the experiment. The latter also enabled the construction of heatmaps and the Visual Transformers model [14].

The experiment with participants followed this protocol: First, it was verified that participants were individuals (regardless of gender) over 18 years old, regardless of profession or education level, and without any vision problems (especially near vision, less than 1.3 meters). Of the total participants in the experiment, 93.3% belonged to the age group of 18 to 25 years, while only 6.7% were over 25 years old. In the 18 to 25 age group, females predominated at 60.7%, followed by 37.5% male participants and 1.8% who preferred not to answer. In the group over 25 years old, 75.0% identified as male and 25.0% as female, with no individuals opting not to disclose their gender. Overall, 58.3% of the participants were women, 40.0% were men, and 1.7% chose not to indicate their gender (see Figure 3).

Next, a brief explanation of the experiment was provided to the participants, followed by the reading of the informed consent letter and the confidentiality agreement. After agreeing to the terms outlined in these documents, both the investigator and the participant signed letters, and a copy of each was provided to the participant.[1] The participant was then instructed to sit in a chair positioned in front of the television (TV), adjusting the distance between the participant and the TV to 1.3 meters. The eye tracker was immediately placed on the participant, with the lens adjusted to ensure comfort. Next, the calibration process for the eye tracker began, requiring each participant to observe points (targets) on the screen appearing sequentially from 1 to 5 (see Figure 4).

Once the calibration was completed, the experiment began by displaying the images on the screen, interspersed with a white screen featuring a red dot at the center between images to prompt the participant to fix their gaze on that point before each image (see Figure 4). The images presented to the participants on the screen were displayed for 10 seconds each before the white screen with the red center dot appeared. The entire experiment lasted approximately 20 minutes per participant.

After setting up the equipment and presenting the protocol to the participant, the experiment was conducted to collect fixation data. Each participant observed images on the screen related to various handcrafted products (basketry and jars), with the primary objective of building heatmaps associated with the participants' eye movements and fixations on each aspect of the image. This approach provided valuable insights into inherent visual patterns without the influence of a specific task.

To reduce the central fixation bias introduced by the presence of the red dot at the center of the screen before each image, a temporal cropping strategy was applied to the fixation data. Specifically, 10 frames were excluded from the beginning and 10 frames from the end of each viewing sequence for every image and participant. This procedure aimed to eliminate the initial fixation driven by the red cue as well as any potential central bias that might occur toward the end of the viewing period, ensuring that the fixation data used for the heatmap generation more accurately reflected spontaneous and unconstrained visual attention.

Based on the collected visual fixation data, a data cleaning and structuring process was carried out to ensure the coherence, clarity, and consistency of the information. This involved applying cleaning, transformation, and normalization procedures to adequately prepare the data for analysis in the subsequent stages of the study.

[1] This protocol has been approved by the Ethics Committee of Universidad Adolfo Ibáñez (certificate 57/2023).

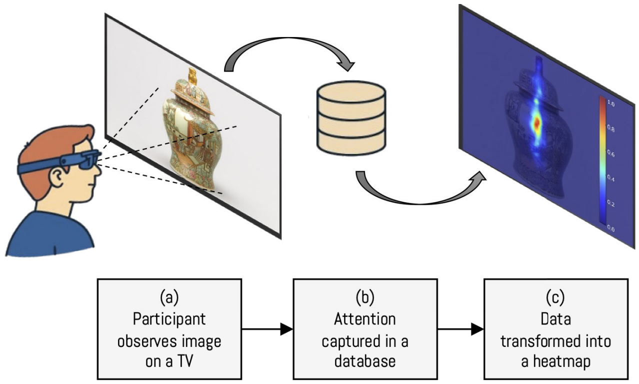
(a)
Participant observes image on a TV
(b)
Attention captured in a database
(c)
Data transformed into a heatmap

FIGURE 5. Diagram explaining the process of generating heat maps for participants.

B. HEAT MAP GENERATION

The heatmap generation phase plays a crucial role by providing a visual and quantitative representation of the participants' attention patterns, as well as the intrinsic attention mechanism of the Transformers (ViT) model during image observation. The implementation of this process was carried out separately and distinctly for both.

1) PARTICIPANT HEATMAPS

This process aimed to record and capture eye movements and fixations by measuring the gaze position and the areas where participants focused their attention during image viewing (fixations) using an eye tracker. The data obtained from tracking eye movements and fixations were filtered and normalized to generate heatmaps that visually represent the areas of greatest fixation and/or attention. This allowed for a deeper understanding of the visual patterns associated with observing handcrafted products (see Figure 5).

The eye-tracking and heatmap generation process integrated seamlessly with the participant's visual experience, ensuring the collection of accurate and relevant data for subsequent analysis. This robust methodological strategy ensured the quality and reliability of the results, providing a detailed view of the connection between visual perception and attention patterns in the context of images featuring handcrafted products such as jars and basketry.

2) TRANSFORMER HEAT MAP

To extract the model-based visual attention, we used a Vision Transformer (ViT-Base) architecture pretrained with the self-supervised DINO framework (Self-Distillation with No Labels) developed by Facebook Research [10]. The model was implemented following the architecture described by Dosovitskiy et al. [14], using a patch size of 16 and an image input size of 224 × 224.

Each experimental image was divided into patches and passed through a patch embedding layer. A learnable [CLS] token was prepended to the sequence of patch embeddings to enable a global representation of the image. Positional embeddings were added to the tokens to encode spatial information, and the resulting sequence was processed by a stack of Transformer blocks.

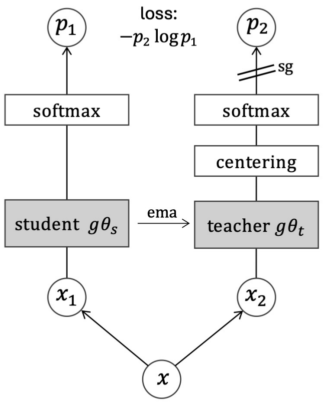

FIGURE 6. Overview of DINO self-supervised training process [10].

For attention extraction, we followed the interpretability procedure proposed in the original DINO implementation. Self-attention weights were extracted from the final Transformer block, specifically focusing on the attention from the [CLS] token to all image patches. Specifically, we extracted the attention values from the [CLS] token to all other patch tokens, which yields a spatial attention map indicating which regions of the image contributed most to the model's global representation.

The attention weights from each head were collected and reshaped to form a 2D grid. We computed the mean attention across all heads to obtain aggregate attention maps. These were then upsampled via nearest-neighbor interpolation to match the original image resolution. The resulting attention heatmaps were saved as images and CSV files for further comparison with human eye-tracking data.

This procedure ensures that the attention maps are spatially aligned with the original stimuli and interpretable in terms of localized visual importance, allowing us to compute divergence metrics between model-based and human attention.

C. EVALUATION

In this stage, a comparison was made between the distributions obtained from the heatmaps generated in the previous stage for each group (participants and the ViT model). Five metrics were used for the comparison, as detailed below.

1) KULLBACK-LEIBLER DIVERGENCE (KLD)([30], [31])

Comparing two heatmaps involves comparing two histograms or distributions. A widely used tool for measuring the difference between distributions is the Kullback-Leibler Divergence (KLD). This measure originates from information theory and is commonly used to compare distributions. Given two heatmaps with distributions P and Q (in some space X), the KL divergence from Q to P is denoted as $DKL(P||Q)$ and is defined as follows:

$$D_{KL}(P\|Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(X)}{Q(X)}\right]$$

First, both heatmaps must be normalized so that they represent valid probability distributions. The KLD formula is then applied to each pair of corresponding values in the two maps, and all the calculated values are summed to obtain the total KLD; a low value (close to 0) indicates greater similarity between the maps, while a high value indicates greater discrepancies between the maps (taking Q as the reference).

2) JENSEN-SHANNON DIVERGENCE (JSD) [32]

The Jensen-Shannon Divergence (JSD) is a statistical measure that quantifies the similarity between two probability distributions. It is based on the concept of Kullback-Leibler Divergence; however, JSD has certain advantages over KLD, particularly its symmetric nature and bounded range. The Jensen-Shannon Divergence between two probability distributions P and Q is mathematically defined as follows:

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

where $M = \frac{1}{2}(P+Q)$ is the average of the two distributions, and KLD represents the Kullback-Leibler divergence. This formulation highlights the symmetric nature of JSD, as it treats both distributions equally, unlike Kullback-Leibler divergence, which is inherently asymmetric. JSD values are bounded between 0 and 1, where 0 indicates maximum similarity and 1 indicates complete dissimilarity.

3) HELLINGER DISTANCE (HD) [23]

The Hellinger Distance is a metric that measures the similarity between two probability distributions. It is based on the square root of the distributions, making it less sensitive to slight differences compared to other metrics such as KLD. Mathematically, it is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{x}\left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2}$$

This distance takes values between 0 (when P and Q are identical) and 1 (when they are completely disjoint). It is useful in contexts where it is important to penalize slight differences between distributions less heavily.

4) SOBOLEV DISTANCE (SD) [1]

The Sobolev distance compares two distributions by considering both pointwise differences and their derivatives (or gradients), making it sensitive to the spatial variation of the distributions. It is particularly suitable for heatmaps or spatial distributions where not only value matches matter, but also the "smoothness" or "texture" of the differences. A common formulation for the Sobolev distance between P and Q is:

$$D_{\text{Sob}}(P, Q) = \sqrt{\int_X |P(x) - Q(x)|^2 dx + \lambda \int_X |\nabla P(x) - \nabla Q(x)|^2 dx}$$

where λ is a weight parameter that controls the influence of differences in the gradients. Values close to zero indicate high similarity in both the intensities and the gradient structures.

5) KOLMOGOROV-SMIRNOV DISTANCE (KSD) [28], [44]

The Kolmogorov-Smirnov distance measures the largest absolute discrepancy between the cumulative distribution functions (CDF) of two distributions. For two distributions P and Q with CDFs FP(x) and FQ(x), the metric is defined as:

$$D_{KS}(P, Q) = \sup_x |F_P(x) - F_Q(x)|$$

This metric is widely used to compare one-dimensional distributions and is especially useful when identifying significant local differences in the tails or the shapes of the distributions. It also forms the basis for the Kolmogorov-Smirnov statistical test used in nonparametric analyses.

IV. RESULTS

This section presents the results obtained from the comparison between the human attention heatmaps, generated from data collected using eye-tracking from sixty participants, and the attention maps generated by the Visual Transformers (ViT) model described by Dosovitskiy et al. The experiment was conducted using a set of images corresponding to handcrafted objects, specifically handbags and jars. For the quantitative comparison, multiple distance and divergence metrics were used, including Kullback-Leibler Divergence (KLD), Jensen-Shannon Divergence (JSD), Hellinger Distance (HD), Sobolev Distance (SD), and Kolmogorov-Smirnov Distance (KSD). These metrics allowed for the analysis of both global and local similarities and differences between the attention distributions of both approaches (human and artificial). Additionally, the results were visualized using pixel-by-pixel difference maps, which allowed for the identification of areas of greatest discrepancy (or divergence) between the participants' attention patterns and the attention estimated by the ViT. The results are discussed in detail below.

A. RESULTING PARTICIPANT HEAT MAPS

The results of the human experiment revealed that, on average, there was a predominant concentration of fixations in the central area of the handbags, with saccadic movements primarily occurring from top to bottom (or vice versa), where distinctive elements such as zippers, clasps, or different textures are generally located. This suggests manifest attention focused on visual features relevant to the object's functionality or aesthetics. In some cases, such as in cesteria_03 and cesteria_04, additional fixations and saccadic movements were observed distributed toward the lateral or upper areas, which could be related to additional details such as finishes or decorative patterns (see Figure 7). In general terms, the attention patterns tended to be vertically centered along the mid-axis of the handbags, which is consistent with the natural visual scanning strategy used for symmetrical objects [5], [24], [54].

For the jars, on average, the heatmaps revealed a vertical attention pattern that followed the elongated shape of the object, with high-density points in the middle region of the

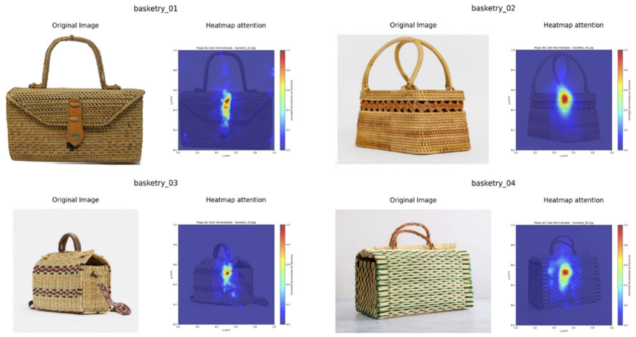

FIGURE 7. Example of participants' average attention heatmaps for basketry (first four images).

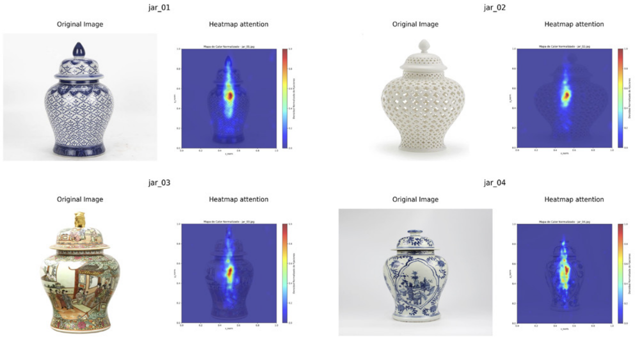

FIGURE 8. Example of participants' average attention heatmaps for the jars (first four images).

jars' bodies—an area where the jars have a more rounded and wider structure. Dispersed attention was also noted in the upper parts of some jars, which could be due to visual exploration of the lids or rims, indicating attention to structural or aesthetic features of interest to the participants (see Figure 8).

In general terms, when analyzing the heat maps, a recurring pattern was observed in which participants tended to focus their attention primarily on the central areas of the images. This phenomenon was consistent in both the basketry and jar photographs (see Figure 9), showing a visual inclination toward the vertical line running through the center of the objects.

While some fixations were detected in the peripheral areas, they were much more dispersed and had lower density. This reinforces the idea that observers' gaze preferentially focused on the main object, with less interest in exploring the surroundings or edges of the images.

These results align with well-documented patterns of human visual attention, where observers tend to focus their gaze on regions containing the greatest amount of perceptual information or perceived as structurally relevant within the stimulus. This tendency has been associated with concepts such as the perceptual center of mass, a notion describing the visual system's natural inclination to target area of balance or visual prominence within a scene [5], [24], [45]. Consequently, it is not surprising that, in this analysis, both the areas richest in information and the structurally prominent regions attracted a higher proportion of fixations.

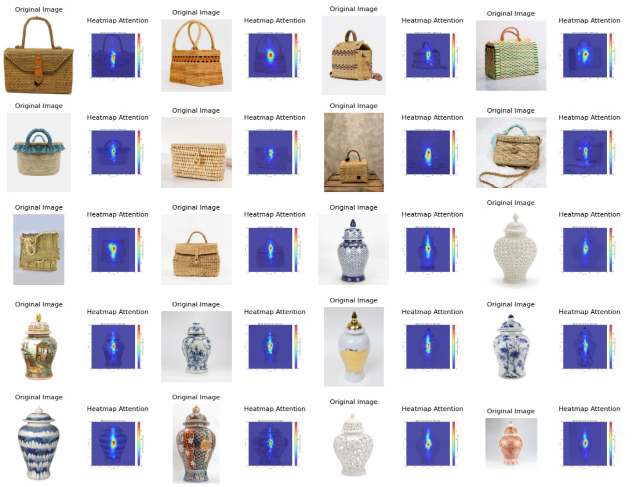

FIGURE 9. Heatmaps of average human fixations for basketry and jars images.

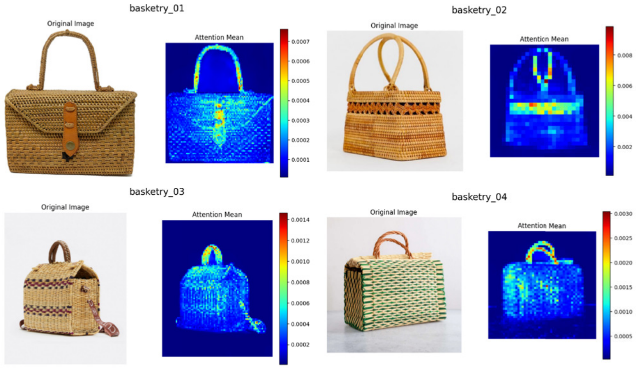

FIGURE 10. Example of ViT average attention heat maps for basketry (first four images).

B. HEAT MAPS RESULTING FROM APPLYING THE VIT MODEL

For the basketry images processed by the ViT model, attention was distributed in the central and upper areas of the handbags. For example, in the cesteria_01 image, on average, the model directed attention toward the central clasp and its vertical structure, as well as toward its handle. In cesteria_02, the model focused attention particularly on the decorative upper stripe of the basket and its handles. In images such as cesteria_03 and cesteria_04, attention was similarly focused on the handles and upper edges, highlighting the ViT's sensitivity to the structural boundaries of the baskets (see Figure 10).

In general, the model tends to identify distinctive elements such as handles, closures, and textile details in the upper and middle parts of the baskets, activating specific areas containing relevant geometric details or textures.

In the case of the jars, the model's attention maps revealed a marked preference for the upper areas of the objects. For example, in jarra_01 and jarra_02, the model concentrated attention on areas around the lids and upper edges, highlighting interest in the shapes and reliefs concentrated in those regions. In images such as jarra_03 and jarra_04,

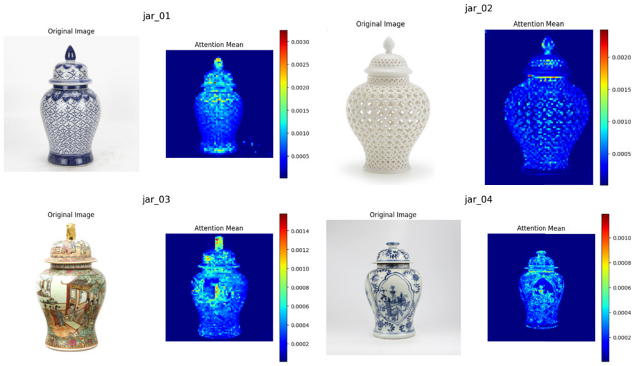

FIGURE 11. Example of ViT average attention heatmaps for the jars (first four images).

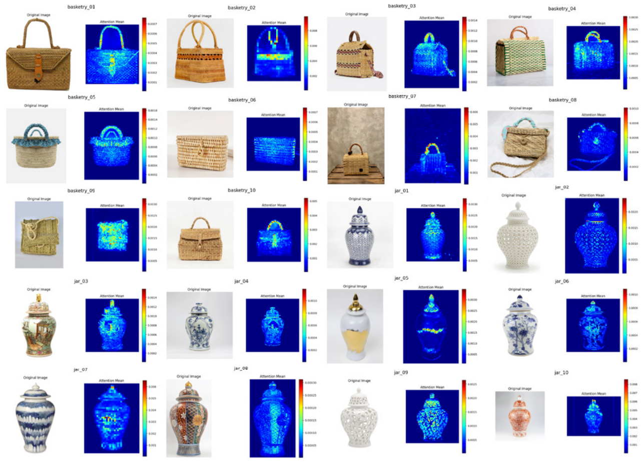

FIGURE 12. ViT-generated average attention heatmaps for the basketry and jar images.

ViT focused its attention on areas with complex decorative patterns located in the middle part of the jar (see Figure 11).

In general terms, the ViT prioritized decorative areas, reliefs, and edges, distributing its attention both to the upper regions and to specific patterns located on the body of the jar.

Overall, the heatmaps generated by the ViT showed a marked tendency to focus their attention on the structural areas and edges of the objects present in the images. In handbags, for example, the model commonly highlighted the handles, upper edges, and those areas where functional details such as clasps are typically located, particularly in zones where reliefs or geometric patterns are concentrated (see Figure 12).

When analyzing the images of the jars, the behavior was different: the ViT's attention was distributed more broadly along the contour and surface of the object. The model placed particular emphasis on the decorated areas and the reliefs adorning the jar's structure, highlighting complex visual patterns and differences in texture or color that enrich the object's surface (see Figure 12).

Overall, it can be observed that ViT consistently prioritized the edges and those areas with greater spatial complexity, such as contours, handles, or ornamental details. Although in some cases attention was also observed on the central body of the objects, the greatest concentration of attention fell on the areas that defined or structured the figure.

This suggests that ViT organizes its attention primarily on those visual elements that are most informative for defining and segmenting the objects, responding to their design, which is oriented towards identifying spatial patterns and relevant structural relationships within the scene.

C. HUMAN ATTENTION VS. VIT ATTENTION

An initial visual analysis between the attention heatmaps generated by the participants and those produced by the ViT model revealed notable differences in the visual exploration mechanisms employed by each. In the case of human participants, a concentration of fixations (overt attention) was observed in the central areas of the objects. This behavior was not only consistent across all basketry and jug images, but also followed typical patterns, such as seeking symmetry and prioritizing regions with high density. Saccadic movements, which correspond to rapid shifts of gaze between fixation points, tended to be limited in extent, remaining mostly within the fixation density. This suggests that participants prioritized certain structural or decorative features in both the bags and the jugs and did not focus their exploration on the background or edges of the images, which were not the subject of fixation by the participants.

In contrast, heat maps generated by the Visual Transformers (ViT) model reflect a distribution of attention based on weights learned about different regions of the image, without simulating fixations or saccadic movements as humans do. Instead of focusing on specific points, the model distributes its attention along edges, contours, and areas with high spatial frequency, such as textures and reliefs, thus covering the entire figure of the object. This strategy responds to an internal mechanism that assigns greater relevance to significant spatial patterns to optimize visual processing.

Therefore, in the case of human participants, attention was expressed through fixations and saccadic movements, mainly focused on the central and structurally relevant areas of objects. The ViT model simulates attention maps through a learned distribution of weights over regions of the image, based on the relevance of internal features. These attention distributions highlight contours, textures, and edges, aligning with the patterns that contribute most to the model's overall representation.

Based on the methodology described in this study, after representing the attention heatmaps (on average) for both the participants and the ViT model, the previously mentioned metrics were calculated between the results, using the participants' average attention distribution as reference. This revealed systematic differences between the two distributions for both the basketry and jar sets (see Table 1).

For the basketry images, the Kullback-Leibler Divergence (KLD) values tended to be higher (mean $\approx$ 1.98), with maximum values reaching up to 2.65 in cesteria_06, suggesting greater discrepancies between the human and ViT attention distributions. Consistently, the Jensen-Shannon

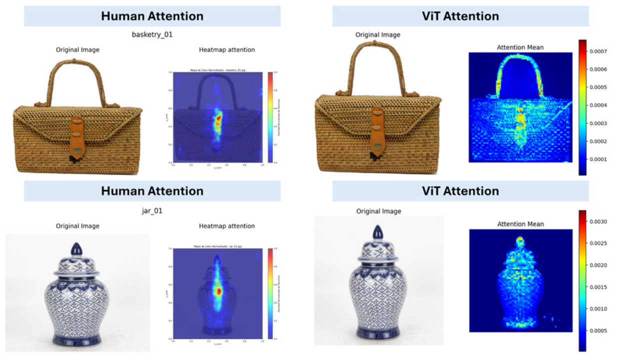

FIGURE 13. Example of comparison between human average heatmaps and ViT average heatmaps.

TABLE 1. Results of comparison metrics between human average heatmaps and ViT average heatmaps.

Imagen	KL	JSD	Hellinger	KS	Sobolev
			Cestería		
Basketry_01	2.2071	0.4611	0.7594	0.8201	0.0592
Basketry_02	1.5990	0.3705	0.6786	0.8185	0.0836
Basketry_03	1.9228	0.4315	0.7345	0.8781	0.0542
Basketry_04	1.7100	0.3907	0.6949	0.8076	0.0752
Basketry_05	2.0635	0.4494	0.7475	0.8560	0.0531
Basketry_06	2.6499	0.5033	0.7965	0.9000	0.0412
Basketry_07	2.1163	0.4595	0.7603	0.9239	0.0661
Basketry_08	2.0458	0.4346	0.7327	0.8278	0.0429
Basketry_09	1.5894	0.3695	0.6693	0.8656	0.0685
Basketry_10	1.6659	0.3895	0.6944	0.8630	0.0814
			Jarras		
Jar_01	1.5818	0.3778	0.6820	0.8882	0.0638
Jar_02	1.8600	0.4141	0.7157	0.8379	0.0581
Jar_03	2.3089	0.4670	0.7666	0.9022	0.0635
Jar_04	1.7085	0.3995	0.7056	0.9048	0.0392
Jar_05	1.8716	0.4020	0.7074	0.8574	0.0576
Jar_06	1.9654	0.4046	0.7065	0.8417	0.0452
Jar_07	1.6411	0.3684	0.6659	0.7139	0.0818
Jar_08	2.0871	0.4236	0.7209	0.7867	0.0308
Jar_09	1.9451	0.4242	0.7270	0.8806	0.0669
Jar_10	1.3852	0.3309	0.6351	0.8846	0.0770

Divergence (JSD) and Hellinger Distance metrics also showed notably high values for this set, with averages of 0.42 and 0.72 respectively, indicating that the artificial model prioritized peripheral or high spatial frequency visual elements, while human participants focused their attention on functional or central regions, such as the clasps on the handbags (see Figure 14).

For the jar images, the values of these same metrics tended to be slightly lower (mean KLD ≈ 1.82; JSD ≈ 0.40; Hellinger ≈ 0.70), with notable images such as jarra_10, which showed a KLD of only 1.38 and a Hellinger of 0.63, suggesting greater similarity between human and artificial attention in this set. This pattern could be explained by a coincidence in the orientation of attention toward ornamental patterns and the pronounced shape of the main area of the jar, which were present in both forms of analysis (manifest and covert).

For the Kolmogorov-Smirnov (KS) and Sobolev metrics, the KS metric, which measures the maximum cumulative discrepancy between two distributions, showed considerable variation in the basketry images, with values ranging from 0.81 (cesteria_04) to 0.92 (cesteria_07). In the jar images, a slightly wider range was observed, from 0.71 (jarra_07) to 0.90 (jarra_03). This dispersion indicates that significant local differences exist between the human and ViT attention maps, particularly in some jars whose morphology or decoration results in more similar distributions between the two approaches.

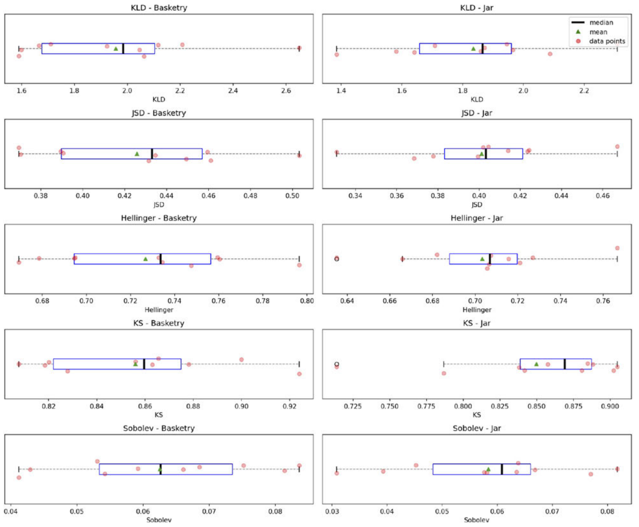

FIGURE 14. Boxplots of divergence metrics comparing attention maps from human participants and ViT, separated by object category (basketry and jars). Each panel shows the distribution of a specific metric across the 10 images in each category.

Meanwhile, the Sobolev metric, which evaluates differences in spatial gradients (that is, changes in attention across space), also revealed that for the basketry images, the values tended to concentrate between 0.041 (cesteria_06) and 0.083 (cesteria_02), suggesting that the ViT model exhibited moderate differences compared to humans in terms of how it shifted attention across regions of interest. In contrast, the jar images showed a wider range, from 0.030 (jarra_08) to 0.081 (jarra_10), indicating that in some cases (such as jarra_08), the ViT exhibited spatial attention very similar to that of humans in terms of visual focus transition, while in other cases (such as jarra_10), greater discrepancies were observed in how attention shifted between regions.

Overall, these results show that local metrics, such as KS and Sobolev, capture more subtle differences between distributions compared to KLD, JSD, and Hellinger metrics. In the basketry dataset, local differences tended to remain stable, while in the jug dataset, more pronounced variations were observed. This may be due to the diversity of decorative patterns and the more defined geometric structure of the jugs, which could influence how the ViT model internally distributes its attention (i.e., how it assigns its attention weights).

In general, the metrics supported the visual results described above. The basketry images showed greater discrepancies between the ViT attention maps and human

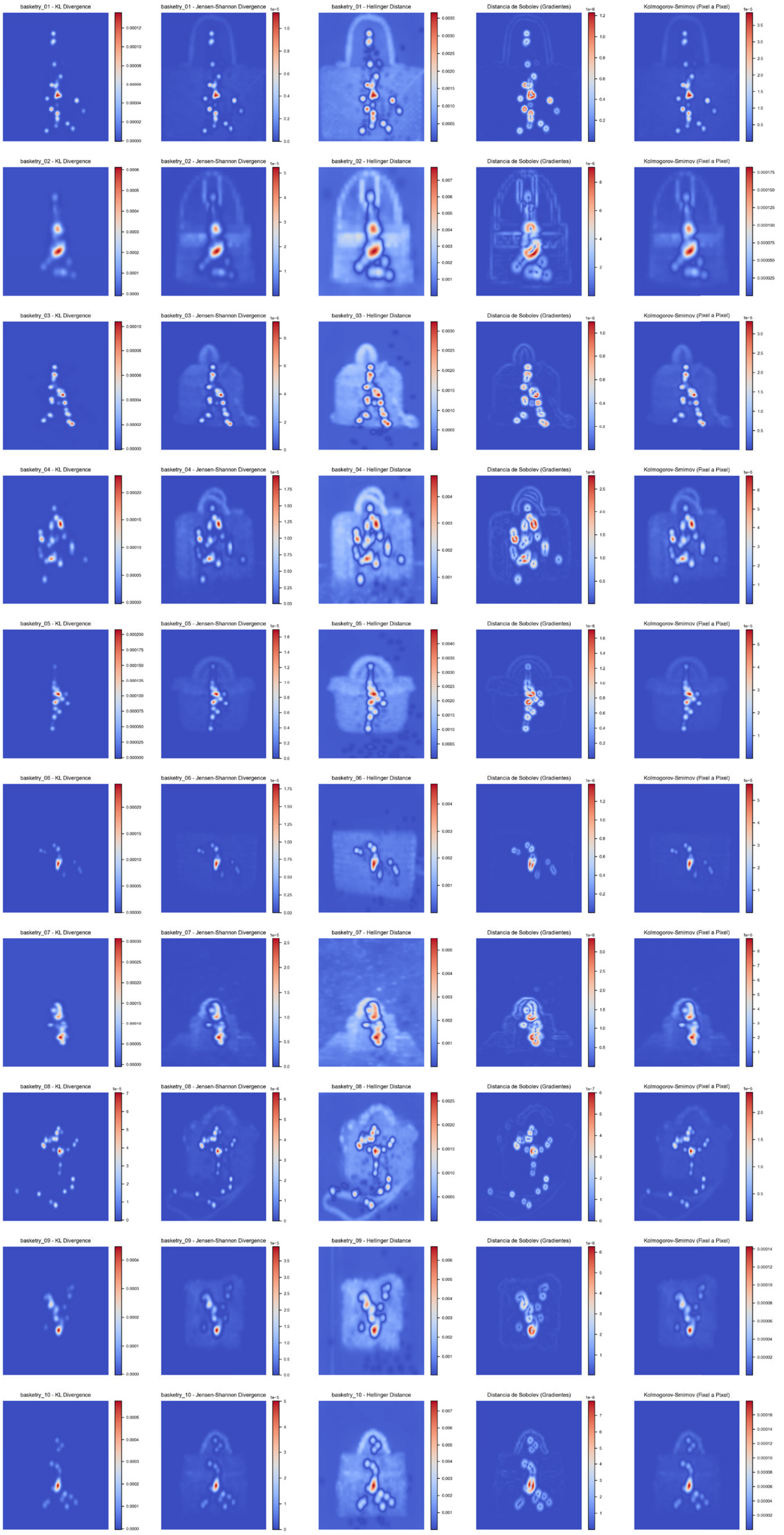

FIGURE 15. Pixel-by-pixel differences between the distributions representing human attention heatmaps and the distribution representing ViT attention heatmaps for basketry according to different metrics.

patterns, which can be attributed to the structural complexity of these objects. While the ViT tended to highlight edges and lines through a learned distribution of weights over the image, human participants focused their attention on functional or

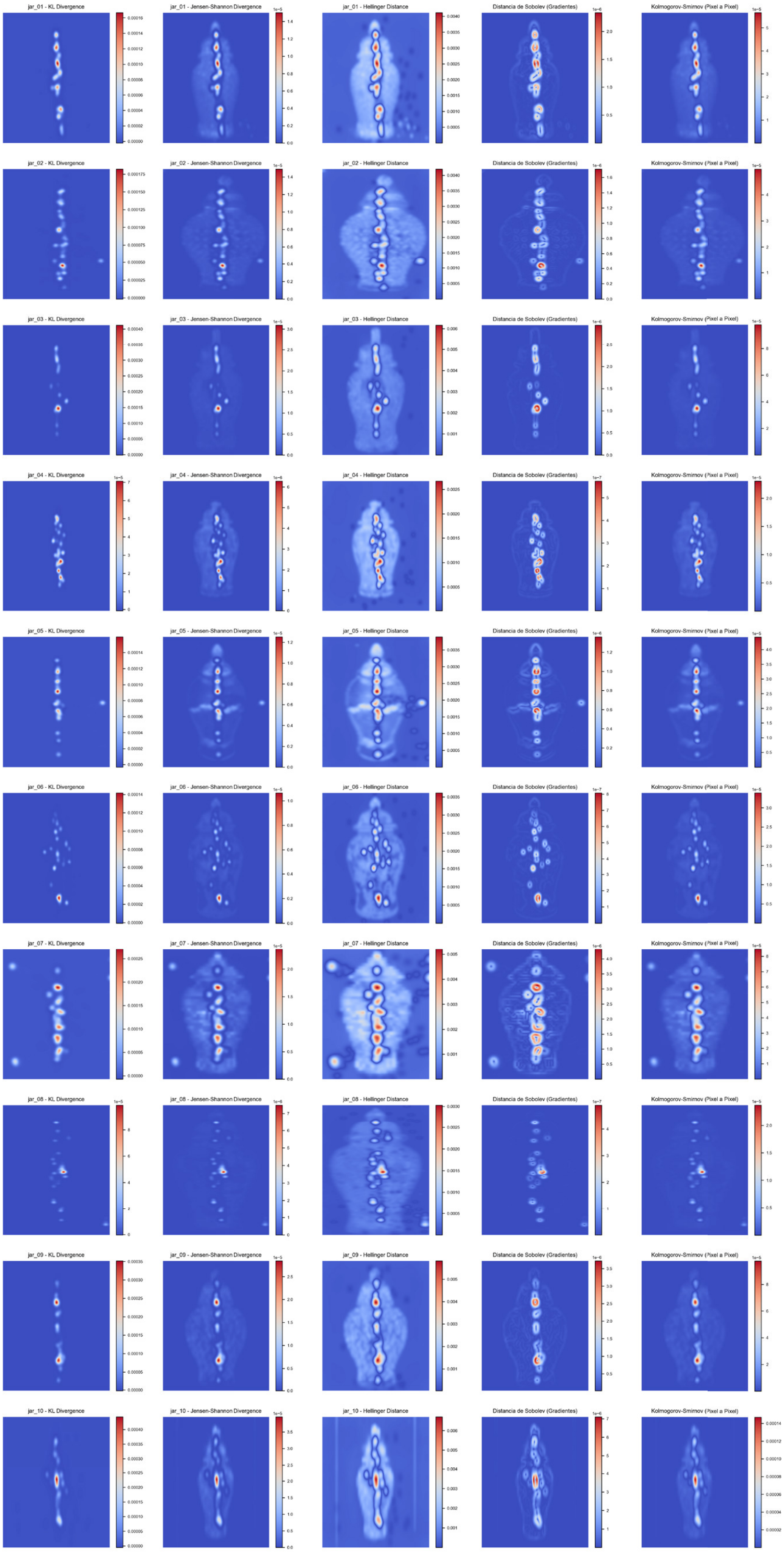

FIGURE 16. Pixel-by-pixel differences between the distributions representing human attention heatmaps and the distribution representing ViT attention heatmaps for jar according to different metrics.

perceptually relevant areas. In the case of the jugs, there was greater agreement between the two types of attention, possibly because both the ViT and humans focused on the ornamental patterns on the body and upper edges. However, in the central region of the jars' bodies—which is more rounded and wider—there were still notable differences in the distribution of attention (see Figures 15 and 16).

V. CONCLUSION

The comparative analysis between the heatmaps representing human visual attention and those generated by the Visual Transformers (ViT) model demonstrated that, although computational models have advanced, they still present significant limitations in replicating the perceptual behavior of human fixations. Despite recent advances in visual attention models such as Visual Transformers, multiple studies have noted that these models tend to capture general attention patterns but fail to match human visual perception when confronted with complex scenes [10], [24], [45]. Human visual attention does not depend solely on the physical characteristics of the stimulus but is profoundly influenced by cognitive factors such as context, intention, and the observer's prior experience, adding a layer of complexity that is difficult to emulate computationally [24], [45], [54].

The images used in the experiment, focused on handcrafted objects such as handbags and jars, allowed us to observe how participants tended to fix their attention on specific areas, particularly the central parts of the objects, following vertical saccadic movements (from top to bottom or vice versa), where decorative, structural, or functional elements are usually located. In the case of the handbags, fixations were concentrated on details such as clasps, handles, and textures, whereas for the jars, attention focused on formal elements such as the lids and the curves of the object's body. The ViT model, for its part, managed to capture some of these areas of interest, showing a tendency to distribute attention more uniformly across the entire image, including regions that did not necessarily coincide with human fixations. This suggests that artificial attention still lacks the contextual sensitivity that characterizes human perception and actively guides gaze direction.

The metrics used to assess the similarity between the heatmaps (Kullback-Leibler Divergence (KLD), Jensen-Shannon (JSD), Hellinger Distance (HD), among others) revealed differences between the two approaches. Greater alignment was observed in the central regions of the images, where fixations are usually more prominent, but greater divergence appeared in the peripheral areas of the images, especially those with lower visual load. These observations were reinforced by the pixel-by-pixel difference maps, which visually identified the areas where the ViT differed the most from human attention. This type of analysis provides empirical evidence that the ViT model still fails to fully replicate human perceptual fixations, which could limit its applicability in contexts requiring a high degree of manifest visual interpretation.

In summary, while the ViT model represents a significant advancement in simulating visual attention, the results of this research show that there is still considerable progress to be made in approaching the behavior of human perception. Understanding the differences between the two approaches not only helped identify weaknesses in current models but also encouraged reflection on the aspects of human perception that artificial intelligence has yet to adequately model. This critical understanding is essential if such technologies are to be applied in sensitive fields such as education, advertising, digital art, or human-machine interaction, where well-directed visual attention can make a significant difference in the effectiveness and comprehension of the presented content.

VI. DISCUSSIONS

In the future, it is expected that this topic will continue to be studied, taking into account the difference between the distributions of the attention fixations of the participants and those of the ViT model to train the model in order to minimize this difference, considering that equivalent behaviors were obtained for the metrics. The heads of the attention mechanism were also analyzed to verify which of them behaved as close as possible to the participants' results.

Additionally, although this study focused on the direct comparison between human attention and the attention generated by a Vision Transformer trained with a self-supervised approach (DINO), we recognize the value of contrasting these results with more classical models of visual saliency or attention, such as Grad-CAM or Itti-Koch. These models could provide additional context to better understand how ViT-based attention differs in nature and focus. We have not included such comparisons in the present work to maintain a clear scope, but we plan to incorporate them in future research as part of a broader evaluation framework.

On the other hand, it would be pertinent to extend the analysis to other transformer-based architectures beyond the original ViT, such as Swin Transformer, DeiT, or TNT. These models share basic attention mechanisms but include architectural innovations that may influence their attention behavior. Studying whether the similarities and discrepancies observed with human attention hold across all these variants remains an important direction for future work.

Furthermore, we propose generalizing these studies, using the same (or similar) data and the same principles, but comparing large multimodal language models (LLMs) (such as ChatGPT, Grok, Gemini, or Claude) that have just made public their visual ability to distinguish which ones exhibit behaviors close to human fixations.

Finally, it is important to note that while our study is based exclusively on external visual data and does not include neurological measurements, recent literature suggests potential analogies between the operation of ViT models and certain neural processes. In particular, Ramezanpour and Fallah propose that ViTs may function similarly to perceptual systems mediated by the mid-level temporal cortex, especially in their ability to capture global visual features [42]. This interpretation is consistent with the spatial distribution observed in the attention maps generated by the ViT model in our study. Buschman and Miller. emphasize the role of the prefrontal cortex in top-down attentional control [8], our data do not allow us to draw conclusions about such higher-order mechanisms. Nonetheless, the parallel with mid-level perceptual processing offers a compelling framework for understanding the behavior of ViTs in relation to human attention and opens future lines of inquiry into the biological plausibility of transformer-based vision models.

VII. ADDITIONAL EXPERIMENT RESULTS

The appendices present the final results of applying the metrics described in Section IV to compare the heatmaps of human fixations with those obtained from the Visual Transformer model. Figures 15 and 16 illustrate the results for the experiment discussed in Section IV.

A. DATA AVAILABILITY STATEMENT

https://github.com/luis-oliveros/Visual_Attention_and_ViT (accessed on 05 August 2025)

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, "Pure and applied mathematics," in *Sobolev Spaces*, 2nd ed., Amsterdam, The Netherlands: Elsevier, 2003.

[2] N. Akhtar and J. T. Enns, "Relations between convert orienting and filtering in the development of visual attention," *J. Experim. Child Psychol.*, vol. 48, no. 2, pp. 315–334, Oct. 1989.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[4] D. Belyusar, A. C. Snyder, H.-P. Frey, M. R. Harwood, J. Wallman, and J. J. Foxe, "Oscillatory alpha-band suppression mechanisms during the rapid attentional shifts required to perform an anti-saccade task," *NeuroImage*, vol. 65, pp. 395–407, Jan. 2013.

[5] J. G. R. Berrío, *Percepción Visual*, 2nd ed. Medellín, Colombia: Fondo Editorial Pascual Bravo, 2019.

[6] J. W. Bisley, "The neural basis of visual attention," *J. Physiol.*, vol. 589, no. 1, pp. 49–57, Jan. 2011.

[7] C. D. Blair and J. Ristic, "Attention combines similarly in covert and overt conditions," *Vision*, vol. 3, no. 2, p. 16, Apr. 2019.

[8] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, Mar. 2007.

[9] M. Cadoni, S. Nixon, A. Lagorio, and M. Fadda, "Exploring attention on faces: Similarities between humans and transformers," in *Proc. 18th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2022, pp. 1–8.

[10] M. Caron, H. Touvron, I. Misra, H. Jeǵou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Apr. 2021, pp. 9630–9640.

[11] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021.

[12] N. Cowan, "Activation, attention, and short-term memory," *Memory Cognition*, vol. 21, no. 2, pp. 162–167, Mar. 1993.

[13] T. J. Crawford and H. J. Müller, "Spatial and temporal effects of spatial attention on human saccadic eye movements," *Vis. Res.*, vol. 32, no. 2, pp. 293–304, Feb. 1992.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] J. Driver and C. Spence, "Cross-modal links in spatial attention," *Phil. Trans., Biol. Sci.*, vol. 353, no. 1373, pp. 1319–1331, 1998.

[16] U. Engelke, A. Duenser, and A. Zeater, "Cognitive correlates of overt and covert attention in visual search," in *Proc. IEEE 13th Int. Conf. Cognit. Informat. Cognit. Comput.*, Aug. 2014, pp. 197–202.

[17] A. F. Frielink-Loing, A. Koning, and R. van Lier, "Distinguishing influences of overt and covert attention in anticipatory attentional target tracking," *J. Vis.*, vol. 17, no. 4, p. 3, Jun. 2017.

[18] R. Groner and M. T. Groner, "Attention and eye movement control: An overview," *Eur. Arch. Psychiatry Neurological Sci.*, vol. 239, no. 1, pp. 9–16, Jan. 1989.

[19] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Dec. 2023.

[20] M.-H. Guo, T.-X. Xu, J. Liu, Z.-N. Liu, P.-T. Jiang, T. Mu, S.-H. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[21] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on visual transformer," 2020, *arXiv:2012.12556*.

[22] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102417.

[23] E. Hellinger, "Neue begründung der theorie der quadratischen formen von unendlichvielen veränderlichen," *J. für die reine und Angew. Math.*, vol. 136, pp. 210–271, Apr. 1909.

[24] J. Henderson, "Human gaze control during real-world scene perception," *Trends Cognit. Sci.*, vol. 7, no. 11, pp. 498–504, Nov. 2003.

[25] J. M. Henderson, A. Pollatsek, and K. Rayner, "Covert visual attention and extrafoveal information use during object identification," *Perception Psychophysics*, vol. 45, no. 3, pp. 196–208, May 1989.

[26] A. W. Inhoff, A. Pollatsek, M. I. Posner, and K. Rayner, "Covert attention and eye movements during reading," *Quart. J. Experim. Psychol. Sect. A*, vol. 41, no. 1, pp. 63–89, Feb. 1989.

[27] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1489–1506, Jun. 2000.

[28] A. L. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, Feb. 1933.

[29] L. V. Kulke, J. Atkinson, and O. Braddick, "Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking," *Frontiers Hum. Neurosci.*, vol. 10, Nov. 2016, Art. no. 592.

[30] S. Kullback, *Information Theory and Statistics*. Mineola, NY, USA: Courier Corporation, 1997.

[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[32] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[33] P. Mehrani and J. K. Tsotsos, "Self-attention in vision transformers performs perceptual grouping, not attention," 2023, *arXiv:2303.01542*.

[34] M. Milanova, "Visual attention in deep learning: A review," *Int. Robot. Autom. J.*, vol. 4, no. 3, pp. 154–158, May 2018.

[35] O. Moutik, H. Sekkat, S. Tigani, A. Chehri, R. Saadane, T. A. Tchakoucht, and A. Paul, "Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?" *Sensors*, vol. 23, no. 2, p. 734, Jan. 2023.

[36] N. Park and S. Kim, "How do vision transformers work?" 2022, *arXiv:2202.06709*.

[37] T. Parr and K. J. Friston, "Attention or salience?" *Current Opinion Psychol.*, vol. 29, pp. 1–5, May 2019.

[38] M. Posner, F. Friedrich, J. Walker, and R. Rafal, "Neural control of the direction of covert visual orienting," in *Proc. Meetings Psychonomic Soc.*, 1983, pp. 1–4.

[39] M. Posner S. E. Petersen, "The attention system of the human brain," *Annu. Rev. Neurosci.*, vol. 13, no. 1, pp. 25–42, Jan. 1990.

[40] M. I. Posner, Y. Cohen, and R. D. Rafal, "Neural systems control of spatial orienting," *Phil. Trans. Roy. Soc. London. B*, vol. 298, no. 1089, pp. 187–198, 1982.

[41] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" 2022, *arXiv:2108.08810*.

[42] H. Ramezanpour and M. Fallah, "The role of temporal cortex in the control of attention," *Current Res. Neurobiol.*, vol. 3, Apr. 2022, Art. no. 100038.

[43] S. Satoh and S. Miyake, "A model of overt visual attention based on scale-space theory," *Syst. Comput. Jpn.*, vol. 35, no. 10, pp. 1–13, Sep. 2004.

[44] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Statist.*, vol. 19, no. 2, pp. 279–281, Jun. 1948.

[45] B. W. Tatler and B. T. Vincent, "Visual attention and cognitive control: Evidence from eye movements in scene perception," *Trends Cognit. Sci.*, vol. 11, no. 11, pp. 520–528, 2007.

[46] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[47] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?" 2021, *arXiv:2105.07197*.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Sep. 2017, pp. 5998–6008.

[49] L. M. Ward, "Covert focussing of the attentional gaze," *Can. J. Psychol. /Revue canadienne de psychologie*, vol. 39, no. 4, pp. 546–563, Dec. 1985.

[50] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 72227–72249, 2023.

[51] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, vol. 3, pp. 2048–2057.

[52] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, Mar. 2022.

[53] Y. Yang, L. Jiao, X. Liu, F. Liu, S. Yang, Z. Feng, and X. Tang, "Transformers meet visual learning understanding: A comprehensive review," 2022, *arXiv:2203.12944*.

[54] A. L. Yarbus, *Eye Movements and Vision* (1967). New York, NY, USA: Springer, 1967.

LUIS GUILLERMO OLIVEROS PIÑERO (Student Member, IEEE) received the Licenciatura degree in mathematics and the M.S. degree in applied mathematics from the Universidad de Carabobo, Venezuela, in 2011 and 2017, respectively, and the Diploma degree in finance from the Universidad del Zulia, in 2021. He is currently pursuing the Ph.D. degree in data science with Universidad Adolfo Ibáñez, Santiago, Chile. Since 2023, he has been a Professor and an Assistant Professor in various undergraduate and graduate courses at the Universidad Adolfo Ibáñez, including probability and statistics, exploratory data analysis, predictive analytics, and applications of deep learning. He is proficient in Python, R, SQL, LaTeX, Tableau, Power BI, and advanced spreadsheet modeling. His research interests include statistical learning, data visualization, machine learning applications in industrial, academic and business contexts, and collaborative tools for data-driven decision making.

MIGUEL CARRASCO received the Ph.D. degree (Hons.) in informatics from Institut des Systèmes Intelligents et de Robotique (ISIR), Pierre et Marie Curie University - Paris 6, in 2010, and the Ph.D. degree in engineering sciences in computer science from Department of Computer Science, Pontificia Universidad Católica de Chile, under a jointly supervised Ph.D. program. He is currently teaching at the Escuela de Ingeniería Informática y Telecomunicaciones, Universidad Diego Portales (UDP). His research interests include the development of automatic algorithms based on image processing and computer vision theory. His main research topics are image processing for biology (virus and pollen segmentation), computer vision for industrial inspection, industrial Internet of Things (IIoT), and failure prediction based on pattern recognition. Past and current research includes human-computer interaction, users' gaze gesture prediction with eye-trackers, automatic multiple visual inspection, and telemedicine based on IoT (heart sensors). He was awarded a scholarship from the Collège Doctoral Franco-Chilien in 2007.

JOSÉ ARANDA (Member, IEEE) received the B.S. degree in industrial engineering and the M.S. degree in data science from Adolfo Ibáñez University, Santiago, in 2024. From 2021 to 2024, he was a Research Assistant at the Computer Vision Laboratory related to vision transformers. His research interests include the comparison of the ViT attention module with human vision attention, where he has a strong focus on neuroscience and neural network architecture. Also, he is part of a research team focused on projects around the use of vision attention in the mining industry, including automated personal protective equipment detection and document analysis AI agents.

CÉSAR GONZÁLEZ-MARTÍN received the degree in fine arts and the Ph.D. degree in arts from the University of Granada, Spain. He is currently a Lecturer and a Researcher with the University of Cordoba, Spain. He has participated in different European projects, such as RRREMAKER—Reuse Reduce Recycle AI-based platform for automated and scalable Maker culture in Circular economy (H2020-MSCA-RISE), WARMEST—loW Altitude Remote sensing for the Monitoring of the state of Cultural hEritage Sites: Building an inTegrated model for maintenance (H2020-Marie SkŁodowska-Curie Actions-RISE-2017), and GLOCALFINEART—Global cOntemporary art market: The intrinsiC and sociologicAL components of FINancial and artistic valuE of ARTtworks (FP7). He has also collaborated in other research projects, such as ARTAPP—Artes Visuales, Gestión del Talento y Marketing Cultural: Estrategias Para la Construcción de Marca y Desarrollo de Una Red Para la Promoción y Difusión de Jóvenes Artistas by the Spanish Research Program, and Estrategias Para la Comunicación del Patrimonio a Través del Arte Contemporáneo Transferencias Culturales Entre Artesanía, Arte y Diseño' funded by CEIBioTic, University of Granada. He received the International Mention and Extraordinary Award from the University of Granada for the Ph.D. degree.