A Multi-Criteria Approach for Gaze Analysis Similarity in Paintings

Tess Masclef[a], Mihaela Scuturici[b], Tetiana Yemelianenko[c] and Serge Miguet[d]
Université Lumière Lyon 2, CNRS, Ecole Centrale de Lyon, INSA Lyon, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69007 Lyon, France
{*tess.masclef, mihaela.scuturici, tetiana.yemelianenko, serge.miguet*}*@univ-lyon2.fr*

Keywords: Content-Based Image Retrieval, Gaze Estimation, Digital Humanities, Gaze Analysis, Paintings, Object Detection, Depth Estimation.

Abstract: In the fields of art history and visual semiotics, analysing gazes in paintings is important to understand the artwork, and to find semantic relationships between several paintings. Thanks to digitization and museum initiatives, the volume of datasets on artworks continues to expand, enabling new avenues for exploration and research. Artificial neural networks, trained on large datasets are able to extract complex features, and visually compare artworks. This comparison could be done by focusing on the objects present in the paintings, and matching paintings with high object co-occurrence. Our research takes this further by studying the way objects are viewed by characters in the scene. This study proposes a new approach that combines methods for gaze-based and visual-based similarity, to encode and use gaze information for finding similar paintings, while maintaining a close visual aspect. Experimental results which integrate the opinions of domain experts, show that these methods complement each other. Quantitative and qualitative assessments confirm the results from the combination of gaze and visual analysis. Thus, this method improves existing visual similarity queries and opens up new possibilities to retrieve similar paintings according to user-specific criteria.

1 INTRODUCTION

In the process of analysing a painting, specialists use similar artworks, with several criteria including: colours, objects and characters disposition, and the gaze of characters inside the painting. In this paper, we explore the link between visual and gaze similarities in artworks.

For centuries, gaze has been a crucial means for artists to convey messages, narratives, emotions and social or cultural aspects of their time. By analysing the interplay of gazes in a scene, one can decipher the artist's intention; revealing emotions, thoughts, and underlying themes that eyes communicate without words.

In particular, a character gazing at an object conveys a specific symbolic or literal meaning. For example, looking at a book may represent the search for knowledge, spirituality or intellectual exploration, which allows a layered interpretation of the artwork.

Beyond the visual, the presence of a particular object plays a role in the enrichment of the composition and narrative through symbolism and cultural context. The recurring presence of certain objects in the paintings may also help us to classify the painting's genre. Indeed, some objects are specific to a particular genre such as the bowl of fruit in still-life paintings.

Searching for similarities between artworks in a large corpus is a time-consuming task. The increasing digitization of paintings has enabled the creation of datasets that can be used by artificial neural networks (ANNs). Typically, we use these ANNs to extract complex features from images or from multi-modal data including images and texts. Among recent architectures, CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021) uses a ViT-type transformer for visual features and a language model for textual features, projecting both into a common space to capture nuances and contextual details. This network makes easier precise image comparison, classification, and natural language interpretation, making CLIP ideal for applications needing deep image understanding. We only use CLIP to extract the visual features of images. We do not exploit the associated textual data. The model is used as a pre-trained visual encoder to generate embeddings that capture rich semantic information from images. These features are

[a] https://orcid.org/0009-0008-2783-8643
[b] https://orcid.org/0009-0009-6127-8843
[c] https://orcid.org/0000-0003-4559-0748
[d] https://orcid.org/0000-0001-7722-9899

Masclef, T., Scuturici, M., Yemelianenko, T. and Miguet, S.
A Multi-Criteria Approach for Gaze Analysis Similarity in Paintings.
DOI: 10.5220/0013152800003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 123-134
ISBN: 978-989-758-728-3; ISSN: 2184-4321

then used by an algorithm for efficient high dimensional search space, like $k-$NN ($k-$Nearest Neighbours) and search trees algorithms. The distance used in this algorithm is the Angular distance (based on the cosine similarity). Calculating the distance between the query vector and each vector in the dataset is very costly. So to reduce the complexity of the computation time we choose to use an approximate nearest neighbour search algorithm, ANNOY[1] (Li et al., 2019) proposed in 2015 for the Spotify platform by Erik Bernhardsson.

As shown in Figure 1: the main character in the request image is looking at a crucifix on a skull, while the characters in the first and third retrieved images are looking out of frame with the skull present in the scene. The character in the second retrieved image is looking at a book. This example shows that existing ANNs using CLIP approaches do not encode relational information such as the links between the gaze and the objects observed.

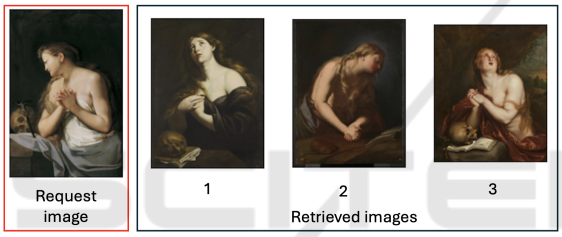


Figure 1: Results for a visual similarity search (3 retrieved images) using CLIP approach: in the request image, the objects looked at are the crucifix and the skull.

Just as it makes sense to match two texts based on word co-occurrences, two images with common objects can also be matched. But it is even more relevant to match two images where the main characters are looking at the same objects, in the same way. The objects looked at by the characters are called objects of contemplation. Therefore, the aim of this study is to develop a tool that identifies images in which the characters are looking similarly at the same object of interest as the query character, using a similarity measure to quantify this information.

This approach allows us to create a corpus of similar "gaze to object" configuration, by associating paintings with the same configuration. In summary, we present the following contributions for paintings containing at least one character:

- We propose to generate a 3D field of view cone from a gaze estimator and an image depth estimator for paintings in order to detect the objects inside the FOV.
- We propose a query tool based on gaze-to-object vector representation.
- We propose to combine global visual similarity and similarity based on objects of contemplation. We show that the relative weight given to visual similarity and objects of contemplation similarity plays, an important role in the perception of resemblance between two paintings.

2 RELATED WORK

Our method is composed of three main parts: gaze analysis, object detection and content-based image retrieval, discussed in the following sections.

2.1 Gaze Analysis Based on 3D Gaze Estimation

Many studies investigating gaze direction are primarily trained on photographic datasets. We propose to apply this study on paintings, which presents a challenge because the image is the result of a creation by a painter and not a physical measurement of reality (photography).

One of the first works to estimate gaze direction is (Recasens et al., 2015). They use an architecture divided into two paths, a saliency path and a gaze path; they are able to select objects in the scene likely to be gazed at by discovering how to extract head position and gaze orientation. Their approach is among the first deep learning approaches for 2D gaze tracking. To evaluate their method, they propose a new reference dataset, *Gaze Follow* that has become a reference used in several works (Chong et al., 2018) (Aung et al., 2018) (Lian et al., 2018).

Using a 2D field of view (FOV) in case of paintings, and images in general, is prone to errors: a background object can be far away from the axis of the gaze and nevertheless be projected in the 2D FOV. To simulate human gaze behaviour in 3D space, (Fang et al., 2021) propose a three-phase approach. A coarse-to-fine strategy determines 3D gaze orientation from head pose, separating it into planar and depth-based components. The Double Attention Module (DAM) then uses planar gaze to set the field of view and mask obstructions. Finally, DAM's dual attention locates whether the gaze target is inside the image and precisely identifies it.

To tackle human biases and physically impossible predictions, (Horanyi et al., 2023) propose to use a 3D depth and probability map of the joint field of view to estimate the joint attention target (JAT) of the people

[1] https://github.com/spotify/annoy

in the scene.

(Tian et al., 2023) develop *FreeGaze*, a framework for estimating gaze in facial videos, using a novel method to detect landmarks and reduce computational costs. Their dual-branch CNN, FG-Net, is tested on MPIIGaze and EyeDiap datasets to analyse the contributions of eye and full-face regions to gaze estimation, providing insights for future network size reduction.

2.2 Object Detection and Recognition in Fine Art

The task of object detection in an image consists of locating and associating a label to each object. A challenging aspect of object detection in historical paintings ($XV^{th} - XVIII^{th}$ century) is that some ancient objects do not appear in modern image datasets.

(Gonthier et al., 2018) developed a multiple instance learning method for weakly supervised object detection in paintings. This approach enables the learning of new classes dynamically from globally annotated databases, thereby eliminating the need for manual object annotation. Additionally, they introduce the IconArt database, specifically designed for conducting detection experiments on unique classes that cannot be learned from photographs, such as religious characters or objects.

In the same way, (Smirnov and Eguizabal, 2018) proposed an automatic detection of objects in images using deep learning, as well as a set of strategies to overcome the lack of labelled data, rare in this field.

A new dataset for the classification of iconography was introduced by (Milani and Fraternali, 2021), notably for the task of identifying saints in Christian religious paintings. For this purpose, they apply a convolutional neural network model, where they achieve good performance. Indeed, they show that the network focuses on the iconic patterns characterizing the saints.

More recently, (Reshetnikov et al., 2022) proposed DEArt, an object detection and pose classification dataset designed to serve as a reference for paintings between the XII^{th} and $XVIII^{th}$ centuries. This dataset contains $15,000$ images annotated on 70 object classes. Their results show that object detectors for the cultural heritage domain can achieve a level of accuracy comparable to state-of-the-art models for generic images, thanks to transfer learning.

A process for training neural networks to locate objects in art images was proposed and evaluated by (Kadish et al., 2021). Using AdaIN style transfer (Huang and Belongie, 2017) and the COCO dataset (Lin et al., 2014), they generate a dataset for training and validation. This dataset is used to fine-tune Faster R-CNN (Ren et al., 2016) object detection network, which is then tested on the existing People-Art test dataset (Westlake et al., 2016).

2.3 Content-Based Image Retrieval (CBIR) in Fine Art

To the best of our knowledge there is no such method for obtaining a list of similar paintings based on objects of contemplation, but a number of methods have been proposed for performing a visual similarity search.

For a given query painting, we want to retrieve similar paintings according to a given search axis (such as visual criteria, painting style and so on.). The most commonly used method is to calculate distances between the representation of the query features and those of a dataset of artistic corpus.

Deep learning can be used for image feature extraction, with artificial neural networks (convolutional or transformer-based) outperforming man made methods in extracting complex features like colour, texture, and composition. These networks, pre-trained on large datasets like ImageNet21k (Ridnik et al., 2021), excel in classification tasks but are often seen as "black-box" systems. To refine similarity searches between paintings, networks can be fine-tuned on tasks such as genre and style recognition (Masclef et al., 2023). (Tan et al., 2021) extracted high-level features from paintings to measure similarity and significantly improve content-based image retrieval. (Zhao et al., 2022) applied CNNs to art-related tasks, showing that fine-tuning pre-trained networks enhances generalization, known as big transfer learning (BiT). Their models effectively retrieve paintings, including computer-generated ones, by analysing various similarity aspects.

To conclude this section, to our knowledge there is no method for obtaining a list of similar paintings based on objects of contemplation. Our proposition is to combine gaze estimation and object detection in paintings to derive a new image retrieval method based on the proximity of looking at similar objects.

3 METHODOLOGY

3.1 Objects of Contemplation Similarity

Our pipeline is split into four steps. First, we estimate the gaze direction using the predicted visual point of interest and face position. Next, we generate a 3D

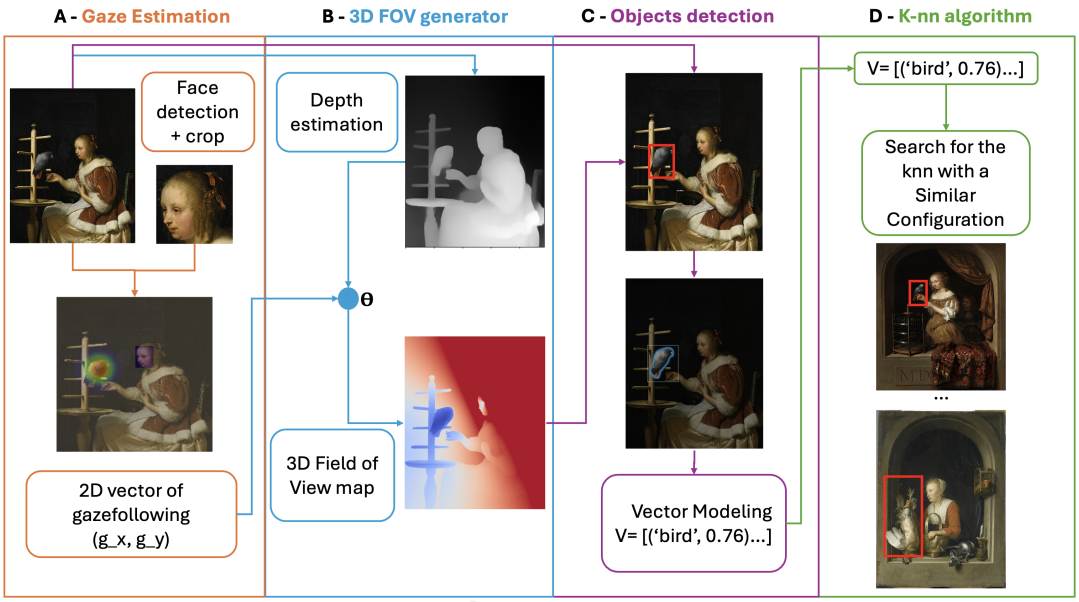


Figure 2: Objects of contemplation similarity pipeline.

field of view (FOV) from this estimation and the depth map. The third step involves detecting objects inside the FOV. Finally, we apply a $k-$Nearest Neighbours algorithm to search for similarities. The pipeline is illustrated in Figure 2.

3.1.1 2D Gaze Estimation

To estimate 2D gaze direction and point of visual interest, we choose to use the architecture of (Gupta et al., 2022). Their architecture is divided into three parts: the human-centric module which takes input from an image and head position, and outputs a 2D gaze vector (which is used to construct a gaze cone); the scene-centric module which processes the original image to produce saliency feature maps; and the prediction module which uses the saliency map to regress a gaze heat map and predict an in-out gaze classification score.

This architecture exploits different modalities: the image, the depth estimation (depth map) and the pose estimation (pose map). This system allows for the individual use of each modality or their combination through attention mechanism. We choose to only use the image modality, which performs almost as well as the combination of three modalities but at a much lower computational cost. Additionally, our results compare favourably with the state-of-the-art, demonstrating the effectiveness of our choice (section 4.2).

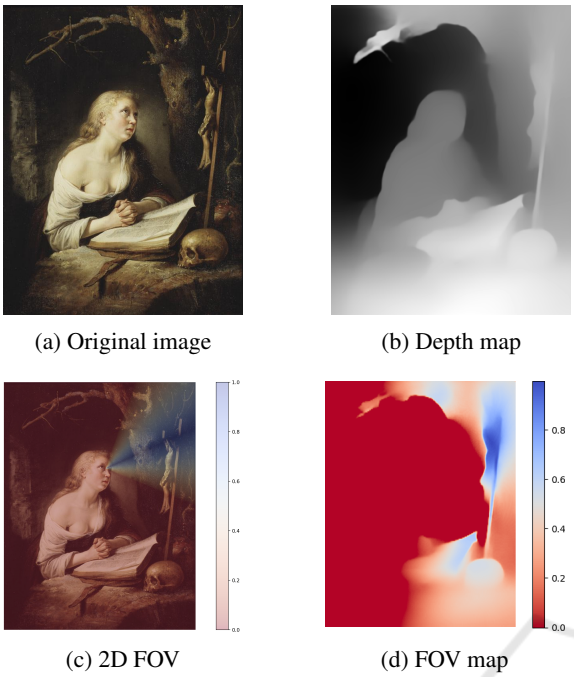
We use the simplified architecture of (Gupta et al., 2022) to predict the gaze direction of characters in paintings by replacing the face detector (FaceBoxes (Zhang et al., 2017)) by YOLO5Face (Qi et al., 2022), which detected more faces in paintings according to our empirical tests. Methods for estimating the visual point of interest in 3D in paintings yield poorer results than those in 2D. Therefore, we propose to start with a 2D estimate, then to build the 3D information, with the help of depth estimation.

3.1.2 3D FOV Generator

The construction of a 2D field-of-view cone (Figure 3c) is less relevant than that of a 3D cone, because in the former case, all objects are perceived by the machine on the same plane, whereas in the latter, depth is taken into account, enabling a more realistic representation of the environment. For example, if a character is looking at the foreground, objects in the background will not be taken into account.

We generate the depth map using **MiDaS v3.1** (Birkl et al., 2023), which produced convincing depth map at reasonable speed according to our empirical tests. Although monocular depth estimation cannot establish an absolute physical scale between the x, y and z dimensions, it nevertheless offers a convincing spatial representation of reality, enabling effective understanding of the relationships between objects in the scene. Figure 3b shows an example of depth estimation.

Given the eye position (h_x, h_y, h_z), the estimated point of visual interest (p_x, p_y, p_z) (h_z and p_z are obtained by taking the value of the depth map at the points (h_x, h_y) and (p_x, p_y) respectively, inspired by (Fang et al., 2021) and (Horanyi et al., 2023) and (g_x, g_y, g_z) the gaze direction obtained by subtracting



(a) Original image (b) Depth map

(c) 2D FOV (d) FOV map

Figure 3: Depth and field of view maps of the original image. In b) depth is encoded in gray levels, lighter pixels corresponding to smaller depths. In c) and d) regions in blue are closer to the viewing axis, regions in white are further away, and regions in red are outside of the field of view.

h from p, we compute the 3D angular difference Θ between the gaze direction and the vector from one point to eyes position, based on:

$$\Theta^{i,j} = \arccos\left(\frac{(i-h_x, j-h_y, k-h_z).(g_x, g_y, g_z)}{||(i-h_x, j-h_y, k-h_z)||_2 . ||(g_x, g_y, g_z)||_2}\right) \quad (1)$$

where (i, j) is the coordinate of each point in Θ (the angle matrix) and k is the depth value at point (i, j). From Θ, we obtain a field of view (FOV) map with opening α:

$$FOV_\alpha(x,y) = \max(1 - \frac{\Theta^{i,j}}{\alpha}, 0) \quad (2)$$

In practice, we choose $\alpha = \frac{\pi}{6}$ radians. Indeed, to approximate human perception of this field, we have reduced the field of view to $\pm 30°$ ($\frac{\pi}{6}$ radians) around the viewing axis, the angle under which colours can be distinguished (Montelongo et al., 2021).

The FOV value ranges from 1, meaning the object is in the centre of the field of view of the character in the paintings, to 0 near $\frac{\pi}{6}$, meaning the object is not in the field of view, as show in Figure 3d.

3.1.3 Object Detection

Once we compute the field of view, we are looking for objects intersecting with it.

First, we start by detecting the objects using YOLO version 7 (Wang et al., 2023). To improve performance, we fine-tune the model with the DEArt dataset (Reshetnikov et al., 2022). This dataset is a reference to detect objects and classify poses for paintings between the XII^{th} and the $XVIII^{th}$ centuries. We then use fine-tuned object detector on the dataset for image retrieval based on objects of contemplation. The output includes bounding boxes and labels for the region of interest. In the bounding box generated by the object detector, there are often unwanted elements that are not part of the object of interest. Therefore, we use the Segment Anything Model (SAM) (Kirillov et al., 2023) that takes this bounding box as input to accurately isolate the object (by segmenting it) and exclude nearby elements that do not represent it. Each point in this region is passed to the FOV function, producing a set of values. The maximum value obtained is retained for further use.

By extension, we can rewrite the FOV function 2 as:

$$FOV_\alpha(O) = max\{FOV_\alpha(x,y) | (x,y) \in O\} \quad (3)$$

where O, being the detected object.

DEArt for object detection can detect up to 70 object classes. For each image and for each main character, several occurrences of these objects are identified in the FOV, some of which may have the same label.

3.1.4 k-NN Algorithm

Let R and S be the representations of two images. Let $R = (R^1, R^2, ..., R^{70})$ and $S = (S^1, S^2, ..., S^{70})$ with $R^i = \{r^i_1, ..., r^i_{n_i}\}$ and $S^i = \{s^i_1, ..., s^i_{m_i}\}$ the set of elements of class i in R and S. r^i_k and s^i_k are the FOV value of the k^{th} occurrence of object i in images R and S respectively. In addition, R^i and S^i are sorted by decreasing FOV value ($r^i_1 \geqslant r^i_2 \geqslant \ldots \geqslant r^i_{n_i}$). n_i and m_i represent the number of occurrences of object i in the two representations. By convention, we can define that $r^i_k = 0$ if $k > n_i$.

The measure we use to compare the images is the Weighted Jaccard measure.

Given two non-negative 70−dimensional real vectors R and S, their Jaccard similarity is defined as follows:

Intersection:

$$\text{Intersection}(R,S) = \sum_{i=1}^{70} \sum_{j=1}^{\min(n_i,m_i)} \min(r^i_j, s^i_j) \quad (4)$$

Union:

$$\text{Union}(R,S) = \sum_{i=1}^{70} \sum_{j=1}^{\max(n_i,m_i)} \max(r^i_j, s^i_j) \quad (5)$$

Weighted Jaccard Measure:

$$J(R,S) = \begin{cases} \frac{\text{Intersection}(R,S)}{\text{Union}(R,S)} & ,\text{if Union}(R,S) > 0 \\ 0 & ,\text{if Union}(R,S) = 0 \end{cases}$$

and the Jaccard distance is defined as $D(R,S) = 1 - J(R,S)$.

This measure has the advantage of considering the presence or absence of objects. It is ideal for data based on sets where the presence or absence of elements is important. It also yields similar results when two characters look at the same objects in a similar manner (with comparable proximity to the central axis of the field of view). By comparing images from this measure, we obtain a list of $k-$nearest neighbours. This list contains the closest images in terms of characters looking at the same objects of interest.

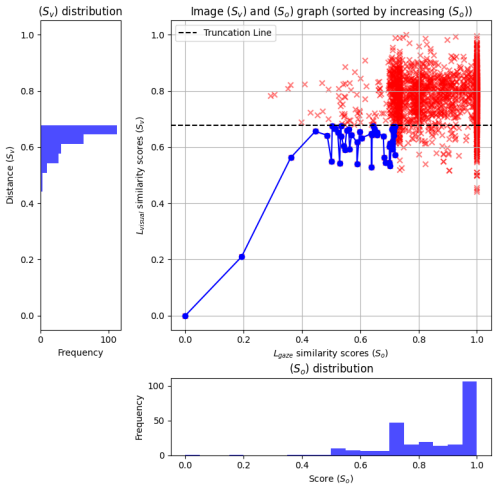
Each of these steps impact the next. A poor estimation of the visual point of interest will result in an inaccurate field of vision map, and the objects intersecting this field of vision may not actually be looked at by the character. Similarly, incorrect object detection will lead to an erroneous vector representation, and the k-nearest neighbours algorithm will associate images where the viewed objects are not visually identical.

3.2 Combination and Compromise Between Visual Similarity and Object of Contemplation Similarity

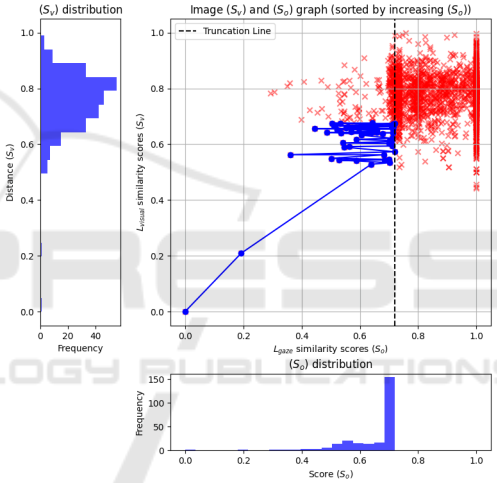
We aim to leverage two complementary approaches: one based on gaze analysis and the other on visual analysis. The goal is to incorporate gaze information while preserving the visual information. An initial approach is to sort the list according to one criterion, to truncate it with a length of $t = 10\%$ of the size of the dataset, and to reorder the remaining images with respect to other. t is a parameter chosen in relation to the length of the dataset or the length of the list of images having at least one viewed object in common with the query image.

3.2.1 Truncated-Reordering

In this approach, we want to highlight images where the objects viewed are the same and where the image visual similarity is strong, and vice versa. We have a first list based on the similarity of objects of contemplation, L_{gaze}. L_{gaze} contains images where the main character looks at the same objects but these images are not necessarily visually similar. Then we have a second list based on visual similarity, L_{visual}. L_{visual} contains images visually similar but do not necessarily contain the same objects. This list is obtained by



(a) Research based on visual similarity, truncated, then reordered according to gaze analysis



(b) Research based on gaze analysis, truncated, then reordered from visual research

Figure 4: Here, we truncate the lists at $t = 250$. Figure (a) illustrates the reordering of the visual search, and Figure (b) illustrates the reordering based on gaze analysis.

calculating the angular distances between the feature vector of the query image and the feature vectors of the images in the dataset. Feature vectors obtained from feature extraction via CLIP.

To obtain a third list based on contemplation object similarity reordered by visual similarity, $L_{gaze_t-visual}$ (with $t = 250$), we truncate the first list to the number of images that have at least one contemplation object in common with the query image. We then perform truncated-reordering based on the visual similarity scores of the images in this list.

Similarly, to obtain a fourth list L_{visual_t-gaze} based on visual similarity reordered by contemplative object similarity, we truncate the visual list and then re-

order from the contemplative object similarity scores. These processes are illustrated in Figure 4, the same data are shown here in a different order.

3.2.2 Compromise

The second approach is to propose a compromise between the two criteria, by weighting them with a parameter λ.

If s_v corresponds to the visual similarity score and s_o to the objects of contemplation similarity score, then

$$C_\lambda = s_v^\lambda . s_o^{(1-\lambda)} \quad (6)$$

with λ a weight between 0 and 1. If λ is close to 0, then more weight is given to objects of contemplation similarity, and if it is close to 1, then more weight is given to visual similarity.

4 EXPERIMENTS

In this section, we present our experiments and results in the search for paintings based on visual field analysis and objects of contemplation.

4.1 Datasets

We mainly use two datasets, one for fine-tuned YOLO v7 object detection in the fine arts domain (DEArt) and another (ArtDL) for content-based image retrieval.

DEArt (Reshetnikov et al., 2022): contains over 15,000 images with approximately 80% non-iconic paintings. The dataset also has manually defined bounding boxes identifying all instances across 70 classes as well as 12 possible poses for objects labelled as human. Of these classes, more than 50 are specific to cultural heritage and therefore do not appear in other datasets; they reflect imaginary beings, symbolic entities and other art-related categories.

ArtDL (Milani and Fraternali, 2021): contains 42,479 images of artworks portraying Christian saints, divided in 10 classes. All images are associated with high-level annotations specifying which iconography classes they belong to (from a minimum of 1 class to a maximum of 7 classes). We choose this dataset because of the strong presence of symbolic objects in religious paintings.

4.2 Gaze Direction and Object Detection Evaluation

To mitigate the lack of quantitative data for our whole method, we evaluate the performance of each separate part: the model for gaze estimation and the model for object detection.

For gaze estimation, we train the architecture of (Gupta et al., 2022) on *Gaze Follow* (photographic image) for the image modality, using the weights of the pre-trained weights for the human-centric module.

The commonly used metrics for evaluating gaze target prediction are AUC (Area Under Curve), L_2 distance, and average precision (AP). The AUC is obtained by comparing the predicted gaze target heatmap to a binarised version of the reference heatmap. This comparison allows for the plotting of a curve representing the true positive rate versus the false positive rate, with the AUC being the area under this curve: a value of 1 indicates perfect performance, and a value of 0.5 indicates random behaviour. The L_2 distance is calculated between focal points of both images. Assuming each image is of size 1×1, distance values range between 0 and $\sqrt{2}$, with a lower value being preferable. When multiple annotations are available for the gaze location (as in *Gaze Follow*), the minimum and average distances are calculated to aggregate all ground truth labels. Finally, average precision (AP) is used to evaluate the classification performance of "in-frame" or "out-of-frame" predictions. AP is calculated over the entire test set, while distance and AUC are evaluated on the subset of images where the gaze target is located within the frame.

Table 1: Comparison with state-of-the-art on *Gaze Follow* dataset for image modality.

Model	AUC ↑	AvgDist ↓	MinDist ↓
(Lian et al., 2018)	0.906	0.145	0.081
(Chong et al., 2020)	0.921	0.137	0.077
(Jin et al., 2021)	0.919	0.126	0.076
(Fang et al., 2021)	0.922	0.124	0.067
Gupta original	**0.933**	**0.134**	**0.071**
Gupta re-trained	**0.9326**	**0.123**	**0.064**

We obtain an AUC precision of 0.933, an average distance of 0.123 and an average minimum distance of 0.0637. These are state-of-the-art results, as shown in Table 1.

Additionally, we fine-tune YOLO v7 with the Dataset of European Art (DEArt). We split the dataset into 70% training, 15% validation, and 15% test sets (training set: 10,500, validation/test sets: 2,250).

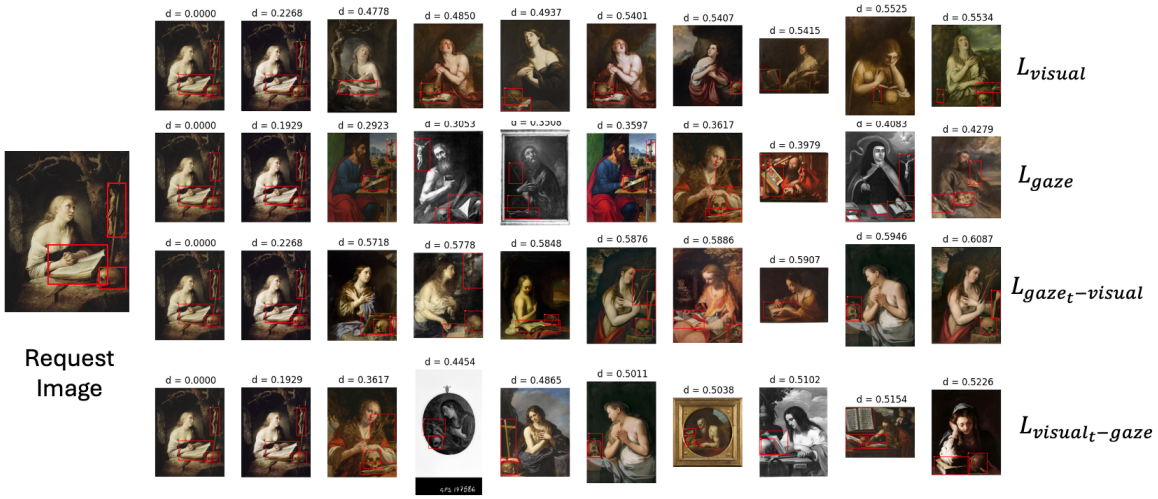


Figure 5: The 10 nearest neighbours obtained for the following 4 criteria with the bounding boxes of objects detected in images: visual with CLIP and Angular distance (L_{visual}), gaze (L_{gaze}), gaze truncated then reordered from visual ($L_{gaze_t-visual}$) and visual truncated then reordered from gaze (L_{visual_t-gaze}) (with $t = 250$). In the request image, the main character has in her field of vision the following objects: a crucifix (0.99), a book (0.67) and a skull (0.53).

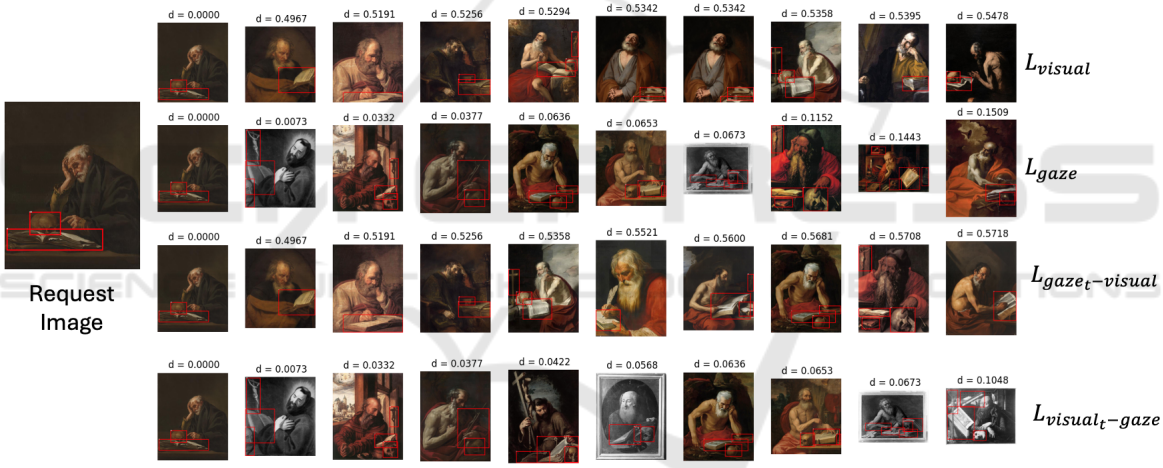


Figure 6: The 10 nearest neighbours obtained for the following 4 criteria with the bounding boxes of objects detected in images: visual with CLIP and Angular distance (L_{visual}), gaze (L_{gaze}), gaze truncated then reordered from visual ($L_{gaze_t-visual}$) and visual truncated then reordered from gaze (L_{visual_t-gaze}) (with $t = 250$). In the request image, the main character has in his field of vision the following objects: a skull on a book.

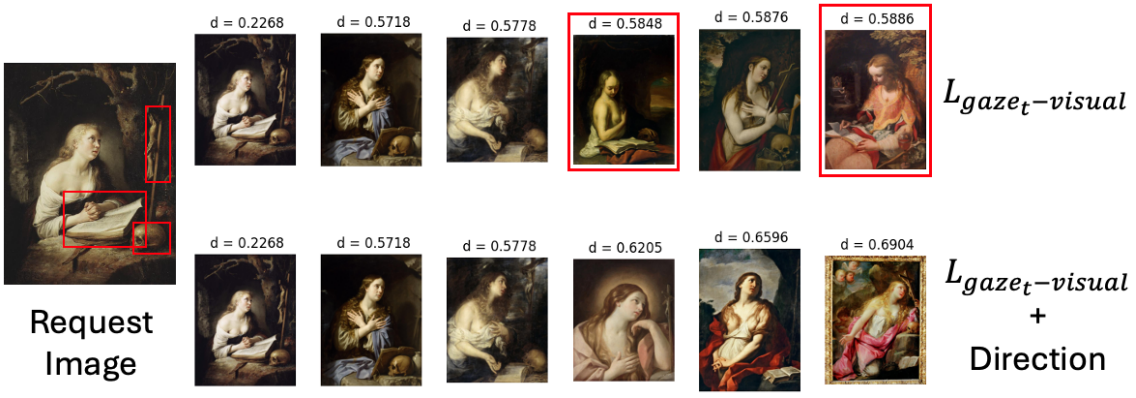
We obtain a mean average precision of 0.523 calculated at an intersection over union (IoU) threshold of 0.5 (mAP@.5) on the test, YOLO v7 performs better than fine-tuned Faster R-CNN used by authors, with a mAP@.5 of 0.312.

4.3 Retrieval Task Evaluation

In our experiments, we only select a subset of 2,310 artworks with a single character to facilitate the qualitative assessment. First, we evaluate painting similarities for each of the gaze and visual features separately. The main idea here is to compare a visual similarity search with a similarity search based on the objects looked at (gaze analysis).

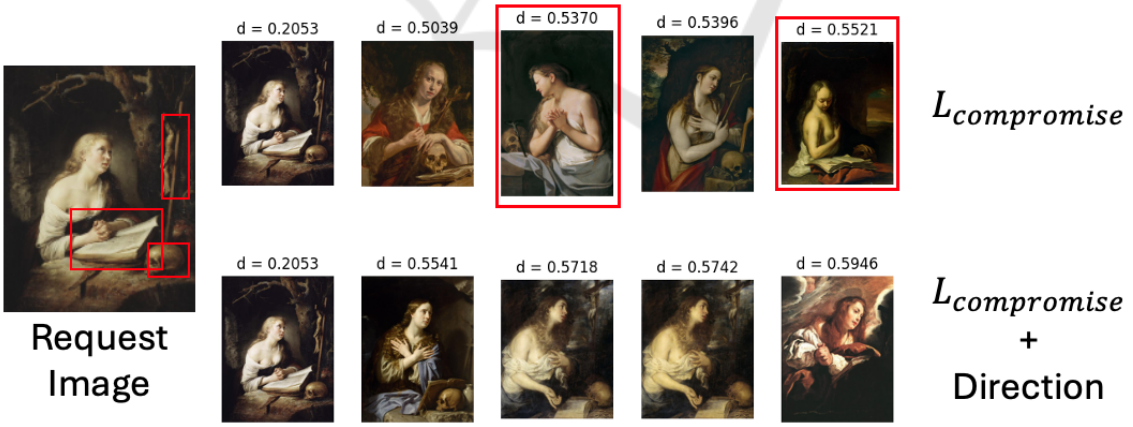
For the gaze analysis, the characters in the images proposed by the tool are not necessarily of the same gender, and their positions may differ, but they all contemplate the same objects. There is a co-occurrence of the objects being looked at, as shown by the second list in Figure 5.

For visual search only, the characters have rather evanescent looks, often gazing out of frame and all of them are the same gender figures in similar positions and surroundings (Figure 5: first list and Figure 6: first list).



(a) Gaze reorganized from visual ($L_{gaze_t-visual}$), gaze reorganized from visual ($L_{gaze_t-visual}$) adding a weight for direction.

(b) Compromise between (L_{gaze}) and (L_{visual}) and Compromise between (L_{gaze}) and (L_{visual}) by adding a weight for direction for a value of $\lambda = 0.25$.



(c) Compromise between (L_{gaze}) and (L_{visual}) and Compromise between (L_{gaze}) and (L_{visual}) by adding a weight for direction for a value of $\lambda = 0.75$.

Figure 7: The nearest neighbours obtained by truncated-reordering and compromise criteria. The images framed in red are those in which the character is not looking at the same direction as in the request image.

To demonstrate that the distributions of these two lists are not identical, we use the Wilcoxon-Mann-Whitney test (Wilcoxon, 1992), which produced a p-value of 0.0165. This result is significantly lower than the conventional 0.05 threshold, allowing us to reject the null hypothesis of equal distributions at a 5% sig-

nificance level. In other words, there is a statistically significant difference between the distributions, indicating that visual similarity does not always capture gaze and objects of contemplation.

However, when visual search is used to reorder the results of a search based on objects of contemplation ($L_{gaze_t-visual}$), this approach produces results that are visually close and aligned on objects of contemplation. In this case, we observe more paintings where the character shares the same gender as the one in the query and contemplates similar objects, as shown in the third row of Figures 5 and 6.

When we do the reverse process, i.e. when we reorder the visual search from the search based on gaze analysis, we obtain an increase in the number of images with objects looked at in common but less visually close. This can be seen in the last list in Figures 5 and 6.

In Figure 5, the second image found by the tool is another version of digitized artwork (not the same luminosity or the same colours) which explains why the distance is different from zero for the four criteria.

We can refine the previous results based on the gaze analysis of character, by giving additional weight to the metric when directions in the images are the same. Indeed, the direction of gaze is an important element in the analysis of an artwork, which is why we have chosen to add a criterion that allows us to take it into account. Figure 7a shows that among the $k-$Nearest Neighbors, the images where the characters were not looking in the same direction were replaced by an image where the character was looking in the same direction.

The final criterion we propose is compromise. This criterion can be configured to favour visual similarity or similarity based on objects of contemplation. By favouring similarity based on objects of contemplation, i.e. with a λ value close to 0, we obtain a list close to (L_{gaze}). We can see this by comparing (L_{gaze}) in Figure 5 and the Figure 7b.

Finally, if we favour visual similarity, i.e. with a λ value close to 1, we obtain a list close to (L_{visual}) and more close to ($L_{gaze_t-visual}$), highlighting the visual similarity between the paintings and keeping a focus on the objects of contemplation, as shown in Figure 7c.

4.4 Qualitative Evaluation on User Preferences

Given the absence of ground truth, we decided to ask the opinion of potential users to evaluate this multi-criteria tool. We asked them to select the lists that are most relevant to them in the similarity search based on objects of contemplation from 8 lists. We surveyed a total of 38 persons. The 38 persons are mainly students (in art and IT) and teacher-researchers. They were asked the following question: "We propose a tool based on artificial intelligence that facilitates the search for paintings with similarities based on the objects observed by the characters in the artwork. The tool generates eight separate lists, each containing paintings with similar subjects and themes. We invite you to select the list or lists that you think contain the most similar artworks, particularly in terms of the objects looked at by the characters, in relation to the requested image." We gave no indication of the nature of the list.

The Figure 8 shows that the lists $L_{gaze_t-visual}$ and $L_{gaze_t-visual} + direction$ are chosen more frequently. The numbers shown in Figure 8 represent the average number of times each list was selected, calculated as a function of the total number of queries. We found that when objects are prominent and visible, users tend to choose lists based on the similarity of the objects of contemplation. On the other hand, when objects are in the background and not easily identifiable, users prefer lists based on visual similarity. Direction also plays a fundamental role in user choice: wherever this option is present, the corresponding list is preferred to one that does not offer it. We can conclude from this that the visual aspect, whether overall or linked to direction, influences users' responses.

5 DISCUSSION

Our method identifies paintings where characters are looking at the same object as the querying character, improving existing results in similarity searches on paintings. This approach highlights the importance of gaze in visual coherence. We sought opinions from experts and users to strengthen our results and obtain evaluations, which are challenging to achieve due to the lack of ground truth in this field. The number of responses we received remains low and would require broader participation to further validate our approach. In our study, we aimed for the closest possible visual similarity, which was confirmed by user choices, indicating that users place significant importance on visual aspects. Our tool also allows for an emphasis on gaze rather than visual similarity, which can help art historians discover works with high gaze similarity but low visual similarity, revealing new and unexpected connections. Additionally, if the goal is to prioritize objects that are intersected by the viewing axis, another metric could be relevant, providing additional flexibility and precision in applying our method. As

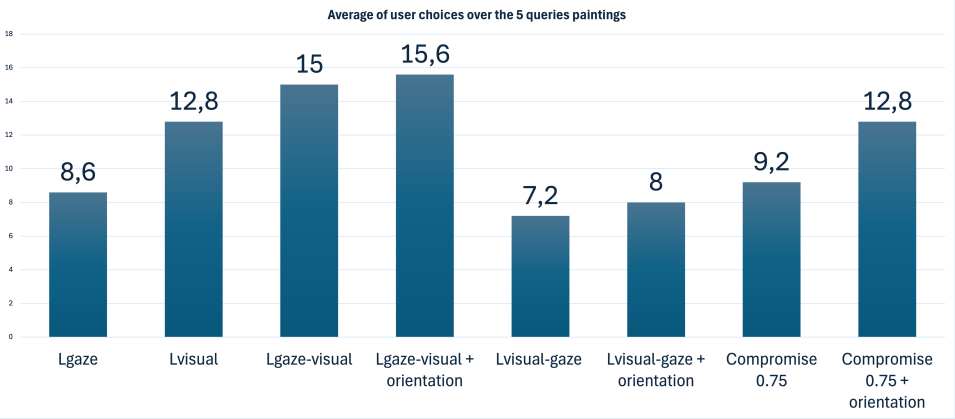


Figure 8: Average responses from 38 users on 5 query paintings, each one having 10 images. The numbers indicate the average number of selections in the list.

mentioned in section 3.1.2, our approach assumes that the depth is consistent with the 2D spatial dimension. A study based on the actual size of known objects could bring more precision to this scaling.

6 CONCLUSIONS

To conclude, the visual features of images extracted using artificial neural networks such as CLIP are not sufficient for gaze analysis. This work showed a multi-criteria tool for painting retrieval based on objects of contemplation and visual similarity. The two pipelines, gaze and visual information, complement each other and improve greatly the visual coherence between a query painting and the list of nearest neighbours. By combining gaze analysis and visual information, we were able to propose paintings similar in terms of objects of contemplation while still being visually close. We plan to extend our method to multi-character paintings and also to use this gaze analysis to estimate the composition of an artwork, either by analysing the convergence of gazes.

ACKNOWLEDGEMENTS

This work was funded by French National Research Agency with grant ANR-20-CE38-0017. We would like to thanks the PAUSE ANR-Program with grant ANR-22-PAUK-0041: Ukrainian scientists support to support the scientific stay of T. Yemelianenko in LIRIS laboratory.

REFERENCES

Aung, A. M., Ramakrishnan, A., and Whitehill, J. R. (2018). Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*.

Birkl, R., Wofk, D., and Müller, M. (2023). Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*.

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., and Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398.

Chong, E., Wang, Y., Ruiz, N., and Rehg, J. M. (2020). Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406.

Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., and Zhai, G. (2021). Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11390–11399.

Gonthier, N., Gousseau, Y., Ladjal, S., and Bonfait, O. (2018). Weakly supervised object detection in artworks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Gupta, A., Tafasca, S., and Odobez, J.-M. (2022). A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5041–5050.

Horanyi, N., Zheng, L., Chong, E., Leonardis, A., and Chang, H. J. (2023). Where are they looking in the 3d space? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In

Proceedings of the IEEE international conference on computer vision, pages 1501–1510.

Jin, T., Lin, Z., Zhu, S., Wang, W., and Hu, S. (2021). Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE.

Kadish, D., Risi, S., and Løvlie, A. S. (2021). Improving object detection in art images using only style transfer. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Li, W., Zhang, Y., Sun, Y., Wang, W., Li, M., Zhang, W., and Lin, X. (2019). Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488.

Lian, D., Yu, Z., and Gao, S. (2018). Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Masclef, T., Scuturici, M., Bertin, B., Barrellon, V., Scuturici, V.-M., and Miguet, S. (2023). A deep learning approach for painting retrieval based on genre similarity. In *International Conference on Image Analysis and Processing*, pages 270–281. Springer.

Milani, F. and Fraternali, P. (2021). A dataset and a convolutional model for iconography classification in paintings. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(4):1–18.

Montelongo, M., Gonzalez, A., Morgenstern, F., Donahue, S. P., and Groth, S. L. (2021). A virtual reality-based automated perimeter, device, and pilot study. *Translational Vision Science & Technology*, 10(3):20–20.

Qi, D., Tan, W., Yao, Q., and Liu, J. (2022). Yolo5face: Why reinventing a face detector. In *European Conference on Computer Vision*, pages 228–244. Springer.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015). Where are they looking? *Advances in neural information processing systems*, 28.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Reshetnikov, A., Marinescu, M.-C., and Lopez, J. M. (2022). Deart: Dataset of european art. In *European Conference on Computer Vision*, pages 218–233. Springer.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Smirnov, S. and Eguizabal, A. (2018). Deep learning for object detection in fine-art paintings. In *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*, pages 45–49. IEEE.

Tan, W. S., Chin, W. Y., and Lim, K. Y. (2021). Content-based image retrieval for painting style with convolutional neural network. *The Journal of The Institution of Engineers Malaysia*, 82(3).

Tian, S., Tu, H., He, L., Wu, Y. I., and Zheng, X. (2023). Freegaze: A framework for 3d gaze estimation using appearance cues from a facial video. *Sensors*, 23(23):9604.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475.

Westlake, N., Cai, H., and Hall, P. (2016). Detecting people in artwork with cnns. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14*, pages 825–841. Springer.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE.

Zhao, W., Jiang, W., Qiu, X., et al. (2022). Big transfer learning for fine art classification. *Computational Intelligence and Neuroscience*, 2022.