Combining object detection and eye tracking to identify points of interest for VR art exhibition visitors

Daria Kulyk
University of Twente
Enschede, Netherlands
d.kulyk@student.utwente.nl

ABSTRACT

Understanding visitors' preferences for artistic content can help enhance their engagement and enjoyment of the paintings. To identify points of interest, we propose to combine an object detection algorithm in artworks with eye-tracking data from users participating in a virtual reality (VR) art exhibition. In the first phase, we fine-tune the object detection model on the two manually collected datasets to locate objects within the VR exhibition paintings. In the second phase, we correlate gaze data with object data for statistical analysis to make inferences about users' regions of interest. Our findings indicate that participants spent more time looking at the meaningful objects of the paintings. Several of these object categories, including Human head, Human hair, Human mouth, Human Eye, and Person achieve precision scores above 50% after fine-tuning the object detection model. This shows that using a computer vision task to identify the areas where participants fixate their gaze holds some promise for gaining insights into user preferences for artistic content.

KEYWORDS

object detection, eye tracking, virtual reality, gaze analysis

1 INTRODUCTION

In recent years, Virtual Reality (VR) has been studied by a large number of researchers and has evolved into an immensely helpful tool in various contexts and industries [1, 40]. In art museums, the use of VR mainly converges to the reconstruction of environments that are no longer accessible, such as an artist's workspace [49] or the creation of new art forms; communicating existing paintings as part of the experience is not widely spread [5]. On that account, the opportunities that can be obtained by exploring current museum paintings in VR have not been completely realized.

One promising direction for museums is to use immersive yet controlled VR environment [43, 46] to identify users' points of interest in exhibition paintings. This allows museums to gain insights into the public's content preferences, which can be integrated into some form of useful knowledge such as knowledge graphs[1]. The stored information can then be utilized to adapt the information provided to exhibition visitors about the artworks. As Swami [50] points out, "more elaborate, relevant, and content-specific information about artworks has the greatest impact on understanding, which in turn affects aesthetic appreciation". Leder et al. [30] likewise emphasize the role of accompanying information in shaping how paintings are perceived. Moreover, given the recent transition from the idea of a collection-oriented museum to that of a visitor-oriented one [42], tailoring painting descriptions to match user preferences could be one way to increase their engagement and satisfaction.

One way to gain insights into users' points of interest is through the analysis of eye tracking data. In this work, we leverage eye tracking data collected during a user study at a VR art exhibition. Eye tracking allows for the collection of objective data [38] since it provides a direct measure of where individuals direct their gaze. Moreover, it has been found to be a good indicator of user preferences in museums by several academics [6, 38]. Notably, participants of the VR study exhibited a strong interest in the objects depicted in the paintings in their post-interviews. Gaze analysis enables us to identify the specific regions within the artwork that drew the most attention. In fact, studies by [48] and [2] have highlighted that art museum visitors often seek information about the content of the paintings.

Yet, eye tracking data on its own does not provide the semantic meaning to the areas of gaze. It can indicate where someone is looking, but it does not contain details about the particular objects that can be found in the paintings. Without this information, it is difficult to infer what type of content users prefer, such as paintings of buildings, people, or animals. It also makes it challenging to discern whether participants' fixations primarily focus on background elements or meaningful areas of the painting. By meaningful areas of the painting, we mean areas containing objects. If there is evidence that participants' gaze was longer on these regions, we can infer that the participants have a genuine interest in the painting.

One possible solution to bring the semantic meaning to the areas of gaze is a manual annotation of objects by human observers. However, it becomes impractical for larger datasets due to the time-consuming nature of the task. For this reason, we propose to combine gaze data with an object detection algorithm to add more

TScIT 39, July 07, 2023, Enschede, the Netherlands

[1] "Knowledge Graphs are very large semantic nets that integrate various and heterogeneous information sources to represent knowledge about certain domains of discourse" [17].

meaning to users' regions of interests. Specifically, our research aims to achieve two objectives: (1) Assess the feasibility of using object detection algorithm to detect areas that gain user attention, and (2) Investigate user content preferences by analyzing gaze data in combination with object data. We plan to achieve our goals by answering the following research questions (RQ):

- **RQ 1:** What is the performance of a state-of-the-art object detection model on a dataset of VR exhibition paintings?
- **RQ 2:** What is the relationship between gaze patterns and meaningful areas of the painting in terms of gaze duration?

In the first phase, we fine-tune the object detection model to identify and locate objects within the VR exhibition paintings. The model is fine-tuned on the two datasets for the specific task of detecting objects in historical artwork. We then perform inference on the dataset of the VR user study paintings to evaluate the *predictions* made by the fine-tuned object detection model. In the second phase, we utilize participants' eye-tracking information collected during the VR user study. Two statistical tests are conducted: (a) to determine whether participants spent more time looking at the meaningful areas of the paintings, and (b) to identify whether the type of object depicted in the painting influenced the duration of their gaze, and which object gained the most attention. To perform the analysis, we manually annotate the paintings from the VR art exhibition with bounding boxes that represent the *ground truth* for object detection. Consequently, we can examine if the eye tracking coordinates fall within the ground truth bounding boxes, revealing the regions of users' focus and the corresponding duration. By combining this information with the earlier object detection predictions, we can evaluate whether the object detection algorithm can successfully detect areas that gain user attention.

Our contributions can be summarized as follows: (1) We propose a method that correlates gaze data and object data to identify users' points of interest at a VR art exhibition, (2) We evaluate our method using data from a user study conducted at a VR art exhibition, showing that participants spent more time looking at the meaningful areas of the paintings, particularly those depicting buildings, (3) We identify categories of Human head, Human hair, Human mouth, Human eye, and Person as having the highest degree of readiness for the combined approach due to their higher object detection precision scores. However, we acknowledge that using computer vision tasks like object detection has limitations, including the lack of art-domain-specific annotated training data. These limitations need to be addressed first in order to be able to effectively correlate the object data with gaze data for preference inferences.

2 RELATED WORK

This section introduces most common object detection frameworks and their applications in the art domain, and presents several related works that use eye-tracking to understand user behavior and attention.

Object detection. Object detection is a computer vision task that has two objectives: determining whether instances of specific object categories (such as people, animals, cars) are present in an image, and localizing the positions of these objects, typically by drawing a bounding box around them [34].

There are two main groups of object detection frameworks. The first group, pioneered by the popular Faster R-CNN model, generates region proposals (i.e. candidate bounding boxes that potentially contain objects), which are then classified into different categories [8]. These models are able to accurately detect objects of various sizes. However, they face increasing computational complexity as the number of object candidates increases [11]. The second group includes models like You Only Look Once (YOLO), which divide the input image into grid cells, followed by classifying categories and predicting bounding boxes for each grid cell. YOLO models are faster since they do not require advanced region proposal generation but may struggle with detecting smaller objects [11].

Recognizing and detecting objects in artworks has been mainly associated with the development of large-scale retrieval systems aimed at supporting historians in their analyses, such as tracing the evolution of an object's portrayal over time [8]. Some progress has been made in the "visual similarity"-based retrieval of paintings, including the retrieval of images depicting same objects or similar iconographic elements [9]. Apart from object detection, deep neural networks have also been widely used to predict artwork attributes. More recently, Castellano et al. [7] introduced an approach for improving the prediction of artwork style and genre by leveraging deep neural networks and knowledge graphs. Eyharabide et al. [16] demonstrated how integrating knowledge graph embeddings with visual image embeddings can enhance object recognition performance on cultural heritage datasets.

Eye tracking. In the VR context, a lot of related work is focused on evaluating visual attention based on eye tracking data. Mu et al. [39] investigated participants' eye gaze while interacting with abstract VR artworks, and the duration of user attention on specific artworks was derived as a result. McNamara [35] employed eye tracking to capture data on the students' visual attention while they engaged with the artworks in the augmented reality (AR). Zhou et al. Al [54] collected eye tracking data with the goal of developing a deep learning model to predict user attention in a virtual museum. A similar research was carried out by Li et al. [31], who have also evaluated other virtual environments besides the virtual museum.

In real-world settings, there has been a notable increase in research focused on predicting a user's object of interest using previously collected gaze data [3, 14, 28]. Building upon that, Cho and Kang [11] advanced the field by introducing a framework that leverages user gaze to enhance object detection performance. In addition, attempts have been made to apply object detection in a real environment to facilitate visual attention analysis. Kumari et al. [29] used object detection to assign mobile eye-tracking data to real objects during a students' lab course, enabling a better understanding of students' visual attention patterns. Rong et al. [47] proposed a method to predict objects that capture drivers' attention while driving. This is achieved by generating attention maps based on driver eye gaze and performing object detection within these areas.

Current research mainly focuses on using eye-tracking to understand user behavior and attention and even proposes several ways for gaze-object mappings for the same purpose. However, to the best of our knowledge, the combined analysis of eye tracking and object detection data has not been performed in the context of cultural heritage (CH) art domain. This presents a new opportunity for using the obtained results to populate the knowledge graphs of various art exhibitions with both detected objects and user content preferences, ultimately enabling the creation of personalized knowledge graphs.

3 VR ART EXHIBITON

In 2020, Museum Rembrandthuis presented the exhibition "HERE: Black in Rembrandt's Time". The goal was to offer a respectful and realistic depiction of black individuals living in and around Amsterdam during the 17th century. The 19 exhibition paintings portray black individuals in central roles, diverging from the prevalent stereotypical representations that emerged in later periods. To investigate user preferences for content, the exhibition was recreated in Unity VR, providing an immersive experience for 31 participants to explore the paintings[2]. Leveraging the capabilities of VR headsets, the user study incorporated eye-tracking technology. The eye-tracking data of each participant was subsequently extracted and stored in the CSV format for further analysis.

The target of our combined method is the dataset of 19 paintings (referred to as VRPaintings dataset from now on). It serves as both the test dataset for the object detection model and the source of eye gaze data obtained from participants.

4 METHODOLOGY

4.1 Object detection model

Faster R-CNN, a Convolutional Neural Networks(CNN)-based object detection framework, relies on a Region Proposal Network (RPN) for efficient region detection within images [44]. The introduction of RPNs made the model much faster than its predecessors, R-CNN and Fast R-CNN [53]. Nevertheless, it still falls short in terms of speed and computational costs compared to models like YOLO v3 and YOLO v4. The original Faster R-CNN architecture [44] adopts the VGG-16 network (16 convolutional layers) as its backbone, but can be replaced by deeper backbones, which leads to better classification results, as well as improved object detection performance [53]. For instance, ResNet, one of the most successful deeper CNNs, outperforms VGG-16, achieving higher accuracy while still maintaining lower complexity [24]. ResNet is available in different depths, including a 50-layer ResNet-50, a 101-layer ResNet-101, and a 152-layer ResNet-15.

For object detection in paintings, the use of R-CNN networks is more common than YOLO models, as evidenced by the works of [13], [26], [22], [36]. Furthermore, through comparative analysis, it has been shown that the R-CNN network, particularly Fast R-CNN, tends to obtain greater accuracy on the People-Art dataset[3] [51]. On the other hand, YOLO has achieved better performance on the Picasso dataset[4], suggesting that it may be more suitable when dealing with abstract forms of art [51]. For our research, we decided to use Faster R-CNN network, given that (1) the majority of the paintings involved in the VR study are realistic portraits of people and (2) smaller object detection may be required for more precise eye gaze correlation.

4.2 Fine-tuning

In the realm of art analysis applications, transfer learning, particularly fine-tuning, has emerged as a prevalent approach that produces state-of-the-art results for different deep learning tasks[9, 21], including object detection. The technique involves taking a pre-trained network and adapting its parameters on the new target dataset [21]. This offers several advantages. First, rather than training a CNN from scratch, which requires a significant amount of annotated data [20] and computational resources, fine-tuning takes advantage of the pre-trained model's data and its' understanding of generic visual features (for example, edges, colour blobs [8]). Second, fine-tuning can be performed on the custom dataset that is specifically tailored to the domain, such as the art domain, and task at hand. In the context of object detection in artworks, fine-tuning is often performed on the dataset of paintings or artistic images, such as in the works of [51] [52]. This is to address the cross-depiction problem [4], wherein object detection models trained on photographic images may exhibit a decline in performance when applied to painting images due to domain shift [8]. Fine-tuning on artwork datasets helps the network learn the art-specific features of the images.

We use fine-tuning in this work to fine-tune the Faster R-CNN pre-trained on MS COCO dataset [33] on our two custom-designed datasets. By using custom datasets, we have the flexibility to select categories that are specifically relevant to our research interests. The pre-trained COCO model, on the other hand, is not optimized to detect these classes. Our datasets are collected from natural images (not paintings). This is due to the fact that VRPaintings dataset mostly consists of realistic portraits of people closely resembling real-life individuals. Additionaly, we aim to detect multiple categories that are currently not sufficiently annotated in existing painting datasets.

4.3 Datasets

For the purpose of fine-tuning the Faster R-CNN network, two datasets are created: one of natural images, the OI dataset, and one containing art-stylized versions of the same images, the StyleOI dataset. The StyleOI dataset aims to limit the effect of the aforementioned cross-depiction problem. In the Results section, we present a comparative analysis of the network's performance on both datasets (subsection 5.2), enabling us to assess which dataset is better suited for our purposes. Table 1 shows the statistics for the final datasets, which are split into training (85%) and validation (15%) sets. Examples of images in the respective datasets can be seen in Figure 1. Each dataset includes a total of 11260 images, 9571 in the training set and 1689 in the validation set. No testing set is created, as the network is intended to be tested on the VRPaintings dataset.

[2] For more details of the study, see [25].

[3] People-Art dataset is an artistic image dataset that contains bounding boxes for the single person class [51].

[4] Picasso dataset is a dataset of 218 Picasso paintings that contains bounding boxes for the single person class [19].

Table 1: The number of instances per class in the OI and StyleOI datasets

Split	Animal	Building	Dress	Hat	Eye	Hair	Hand	Head	Mouth	Person	Tree
Train	248	2258	862	171	4684	4175	4104	5868	2373	11516	1496
Val	46	355	174	29	766	660	731	1039	379	1954	230
TrainVal	294	2613	1036	200	5450	4835	4835	6907	2752	13470	1726

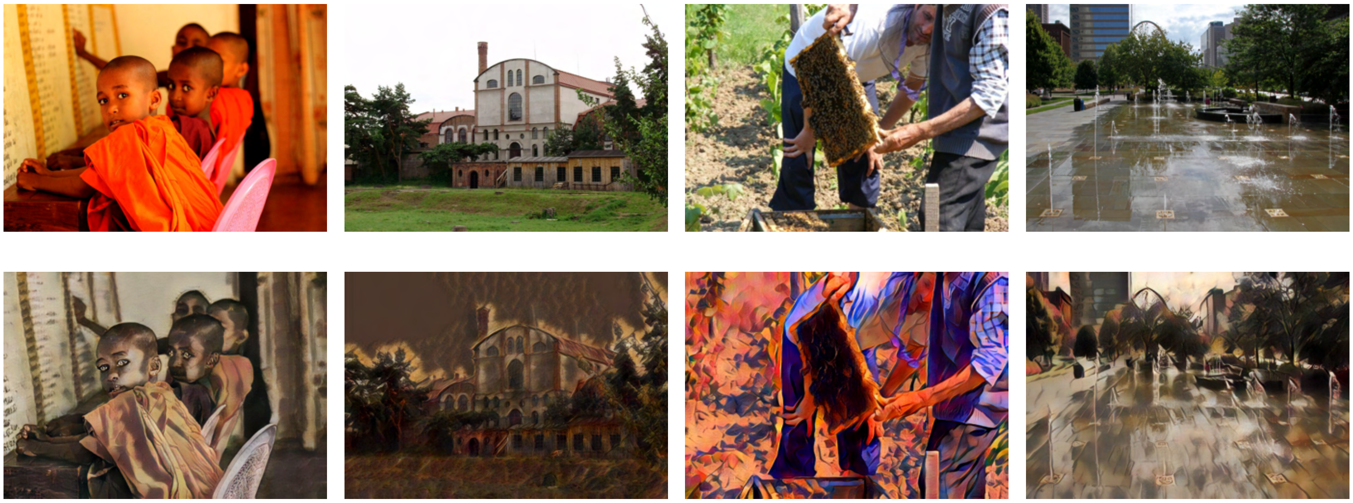


Figure 1: Examples of images from the OI dataset (above) and the StyleOI dataset (below)

OI. The OI dataset is a subset of Open Images V7 dataset[5], which is the largest existing dataset with bounding box annotations. Only 11 classes are used in the OI dataset due to the exhaustive search nature of Faster R-CNN and the associated computational challenges that arise when the number of classes increases [15]. For our class selection criteria, we consider two factors: a sufficient number of bounding box annotated examples in the original Open Images V7 dataset (at least 200), and the category's significance in artworks of the same genre/century and in relation to future gaze analysis. To determine the relevant categories based on genre/century, we utilize the SemArt dataset [18], which contains artistic descriptions for European fine-art paintings from the 8th to the 19th century. We filter paintings from the 17th century that belong to one of the following genres: portrait, genre, landscape. For each genre, we retrieve the 15 most frequently used nouns in descriptions, excluding stop words. After verifying their presence in the OpenImages v7 dataset, we are left with 16 categories for consideration.

Furthermore, in addition to considering classes from SemArt analysis, we decide to incorporate our own ones that can provide interesting insights for the gaze analysis. For instance, SemaArt's-retrieved objects are for the most part higher-level categories, such as Person, Human head; we include more granular ones such as Human eye. This decision arises from the observation that bounding boxes of higher-level categories often cover a substantial portion of the image, potentially limiting the precision of visual attention analysis. By considering a hierarchy of objects, we can look at the results of the attention study from multiple dimensions. For instance, the bounding box of the Head can serve as a reference frame, indicating a general class of interest, while the bounding boxes of the Eye provides more localized region for precise gaze correlation. On the other hand, the inclusion of smaller object classes can affect the accuracy of the object detection model. Hence, we can further investigate the trade-off between object detection performance and the benefits yielded in gaze analysis.

In the end, we allocate 7 spots for categories obtained through SemArt analysis and 4 spots for our own categories. The rationale for our selection is summarized in Table 4 and Table 5.

In the original Open Images V7 dataset, some categories (Animal, Hat) have fewer annotated examples than others. For these categories, we take all the instances present in the original dataset. However, even after doing so, the class imbalance remains, which can cause variations in object detection performance across different categories. In fact, limited data availability is one of the main challenges of the object detection task. The Person category is deliberately oversampled by us. This is because if the model can successfully identify a person in the painting, it can learn to correlate the person's presence with specific body parts such as hands. The model can utilize this knowledge the next time it encounters an image with an identified person to infer the likelihood of certain body parts being present.

StyleOI. The StyleOI dataset consists of the same number of instances and classes as the OI dataset. The key distinction is that

[5] https://storage.googleapis.com/openimages/web/factsfigures_v7.html#object-segmentations

images are stylized to look like artworks by performing AdaIn style transfer using the script by [37]. This methodology aims to address the cross-depiction problem [4] related to the domain shift from natural scenes to art scenes. The stylized images can bridge this gap by preserving the semantic information of the objects, but also capturing the unique visual features inherent in artworks, such as texture and colour[6]. Building upon work by y Kadish et al. [26], who reported significant improvements in detection performance on the People-Art dataset when fine-tuning the Faster R-CNN network on the stylized dataset, our study extends the investigation beyond the single person class. We test whether stylization process can yield better results across multiple object categories. To generate the StyleOI dataset, we use VRPainings dataset as the source of stylized images, while the OI dataset serves as the source of input images. Each input image undergoes a single style modification with a stylization weight of 1, while preserving the image's original size and crop configuration.

To facilitate training and future evaluation, OI and StyleOI datasets are exported from the Open Images annotation format to the widely adopted COCO (Common Objects in Context) [33] format. COCO stores image annotations in JSON format. We also follow a directory structure typical to a COCO dataset, i.e. we have a top-level dataset directory which contains (1) our raw image data and (2) a json file with the relevant image annotations, including bounding box and category information.

4.4 Gaze-object correlation

4.4.1 Gaze data processing. To obtain the gaze data for further gaze-object analysis, we process the data from the 31 participant CSV files (one file per participant). Each painting is represented as a grid of 100 by 100 cells in the CSV file, where each cell represents the duration of gaze at a specific coordinate within the painting (i.e. the coordinate belongs to that cell). The rows and columns in the CSV file are used to calculate the corresponding x and y coordinates for each gaze point - these are the centers of the gaze points. We filter out coordinates with gaze durations less than 0 as we want to focus our analysis on areas of user interest rather than areas that were ignored. The resulting gaze center coordinates $x_1y_1, x_2y_2, ...x_ny_n$, along with their corresponding durations $d_1, d_2, ...d_n$ are stored in gaze arrays $gazeData_{ui,pi} = \{x_1y_1d_1, x_2y_2d_2, ...x_ny_nd_n\}$, where u represents the participant/user and p represents the painting. This enables us to retrieve the data at later stages.

4.4.2 Manual object annotation. Object data is prepared by manually annotating the paintings from the VR art exhibition with bounding boxes that represent the *ground truth* for object detection. CVAT annotation tool[7] is used for this purpose and we export the annotated data to the previously mentioned COCO format. The categories are the same as in the two datasets created to train the object detection model. This is to ensure that we can later evaluate whether the object detection model *predictions* can be used as a substitute for manual annotation of the dataset. We make a distinction between the Person and Human head categories in the following way: if only a bust (head and shoulders) is visible, the painting is only annotated for Human head; if additional body parts are visible, the painting is annotated for both the Person and Human head categories. This distinction enables further investigation into whether viewers are more drawn to paintings that focus on facial features or expressions, or if they prefer paintings in which the human presence is intended to tell a more general story - for example, through specific items of clothing or what the person is holding in their hands.

4.4.3 Visual representation. In the next step, we provide a visual representation of our findings by plotting the bounding boxes and gaze data together over the artwork. Gaze data is represented as heatmap overlays over the paintings; such heatmaps are widely recognized as an effective means of visualizing human attention [39]. To generate the heatmaps, we plot the circles around gaze point centers using the Mathplotlib Python library. The radius of the circle is determined by *radius = min(unitWidth, unitHeight)/2* where *unitWidth* and *unitHeight* are obtained by dividing the painting's width and painting's height by 100, respectively (since there are 100×100 grid cells). This ensures that one gaze point corresponds to one painting unit.

4.4.4 Statistical analysis preparation. Once the gaze data and object data are created, they are related for further statistical analysis. Specifically, for each gaze point, we check whether it falls within any of the object bounding boxes (including their borders) for the painting. In our initial runs, we compared the boundaries of the circle (gaze point center ± radius) with the boundaries of the bounding box, which excluded gaze points whose circles were only partly within the bounding boxes. While this approach may be suitable for larger bounding boxes, where the object within them is typically not located right at the margin, it may not provide the most accurate results for smaller bounding boxes. In the latter scenario where the object is bounded with minimal padding, each gaze point is of high significance, and omitting even one of them can lead to different interpretations in the degrees of user interest. To address this, in our final calculations used for the results section, we examine if the gaze point center (i.e. one x_iy_i coordinate) rather than the entire circle falls within the bounding box. Figure 2 illustrates when the gaze point is considered as being within the bounding box ("on-object") and when it is viewed as being outside the bounding box ("out-object").

For gaze points marked as "on-objects," we collect information on the categories of the bounding boxes that the gaze point falls within, as well as the duration of that gaze point. A single gaze point can be located within multiple bounding boxes (i.e. multiple categories); while we store all the categories associated with the gaze point, we only record its duration once. For gaze points marked as "out-objects," we only collect information on the duration of that gaze point since there is no associated bounding box. To enable statistical analysis of user interest, the gathered data is used to calculate the following values for each pair of painting p and user u:

(1) Total gaze duration on objects: The sum of durations for all gaze points marked as "on-object" within the painting. The cumulative time provides an impression of the general interest of the particular participant in the objects of the particular artwork.

[6] This helps CNN backbone network (that Faster R-CNN relies on) learn and encode the characteristic features associated with those textures and colors.

[7] https://www.cvat.ai

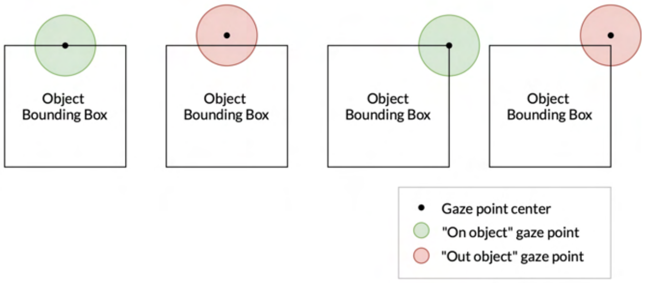


Figure 2: Schematic illustration of deciding whether a gaze point is within a bounding box.

(2) Total gaze duration outside objects: The sum of durations for all gaze points marked as "out-object" within the painting. When compared to the time spent on areas containing objects, it is possible to get a sense of the proportion of attention given to meaningful areas versus the overall painting by each participant.

(3) Average gaze duration on objects: Total gaze duration on objects divided by the number of gaze points marked as "on-objects". We include this measure because, unlike the previous ones, it is less sensitive to outliers in the gaze durations, making it more suitable for performing statistical tests on the combined participant data (discussed in subsection 4.5).

(4) Average gaze duration outside objects: Total gaze duration outside objects divided by the number of gaze points marked as "non-objects". This data is evaluated together with average gaze duration on objects to get an impression of the overall engagement of all participants.

(5) Gaze duration per object: The previous measures do not take into account differences in the amount of time spent on different object categories. Therefore, we also compute average gaze durations per specific object, allowing us to compare the relative engagement with different objects within the painting.

4.5 Gaze-object statistical analysis

4.5.1 Meaningful versus non-meaningful areas. For our statistical analysis, we investigate whether the difference in gaze durations "on-object" and "out-object" is statistically significant. Establishing this not only allows us to assess the participants' level of interest in exhibition paintings, but also helps to determine the relevance of object detection in the context of our study. Our expectation is that the participants spent more time on areas containing objects, in other words, on meaningful parts of the paintings. We decide to conduct an upper-tailed paired t-test which allows us to compare the means of the two groups. The data used for this analysis is defined as follows:

(1) Subjects: The participants involved in the study. Each participant observes the paintings on their own and is independent of others.

(2) Paired measurements: For each participant, we have paired measurements of the average gaze duration on areas with objects (group 1) and areas without objects (group 2).

In total, there are 31 pairs of observations. We conduct two experiments. In the first experiment (Experiment 1), all object categories are considered when calculating average "on-object" gaze duration. In the second experiment (Experiment 2), we exclude Person, Human head and Hair classes. The rationale behind this exclusion is based on the observation that the bounding boxes of these categories often cover the majority of the painting. In other words, we are implying that the whole artwork is meaningful, which makes the results of the Experiment 1 less powerful. By conducting Experiment 2, we can determine whether participants looked specifically at regions within the higher-level Person/Human head/Hair categories, which is more likely to reflect a true interest. Nevertheless, we still perform the Experiment 1 since not all paintings exhibit this characteristic.

4.5.2 Object-Interest analysis. For a more granular analysis, we want to know if the type of object depicted in a painting influences the average time spent on the painting. The average time spent on each painting is calculated from the provided csv files. Given that the paintings are annotated for multiple object categories, a decision has to be made on which object to select for this analysis. To determine the main element of each painting, we consider the average gaze duration per object. It is observed that classes with larger bounding boxes tend to have higher average gaze durations. As previously stated, larger bounding boxes typically correspond to higher-level categories such as "Person" and "Human head", which are the categories that effectively represent the main element of the painting. Consequently, for each artwork, we select the object that most often had the longest average gaze duration. We proceed with conducting a one-way ANOVA test, which allows us to compare the means of multiple groups with the following parameters:

(1) Dependent variable: Average time spent on the painting.

(2) Independent variable: Object category, groups are assigned to the main object in each painting to find out if there is a difference in average time spent on the painting.

5 EXPERIMENTS AND RESULTS

5.1 Training details

We use a Pytorch implementation of Faster R-CNN network with ResNet-50 as the backbone, striking a balance between depth and computational efficiency. We also choose to incorporate Feature Pyramid Network (FPN) into the Faster-R-CNN system to enhance overall accuracy and robustness to object scale variation [32]. The model is pre-trained on the MS COCO dataset [33], originally designed for 80 classes; hence, we modify the output layer to locate the 12 classes (11 classes of interest and one background class). The training batch size is set to 16 and the number of epochs is set to 2. We employ the AdamW optimizer in combination with CosineAnnealingWarmRestarts for learning rate scheduling. The training parameters are the same for both datasets.

Table 2: Average Precision scores (%) on the VRPaintings dataset per each object category

Train set	Metric	Animal	Building	Dress	Hat	Eye	Hair	Hand	Head	Mouth	Person	Tree
OI	**AP**	0	14.9	23.8	13.4	25.8	32.2	24.8	58.1	39.7	27.5	20.0
	AP^{50}	0	30.4	38.7	46.1	59.4	74.1	50.3	91.2	67.0	59.8	27.1
	AP^{75}	0	17.2	38.7	0	17.1	31.5	11.8	74.7	49.4	15.0	27.1
StyleOI	**AP**	0	0	7.4	11.5	27.8	16.0	9.3	50.0	24.7	24.4	2.0
	AP^{50}	0	0	10.6	24.4	58.3	60.0	28.7	85.6	50.0	42.3	4.9
	AP^{75}	0	0	10.6	0	12.5	7.7	6.7	70.8	21.9	23.8	0

5.2 Object Detection Results

5.2.1 Quantitative results. The model performance on the VRPaintings dataset is evaluated according to the standard COCO detection evaluation metrics[8], namely average precision metrics over different Intersection over Union (IoU) thresholds: AP (IoU thresholds from 0.5 to 0.95), AP^{50} (0.5 threshold), and AP^{75} (0.75 threshold). Overall results are summarized in Table 3 and results for each class are shown in Table 2. A few examples of object model predictions and ground truth bounding boxes can found in Figure 6.

Table 3: Average Precision scores (%) on the VRPaintings dataset for all categories

Train set	AP	AP^{50}	AP^{75}
OI	25.5	50.0	25.7
StyleOI	15.7	33.1	14

Considering the fact that the model was faced with the task of detecting a number of classes, including small objects, such as the eye and mouth, we find the model performance to be satisfactory. Notably, categories like Human hair, Human eye, and Human mouth, which were not included in the pre-trained COCO model, achieved some of the highest scores. This serves as evidence of the model's transfer learning capabilities and the applicability of fine-tuning for human-related classes.

On the other hand, the model's performance on non-human related classes is much less accurate. The Animal class, in particular, yielded no positive predictions, which was rather expected due to the limited number of training examples available for this category. In addition, ground truth animals in the VRPaintings dataset are quite small, making them difficult to detect. The Tree class performance has also likely suffered due to the small dimensions of the object. The poor performance in the remaining non-human classes, Building, Hat, Dress, could be attributed to several factors. One possible reason is the unresolved cross-depiction problem. Regardless of whether the images were stylized or not, there is a difference between modern-day items of clothing and buildings in the training dataset and the items of clothing and buildings in the historical paintings used for testing. Furthemore, there is a scarcity of training examples for garments.

Different thresholds. Average precision, the mean average of precision scores which calculate the percentage of correct positive predictions [41], can be measured at different IoU thresholds. In the object detection, the IoU measures the overlap between the predicted bounding box and ground truth bounding box [41]. When the threshold is set to 0.5, the detection is considered correct if $IoU \geq 0.5$. In our case, the impact of a higher threshold is significant overall and is quite significant for most categories. Nonetheless, the AP score for Dress and Tree (only OI dataset) remains the same for both 0.5 and 0.75 thresholds. This suggests that changing the threshold has little effect on the model's ability to accurately predict positive instances for these classes.

Different datasets. The non-stylized OI dataset achieves better results across all classes and nearly all thresholds. The only categories that demonstrate comparable accuracy are the Human Eye and Person categories. This likely indicates that the stylization process had little impact on these objects. However, for the remaining objects, the stylization process appears to have diminished their recognizability. In other words, their distinctive features present in natural images were lost, and the unique features of these categories inherent in artworks were not effectively captured. StyleOI fails to address the cross-depiction problem for our target dataset.

5.2.2 Qualitative results. For quantitative evaluation, we run inference on the VRPAintings dataset to access the predicted bounding boxes. Figure 3 shows examples of successful object predictions with a confidence score above 0.70. A high confidence score is applied to filter out predictions that are likely to be false positives. The observed results are consistent with precision scores; high confidence score bounding boxes have also high precision scores, this includes Human head, Human mouth, Human eye, and Person classes. As seen in Figure 3, the two leftmost paintings demonstrate similar detection results for both the OI dataset (above) and the StyleOI dataset (below). However, this is not the case for the two rightmost paintings, where the OI-trained dataset results in more categories being detected. Interestingly, the leftmost artworks contain larger and more prominent objects that possess a realistic appearance. This suggests that the stylization process may not be as effective for images with smaller figure ratios and/or more abstract representations. After analyzing all the images with the confidence threshold of 0.70, we notice that there are no failed predictions (false positives). We also look at class predictions with lower precision scores. This time we do not set a confidence threshold since the desired categories are likely to have low confidence scores. Futhermore, we only consider the OI dataset, as StyleOI's predictions

[8] Detailed overview of these metrics can be found at https://cocodataset.org/#detection-eval.

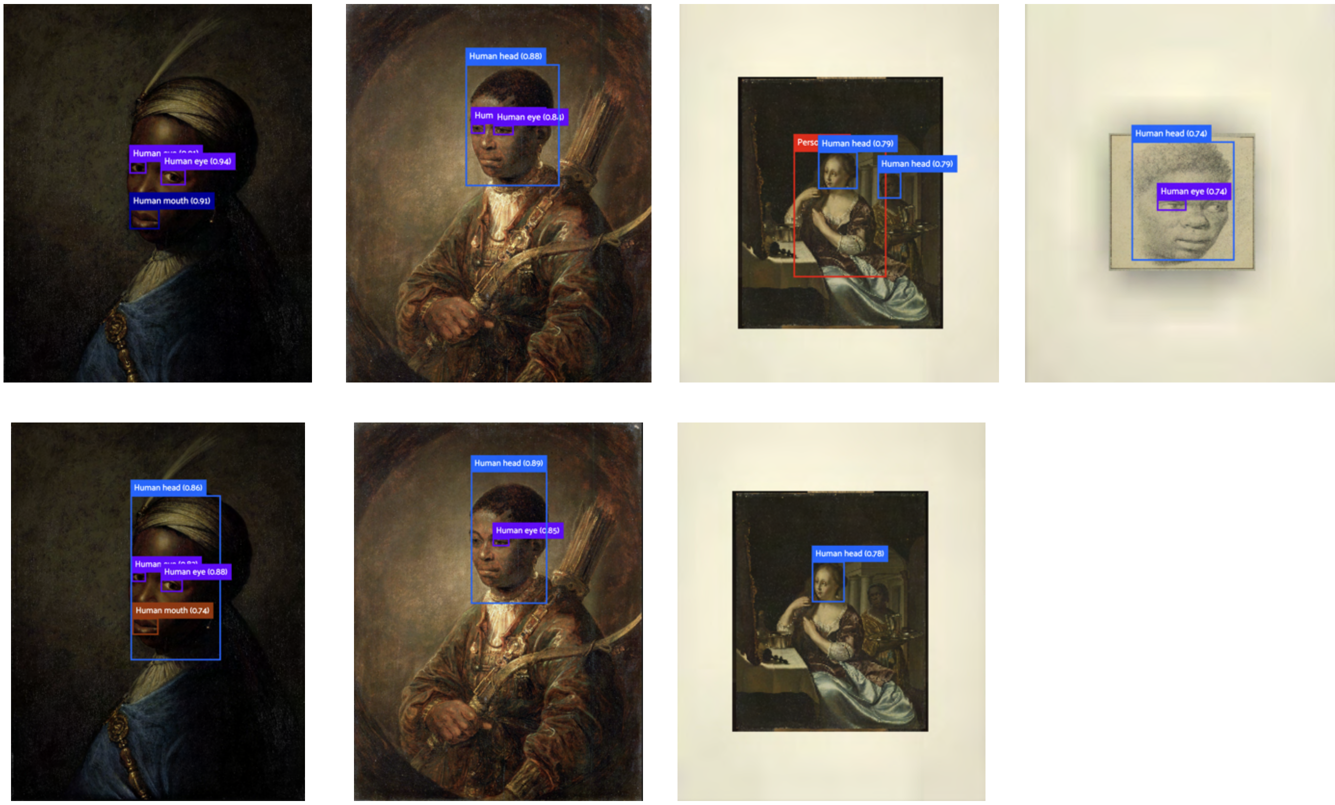


Figure 3: Examples of successful object predictions based on OI training images (above) and StyleOI training images (below). These are detections with a confidence score above 0.70. For the StyleOI dataset, only three images are shown, as the last one had lower confidence scores.

for these objects have for the most part failed. Some interesting observations are shown in Figure 4. One notable tendency is the occasional confusion between the Human hair class and the Hat class. A potential explanation could be that the model was trained predominantly on images of white people and therefore has not learned to accurately recognize the hair of individuals with black ethnicity. Nevertheless, there are a few accurate predictions of Hat class on its own. The "Dress" category has also several successful detections, although it could be argued that the detected "dresses" may not always correspond to actual dresses, depending on the definition. This highlights the importance of establishing a clear semantic understanding of object categories prior to conducting object detection tasks. Some successful recognition examples are seen in the Building category. The detection success is more varying, with paintings featuring several buildings (bottom-left painting in Figure 4) having less precise bounding boxes. The depiction of multiple architectural structures in the historical paintings in not uncommon (consider numerous depictions of market squares), which presents additional challenges for the object detection model in the art domain. It is also worth noting that in certain cases, Person instances are mistakenly identified as Building instances. This is likely due to the way the person is positioned in the painting - a tall stature might bear some resemblance to a building for the object detection model. With the Tree class, there are both failures and successes. Their relatively small size in the paintings may contribute to this. Overall, the trained model has some basic understanding of the aforementioned classes, but the low confidence (and precision) scores for these categories indicate that they may benefit from retraining.

5.3 Gaze Analysis Results

Examples of the generated eye gaze heatmaps with ground truth bounding boxes can be seen in Figure 7. Overall, the visual results reveal that there is some variation in participants' attention given to the paintings (in other words, there may be less/more gaze points, and they may be shorter/longer); however, even for participants who spent less time on the painting, the visual results seem to confirm the first hypothesis that participants spent more time on meaningful areas of the artworks containing objects. Regarding the differences in interest across different paintings and objects, it is difficult to make reasonable inferences based solely on the heatmaps.

5.3.1 Meaningful versus non-meaningful areas. Below we discuss the results of two experiments conducted to draw conclusions about the attention of users given to meaningful areas versus non-meaningful areas. The null hypothesis in our paired t-test is that the

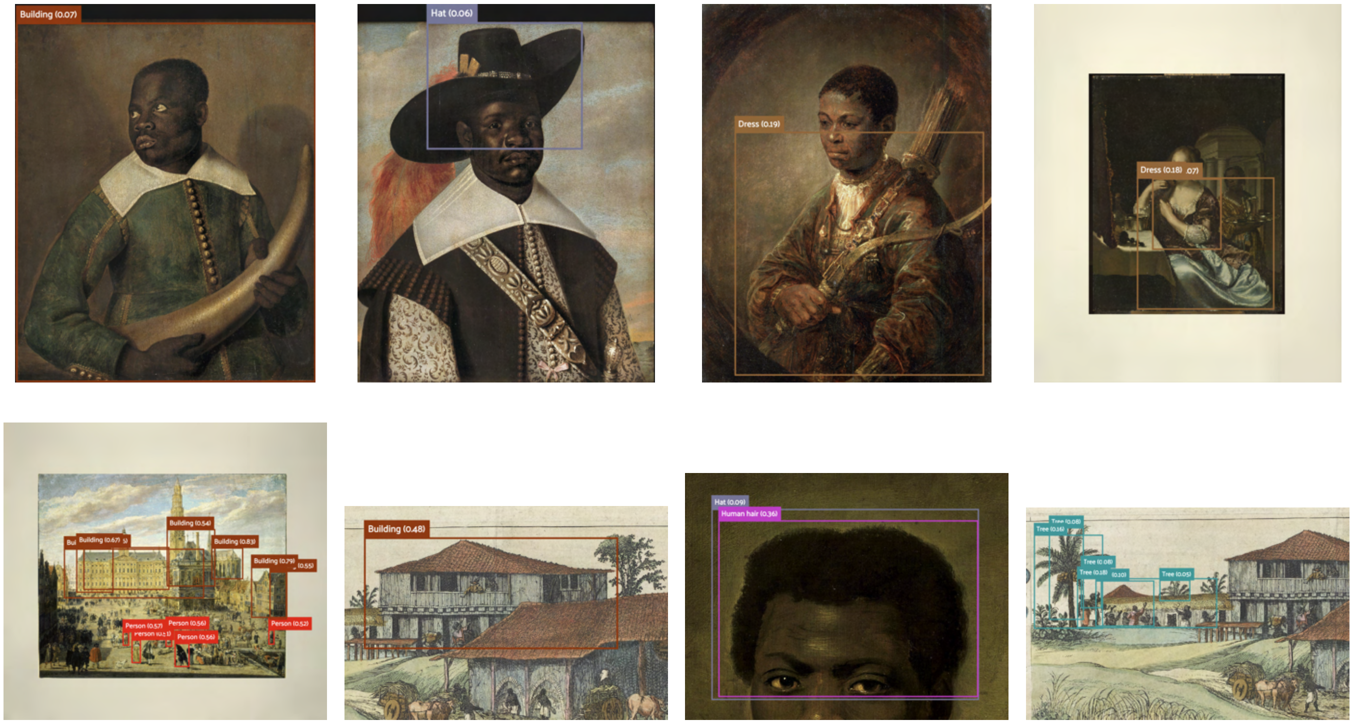


Figure 4: Examples of successful and failed object predictions based on OI training images.

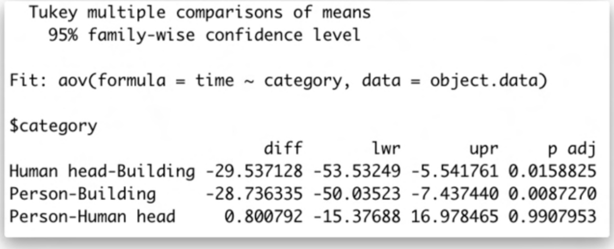
paired population means are equal, written as $H_0 : \mu_1 - \mu_2 = 0$. Our alternative hypothesis is $H_1 : \mu_1 - \mu_2 > 0$ where μ_1 is the average "on-object" gaze duration and μ_2 is the average "out-object" gaze duration.

Experiment 1. In the first experiment, we consider all object categories to be representative of the meaningful areas. To ensure the validity of the paired t-test, it is important to satisfy the normality assumption. Hence, we conduct the Shapiro-Wilk test for normality, which gives us the p-value of 0.63. Since the obtained number is greater than the significance level of 0.05, the normality assumption is supported. After performing the t-test, we find that there is a significant positive average difference between the duration of gaze on objects compared to the duration of gaze on areas without objects ($t_{31} = 6.33, p = 2.75 \times 10^{-7} < 0.05$). This suggests that, on average, participants spent more time looking at meaningful areas of the paintings. If we consider all participants across all paintings, the average gaze duration of participants in areas with objects is higher than in areas without objects in 72% of cases; in the case of total gaze duration this number is 70%.

Experiment 2. In the second experiment, we exclude Person, Human head and Hair classes. The normality assumption for the t-test is likewise satisfied (p = 0.31 > 0.05). The results of the paired t-test again indicate that the null hypothesis can be rejected, confirming that participants spent more time looking at meaningful areas of the paintings ($t_{31} = 1.75, p = 0.04 < 0.05$). Despite the exclusion of certain categories, the result is the same as in Experiment 1, suggesting that considering more granular categories does not impact the conclusion regarding attention given to meaningful areas.

It is also interesting to consider the proportion of attention given to areas with objects versus areas without objects. In Experiment 2, the average gaze duration of participants in areas with objects is higher than in areas without objects in 48% of cases; in terms of total gaze duration, this number is 11% (notice how the difference between the two values is much larger than in Experiment 1). In that sense, Experiment 2 adds valuable insights into the analysis of user interest. In particular, participants spent a significant amount of time fixating on areas without objects, resulting in a lower proportion of total gaze duration on areas containing objects. This is expected as the bounding boxes of more granular objects are smaller and may not accommodate as many gaze points, resulting in a lower cumulative duration. However, when participants did fixate on areas with objects, they tended to have longer fixation durations, leading to a higher proportion of average gaze duration on those regions. This finding suggests that participants *selectively* allocated more time to exploring meaningful areas, indicating their overall interest and engagement with VR exhibition paintings.

5.3.2 Object-Interest analysis. The most popular objects in each paining based on the average gaze duration per object can be seen in Figure 8. We exclude Hat and Human hair groups for the statistical analysis, since there is only one observation in each group. This results in three independent variable groups: Person, Human head, Building. Before proceeding with the one-way ANOVA test, we verify whether the assumptions for conducting the test are met. First, we check whether the average time spent on the painting follows a normal distribution for each group. The Shapiro-Wilk test returns a p-value of 0.37, which is greater than the significance level of 0.05. Next, we carry out Levene's test to to assess the homogeneity of variances among the groups. The p-value from Levene's test is 0.9 (> 0.05), supporting the assumption of homogeneity of variances. The null hypothesis in the one-way ANOVA test is that there is no difference among group means, $H_0 : \mu_1 = \mu_2 = \mu_3$. The alternative hypothesis, H_1, is that at least one group differs significantly from the overall mean of the dependent variable. The one-way ANOVA test shows that there is a statistically significant difference in the average time spent on the painting between the different object categories, $F(2) = 6.607, p = 0.001 < 0.05$. In other words, it is likely that object type has a significant effect on average time spent on the painting, and consequently interest in the painting. To find which specific groups differ significantly, we perform Tukey's Honestly-Significant Difference post-hoc test, results are shown in Figure 5.



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ category, data = object.data)

$category
                           diff       lwr       upr     p adj
Human head-Building -29.537128 -53.53249 -5.541761 0.0158825
Person-Building     -28.736335 -50.03523 -7.437440 0.0087270
Person-Human head     0.800792 -15.37688 16.978465 0.9907953
```

Figure 5: Results of Turkey's HSD post-hoc test

Building category has a significantly higher mean painting time than both Human head and Person categories (p-values of 0.02 and 0.01 are both lower than the significance level of 0.05). The average time spent on Person paintings is slightly higher than the time spent on Human head paintings, but the difference is not statistically significant. This points to the general interest of this particular group of 31 participants, who seem to have a higher level of engagement when presented with artworks showcasing architectural structures.

6 DISCUSSION

6.1 Object detection model

In general, the model performs well in categories related to people (Person class, body parts classes), which partly addresses our needs, given that most of the paintings are portraits. This serves as evidence of the model's transfer learning capabilities and the applicability of fine-tuning for the aforementioned classes.

For other categories, model's retraining is likely required. One of the main limitations of the algorithmic object detection is the scarcity of training examples - we are constrained to categories that have a sufficient amount of bounding box annotations. Furthermore, some categories that appear in the paintings may not have corresponding examples in the existing training datasets. In general, if we want to apply the proposed method to other historical paintings, certain objects that are specific to that time period, such as a cloak, may be consistently missed if the model is trained on modern-day natural images.

To address this limitation, one potential strategy is to use Cultural Heritage (CH)-specific training dataset consisting of *paintings* with annotated objects that are unique to historical contexts. Reshetnikov et al. [45] discuss how these types of datasets are currently lacking and introduce the largest CH dataset to date (to our knowledge) with 15,000 images for 69 classes in an attempt to address the issue. While this holds promise for the future, annotated painting datasets are still largely under development, with the expectation of adding more categories in the course of time. As advancements are made in this field, we can integrate these developments into our framework to improve both the performance and class coverage for our historical painting-object detection model. This is especially relevant given that addressing the domain gap by stylizing the dataset of natural images has not lead to any improvements in the detection of historical-looking items.

An associated problem to class scarcity is the initial imbalanced training dataset. While we could not allocate more examples to the Hat class, there was an opportunity to allocate more instances to other classes, such as the Tree class. Another limitation specific to the Tree class is the relatively small size of the objects. However, as evidenced by the work of [13], detectors trained on natural images have shown the ability to detect quite small objects with precision scores above 50%. This has been demonstrated for classes like Boat, Cow, Dog, Horse, and Sheep. This also suggests one way to enhance the performance of the Animal class by including more specific categories. Notably, in the aforementioned work, most of the classes in the training dataset were associated with around 200-600 images, which is far less than than the number of training examples available for most of our categories. However, their training dataset appears to be more balanced. As such, balancing the dataset could also involve reducing the number of training examples for seemingly over-represented categories. In our specific case, we could have considered reducing the number of instances in people-related categories to avoid over-training on these specific classes, thereby potentially improving the performance of the model for other objects.

Unlitmately, the success of the object detection model in historical paintings depends on the creation of a balanced dataset that includes categories relevant to the specific time period being analyzed. For future work, it may be necessary to create training dataset(s) from multiple existing datasets, as relying solely on one natural images dataset, such as Open Images, might not provide a sufficient number of training examples for each relevant class. By incorporating a diverse set of training data, including both natural images and paintings, we can improve the model's performance in detecting a diverse range of objects. For instance, training the model on natural images has shown promising results for realistic portrait artworks. On the other hand, painting training data can provide better detection results for non-human related categories, where the artistic depictions may differ significantly from real-life examples. Since the number of annotated painting images is still limited, combining the datasets can be seen as a short-term solution to address the problem.

6.2 Gaze-Object Analysis

6.2.1 Meaningful versus non-meaningful areas. Based on the results of gaze analysis, we can conclude that participants of the VR art exhibition dedicated more time to exploring meaningful regions of the paintings that contained objects. This observation aligns well with the objective of object detection algorithm, which aims to identify these meaningful objects rather than the insignificant background elements. Therefore, the results already suggest the potential suitability of using object detection in combination with gaze data. However, it is also important to consider the degree of success of the object detection algorithm in detecting these gaze regions. Its greatest success has been in the human-related categories, suggesting the suitability of using algorithmic object detection to assign participants' gaze data to at least those classes within the paintings. For other classes, the ways to improve algorithm's performance have been mentioned above. By improving the model's accuracy in detecting and localizing other categories, we can assign gaze data to a wider range of meaningful objects within the paintings. This, in turn, will provide additional data (e.g. bounding boxes) for the subsequent statistical gaze-object analysis.

It is also important to establish the constraints of performing the statistical gaze-object analysis for other datasets. When defining meaningful areas, it is necessary to determine which object categories should be included in the analysis. Specifically, in situations where there is a hierarchy of objects in the paintings, such as portrait paintings, it is advised to consider only granular categories. This approach helps mitigate the influence of rectangular bounding boxes that often cover entire regions of the painting for the large (dimension-wise) higher-level classes; this makes the comparison between meaningful and non-meaningful areas less powerful and unreliable. For instance, the results may indicate that participants spent more time looking at areas containing objects when, in reality, they devoted more attention to the background within the bounding box. Nevertheless, it is important to note that the presence of a hierarchical object structure may not always be true. Furthermore, knowing the full content of the test dataset in advance is highly unlikely. Therefore, one of the primary challenges in performing statistical gaze-object analysis lies in the rectangular shape of the bounding boxes.

6.2.2 Object detection considerations. To address this challenge, alternative methods for object annotation can be explored. Semantic segmentation is one such strategy, which generates pixel-level segmentation masks. As every pixel is classified in the image [27], the mask is likely to align more closely with the object's boundaries than the bounding box. On one hand, this can potentially facilitate more precise correlation between "on-object" and "out-object" gaze points. On the other hand, the presence of occlusion poses a significant challenge for object segmentation algorithms, as they struggle to accurately group the regions that have been split into one instance [10]. Consequently, we may not fully rely on segmentation results where objects within paintings overlap (such as in the case of buildings overlapping in one of our testing images). In literature, the application of semantic segmentation in the art domain is lacking. Recently, [12] proposed the first semantic segmentation solution for artistic paintings. They unveiled a new dataset called DRAM, which includes artwork from the movements of Realism, Impressionism, Post-impressionism, and Expressionism. With DRAM as the target dataset and semantic segmentation dataset containing real images as the source dataset, they used style transfer to address the domain gap. Next, they trained the segmentation network using the stylized versions of the images with their original segmentation labels, and applied domain confusion to further refine the segmentation network of each sub-domain using DRAM's original paintings. The method produced state-of-the art results on artistic paintings, however the authors note the same drawback that we previously mentioned. They had to settle for a relatively small number of classes due to differences between modern-day items in photographs and objects commonly found in historical artworks. Furthermore, as with bounding box methods, many existing semantic segmentation approaches require a large number of annotated images with pixel-wise masks. The manual annotation process for these masks is known to be time-consuming and computationally costly [23].

We think the future work can explore the impact of using segmentation masks instead of bounding boxes for combined gaze analysis. This also requires a redefinition of our original categories, since, for example, hierarchical objects can suffer from occlusion. By doing so, we can weigh the potential benefits of more precise gaze analysis against the impact on algorithmic performance.

6.2.3 Object-Interest analysis. The object-interest analysis suggests that the type of object depicted in the painting had an impact on the average time spent on the painting. Specifically, Building category generated the most interest, followed by Person and Human head. Since this type of analysis indicates the preferences of all participants, in the future, more extensive painting descriptions can be provided for the categories that received the most attention. For example, for the Building category, additional information about the architectural style, the artist behind the piece, or the artistic style being used could be included. Moreover, a more focused analysis per participant can be conducted to determine which of the three identified objects captured each individual's attention the most. This paves the way to creating personalized descriptions of paintings for each user, rather than basing them on and generalizing from the preferences of the majority.

It is important to note that the conclusions made about participants' points of interest are based on the results of a relatively small-scale user study involving 31 participants and 19 paintings. To achieve more reliable conclusions, it is necessary to conduct a larger user study with a greater number of participants and a more diverse set of paintings. Such a study would likely include a broader range of object groups and a higher number of observations per group. In our current analysis, the Building group contains only two observations (i.e. Building is the main element of only two paintings), which may not provide sufficient evidence for the generalization of user preferences.

7 CONCLUSION

In this work, we presented and evaluated an approach that can be used to correlate gaze data with object data in order to identify participants' points of interest at a VR art exhibition. We contribute by showing that using a computer vision task to identify the areas where participants fixated their gaze has some promise for gaining insights into user preferences for artistic content; however, this involves first addressing the challenges associated with the chosen computer vision task. In our case, we utilized an object detection algorithm characterized by bounding boxes, which introduced certain limitations. To overcome these limitations and improve the accuracy of the statistical gaze-object analysis, we propose exploring alternative computer vision tasks, such as semantic segmentation. By employing semantic segmentation, which generates pixel-level segmentation masks, we can potentially achieve more precise identification of regions of interest. It is worth noting that the availability of art domain-specific training data poses another challenge, which will require time to address. Once the challenges of computer vision tasks are overcome, further statistical analysis can deliver interesting results into participants' points of interest. While our analysis was limited by a small scale of the user study, the proposed methodology can be applied at the larger-scale VR art exhibitions. Ultimately, the strategy can help understand the general feedback of the exhibition, as well as align painting descriptions with gaze-derived preferences of painting objects.

REFERENCES

[1] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. 2016. State of the art of virtual reality technology. *IEEE Aerospace Conference Proceedings* 2016-June (6 2016), 1–19. https://doi.org/10.1109/AERO.2016.7500674

[2] Fabricio Barth, Heloisa Candello, Paulo Cavalin, and Claudio Pinhanez. 2020. Intentions, Meanings, and Whys: Designing Content for Voice-based Conversational Museum Guides. *ACM International Conference Proceeding Series* (7 2020), 1–8. https://doi.org/10.1145/3405755.3406128

[3] Michael Barz and Daniel Sonntag. 2016. Gaze-guided object classification using deep neural networks for attention-based computing. *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (9 2016), 253–256. https://doi.org/10.1145/2968219.2971389

[4] Hongping Cai, Qi Wu, Tadeo Corradi, and Peter Hall. 2015. The Cross-Depiction Problem: Computer Vision Algorithms for Recognising Objects in Artwork and in Photographs. *arXiv preprint arXiv:1505.00110* (5 2015). https://arxiv.org/abs/1505.00110v1

[5] Marcello Carrozzino and Massimo Bergamasco. 2010. Beyond virtual museums: Experiencing immersive virtual reality in real museums. *Journal of Cultural Heritage* 11, 4 (2010), 452–458. https://doi.org/10.1016/j.culher.2010.04.001

[6] Sylvain Castagnos, Florian Marchal, Alexandre Bertrand, Morgane Colle, and Djalila Mahmoudi. 2019. Inferring art preferences from gaze exploration in a museum. *ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (6 2019), 425–430. https://doi.org/10.1145/3314183.3323871

[7] Giovanna Castellano, Vincenzo Digeno, Giovanni Sansaro, and Gennaro Vessio. 2022. Leveraging Knowledge Graphs and Deep Learning for automatic art analysis. *Knowledge-Based Systems* 248 (7 2022), 108859. https://doi.org/10.1016/J.KNOSYS.2022.108859

[8] Giovanna Castellano and Gennaro Vessio. 2021. Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. *Neural Computing and Applications* 33, 19 (10 2021), 12263–12282. https://doi.org/10.1007/S00521-021-05893-Z

[9] Eva Cetinic and James She. 2022. Understanding and Creating Art with AI: Review and Outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2 2022), 1–22. https://doi.org/10.1145/3475799

[10] Yi Ting Chen, Xiaokai Liu, and Ming Hsuan Yang. 2015. Multi-instance object segmentation with occlusion handling. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 07-12-June-2015 (6 2015), 3470–3478. https://doi.org/10.1109/CVPR.2015.7298969

[11] Dae Yong Cho and Min Koo Kang. 2021. Human gaze-aware attentive object detection for ambient intelligence. *Engineering Applications of Artificial Intelligence* 106 (11 2021), 104471. https://doi.org/10.1016/J.ENGAPPAI.2021.104471

[12] N. Cohen, Y. Newman, and A. Shamir. 2022. Semantic Segmentation in Art Paintings. *Computer Graphics Forum* 41, 2 (5 2022), 261–275. https://doi.org/10.1111/CGF.14473

[13] Elliot J. Crowley and Andrew Zisserman. 2016. The Art of Detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9913 LNCS. Springer International Publishing, 721–737. https://doi.org/10.1007/978-3-319-46604-0{_}50

[14] Stijn De Beugher, Younes Ichiche, Geert Brône, and Toon Goedemé. 2012. Automatic analysis of eye-tracking data using object detection algorithms. In *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Association for Computing Machinery, 677–680. https://doi.org/10.1145/2370216.2370363

[15] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. 2013. Scalable Object Detection using Deep Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (12 2013), 2155–2162. https://doi.org/10.1109/CVPR.2014.276

[16] Victoria Eyharabide, Imad Eddine Ibrahim Bekkouch, and Nicolae Dragoş Constantin. 2021. Knowledge Graph Embedding-Based Domain Adaptation for Musical Instrument Recognition. *Computers 2021, Vol. 10, Page 94* 10, 8 (8 2021), 94. https://doi.org/10.3390/COMPUTERS10080094

[17] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. 2020. Introduction: What Is a Knowledge Graph? In *Knowledge Graphs*. Springer International Publishing, 1–10. https://doi.org/10.1007/978-3-030-37439-6{_}1

[18] Noa Garcia and George Vogiatzis. 2019. How to read paintings: Semantic art understanding with multi-modal retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11130 LNCS (2019), 676–691. https://doi.org/10.1007/978-3-030-11012-3{_}52/TABLES/5

[19] Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik. 2015. Detecting people in cubist art. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8925 (2015), 101–116. https://doi.org/10.1007/978-3-319-16178-5{_}7/COVER

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (6 2014), 580–587. https://doi.org/10.1109/CVPR.2014.81

[21] Nicolas Gonthier, Yann Gousseau, and Saïd Ladjal. 2021. An Analysis of the Transfer Learning of Convolutional Neural Networks for Artistic Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12663 LNCS (2021), 546–561. https://doi.org/10.1007/978-3-030-68796-0{_}39/COVER

[22] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. 2019. Weakly supervised object detection in artworks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11130 LNCS (2019), 692–709. https://doi.org/10.1007/978-3-030-11012-3{_}53/FIGURES/5

[23] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S. Lew. 2018. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval* 7, 2 (6 2018), 87–93. https://doi.org/10.1007/S13735-017-0141-Z/FIGURES/3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (12 2015), 770–778. https://doi.org/10.1109/CVPR.2016.90

[25] Delaram Javdani Rikhtehgar, Shenghui Wang, Hester Huitema hhuitema, Julia Alvares jalvares, Stefan Schlobach ksschlobach, Carolien Rieffe, Dirk Heylen, Hester Huitema, Julia Alvares, and Stefan Schlobach. 2023. Personalizing Cultural Heritage Access in a Virtual Reality Exhibition: A User Study on Viewing Behavior and Content Preferences. *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (6 2023), 379–387. https://doi.org/10.1145/3563359.3596666

[26] David Kadish, Sebastian Risi, and Anders Sundnes Lovlie. 2021. Improving Object Detection in Art Images Using only Style Transfer. *Proceedings of the International Joint Conference on Neural Networks* 2021-July (7 2021), 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534264

[27] Jaskirat Kaur and Williamjeet Singh. 2022. Tools, techniques, datasets and application areas for object detection in an image: a review. *Multimedia Tools and Applications 2022 81:27* 81, 27 (4 2022), 38297–38351. https://doi.org/10.1007/S11042-022-13153-Y

[28] Jung Hwa Kim, Seung June Choi, and Jin Woo Jeong. 2019. Watch do: A smart IoT interaction system with object detection and gaze estimation. *IEEE Transactions on Consumer Electronics* 65, 2 (5 2019), 195–204. https://doi.org/10.1109/TCE.2019.2897758

[29] Niharika Kumari, Verena Ruf, Sergey Mukhametov, Albrecht Schmidt, Jochen Kuhn, and Stefan Küchemann. 2021. Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4. *Sensors 2021, Vol. 21, Page 7668* 21, 22 (11 2021), 7668. https://doi.org/10.3390/S21227668

[30] Helmut Leder, Claus Christian Carbon, and Ai Leen Ripsas. 2006. Entitling art: Influence of title information on understanding and appreciation of paintings. *Acta Psychologica* 121, 2 (2 2006), 176–198. https://doi.org/10.1016/j.actpsy.2005.

08.005

[31] Xiangdong Li, Yifei Shan, Wenqian Chen, Yue Wu, Praben Hansen, and Simon Perrault. 2021. Predicting user visual attention in virtual reality with a deep learning model. *Virtual Reality* 25, 4 (12 2021), 1123–1136. https://doi.org/10.1007/s10055-021-00512-7

[32] Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2017-January. IEEE Computer Society, 936–944. https://doi.org/10.1109/CVPR.2017.106

[33] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS, PART 5 (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1{_}48/COVER

[34] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2019. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision 2019 128:2* 128, 2 (10 2019), 261–318. https://doi.org/10.1007/S11263-019-01247-4

[35] Ann M. McNamara. 2011. Enhancing art history education through mobile Augmented Reality. In *Proceedings of VRCAI 2011: ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications to Industry*. 507–512. https://doi.org/10.1145/2087756.2087853

[36] Alexis Mermet, Asanobu Kitamoto, Chikahiko Suzuki, and Akira Takagishi. 2020. Face Detection on Pre-modern Japanese Artworks using R-CNN and Image Patching for Semi-Automatic Annotation. *SUMAC 2020 - Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents* (10 2020), 23–31. https://doi.org/10.1145/3423323.3423412

[37] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, Wieland Brendel, Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv* (7 2019), arXiv:1907.07484. https://doi.org/10.48550/ARXIV.1907.07484

[38] PPP Morantes, SA Peñarete, G Arbelaez, M Camargo, and L Dupont. 2016. Understanding museum visitors' experience through an eye-tracking study and a living lab approach. In *International Conference on Engineering, Technology and Innovation/IEEE lnternational Technology Management Conference (ICE/ITMC)*. IEEE, 1–6. https://doi.org/10.1109/ICE/ITMC39735.2016.9025900

[39] Mu Mu, Murtada Dohan, Alison Goodyear, Gary Hill, Cleyon Johns, and Andreas Mauthe. 2022. User attention and behaviour in virtual reality art encounter. *Multimedia Tools and Applications* (7 2022), 1–30. https://doi.org/10.1007/S11042-022-13365-2/FIGURES/22

[40] Muhanna A. Muhanna. 2015. Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions. , 344–361 pages. https://doi.org/10.1016/j.jksuci.2014.03.023

[41] Rafael Padilla, Sergio L. Netto, and Eduardo A.B. Da Silva. 2020. A Survey on Performance Metrics for Object-Detection Algorithms. *International Conference on Systems, Signals, and Image Processing* 2020-July (7 2020), 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

[42] Davide Pantile, Roberto Frasca, Antonio Mazzeo, Matteo Ventrella, and Giovanni Verreschi. 2017. New Technologies and Tools for Immersive and Engaging Visitor Experiences in Museums: The Evolution of the Visit-Actor in Next-Generation Storytelling, through Augmented and Virtual Reality, and Immersive 3D Projections. In *Proceedings - 12th International Conference on Signal Image Technology and Internet-Based Systems, SITIS 2016*. Institute of Electrical and Electronics Engineers Inc., 463–467. https://doi.org/10.1109/SITIS.2016.78

[43] Thomas D. Parsons. 2015. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience* 9, DEC (12 2015), 660. https://doi.org/10.3389/fnhum.2015.00660

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (6 2015), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[45] Artem Reshetnikov, Maria-Cristina Marinescu, Joaquim More Lopez, Artem Reshetnikov, Maria-Cristina Marinescu, and Joaquim More Lopez. 2022. DEArt: Dataset of European Art. *arXiv* (11 2022), arXiv:2211.01226. https://doi.org/10.48550/ARXIV.2211.01226

[46] A Rizzo and Bouchard Stéphanee. 2019. *Virtual Reality Technologies for Health and Clinical Applications Virtual Reality for Psychological and Neurocognitive Interventions*. Springer. http://www.springer.com/series/13399

[47] Yao Rong, Naemi Rebecca Kassautzki, Wolfgang Fuhl, and Enkelejda Kasneci. 2022. Where and What: Driver Attention-Based Object Detection. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (5 2022), 1–22. https://doi.org/10.1145/3530887

[48] Stefan Schaffer, Aaaron Ruß, Mino Lee Sasse, Louise Schubotz, and Oliver Gustke. 2021. Questions and Answers: Important Steps to Let AI Chatbots Answer Questions in the Museum. In *ArtsIT, Interactivity and Game Creation: Creative Heritage. New Perspectives from Media Arts and Artificial Intelligence. 10th EAI International Conference, ArtsIT 2021, Virtual Event (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 422)*, Matthias Wölfel, Johannes Bernhardt, and Sonja Thiel (Eds.). Springer, Cham, 346–358. https://doi.org/10.1007/978-3-030-95531-1

[49] Maria Shehade and Theopisti Stylianou-Lambert. 2020. Virtual Reality in Museums: Exploring the Experiences of Museum Professionals. *Applied Sciences 2020, Vol. 10, Page 4031* 10, 11 (6 2020), 4031. https://doi.org/10.3390/APP10114031

[50] Viren Swami. 2013. Context matters: Investigating the impact of contextual information on aesthetic appreciation of paintings by Max Ernst and Pablo Picasso. *Psychology of Aesthetics, Creativity, and the Arts* 7, 3 (8 2013), 285–295. https://doi.org/10.1037/a0030965

[51] Nicholas Westlake, Hongping Cai, and Peter Hall. 2016. Detecting people in artwork with CNNs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9913 LNCS (2016), 825–841. https://doi.org/10.1007/978-3-319-46604-0{_}57/TABLES/4

[52] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (4 2017), 1211–1220. https://doi.org/10.1109/ICCV.2017.136

[53] Hao Zhang and Xianggong Hong. 2019. Recent progresses on object detection: a brief review. *Multimedia Tools and Applications* 78, 19 (10 2019), 27809–27847. https://doi.org/10.1007/S11042-019-07898-2/TABLES/3

[54] Yunzhan Zhou, Tian Feng, Shihui Shuai, Xiangdong Li, Lingyun Sun, and Henry Been Lirn Duh. 2022. EDVAM: a 3D eye-tracking dataset for visual attention modeling in a virtual museum. *Frontiers of Information Technology and Electronic Engineering* 23, 1 (1 2022), 101–112. https://doi.org/10.1631/FITEE.2000318

A APPENDIX

Table 4: Accepted categories for training datasets for object detection model. Categories marked with * are own categories. Other categories are derived based on the analysis of the SemArt dataset.

Accepted Categories	
Category	**Rationale**
Person	Captures one of the most repeated categories (Man, Woman, Boy) at once
Human hand	Second most repeated element in portrait paintings; opens up possibilities for a more comprehensive analysis of user interest due to the potential for participants' gaze to be directed towards objects held in the hand
Human head	Provides a general reference frame for more granular categories such as Human eye and Human mouth; its performance can be compared to the Person category to understand suitability for analyzing gaze patterns
Dress	Enables to assess model's ability to overcome domain shift between modern and medieval dresses
Building	Potentially represents a meaningful object in the paintings; captures a broader range of architectural elements than the House category, which can lead to better detection performance
Animal	Potentially represents a meaningful object in the paintings
Tree	Among the remaining landscape categories, has the highest number of occurrences
Human eye*, Human mouth*, Human hair*	Allows granular analysis of eye gaze patterns within portrait paintings
Hat*	Enables the analysis of participant perception of historical style attributes; more suitable for object detection model than e.g. jewelry due to the bigger size

Table 5: Rejected categories for training datasets for object detection model

Rejected Categories	
Category	**Rationale**
Man, Woman, Boy	Summarized under Person category to save class space; however, may be interesting for future analysis to explore participants' gaze patterns in relation to different sexes depicted
Human face	Similar to Human head category, no additional insights for gaze analysis
House	Similar to Building category, no additional insights for gaze analysis
Boat	Has fewer occurrences than the selected categories from landscape paintings (Tree and Building)
Table	Unlikely to represent a prominent or meaningful object in the paintings
Window	Unlikely to represent a prominent or meaningful object in the paintings
Musical instrument	Rejected due to space constraints, has potential for future analysis

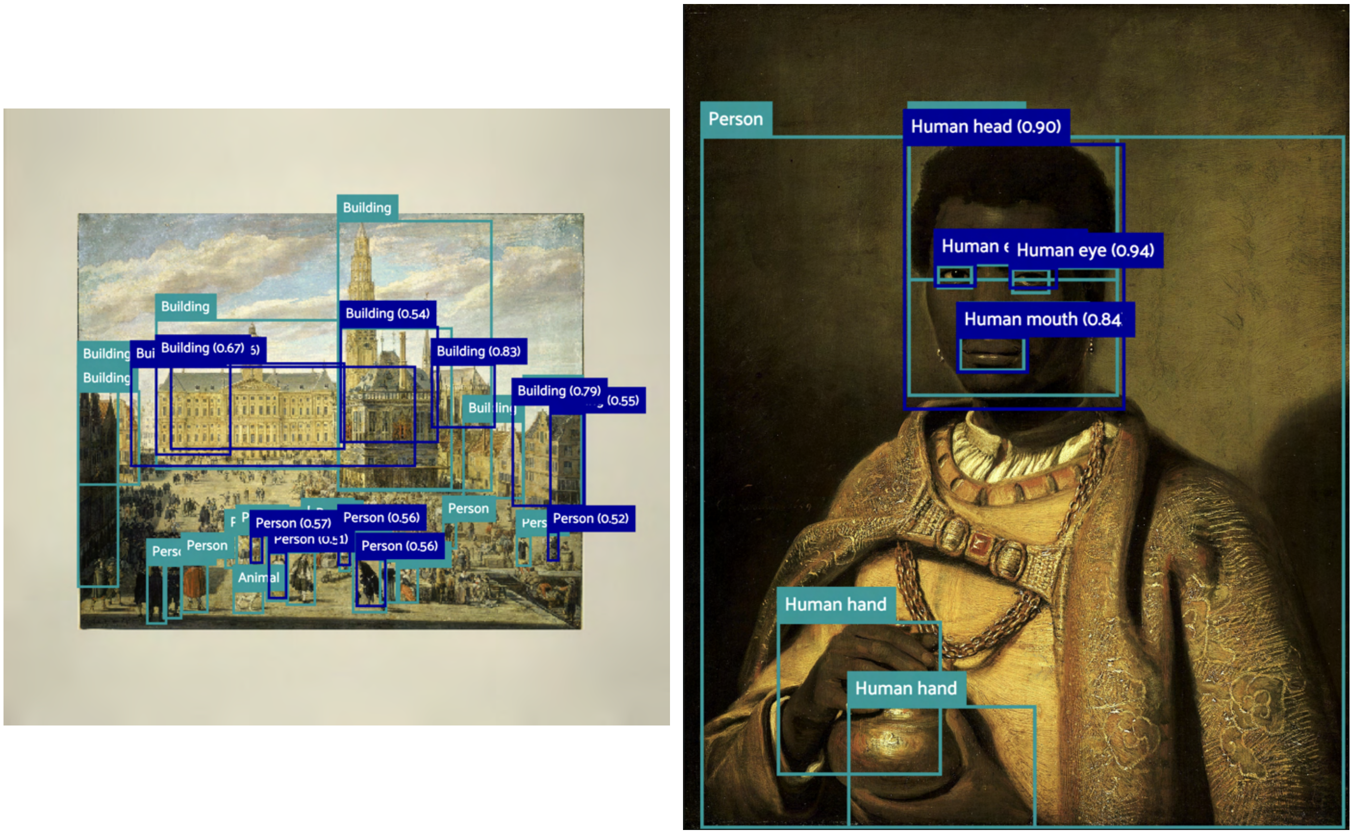
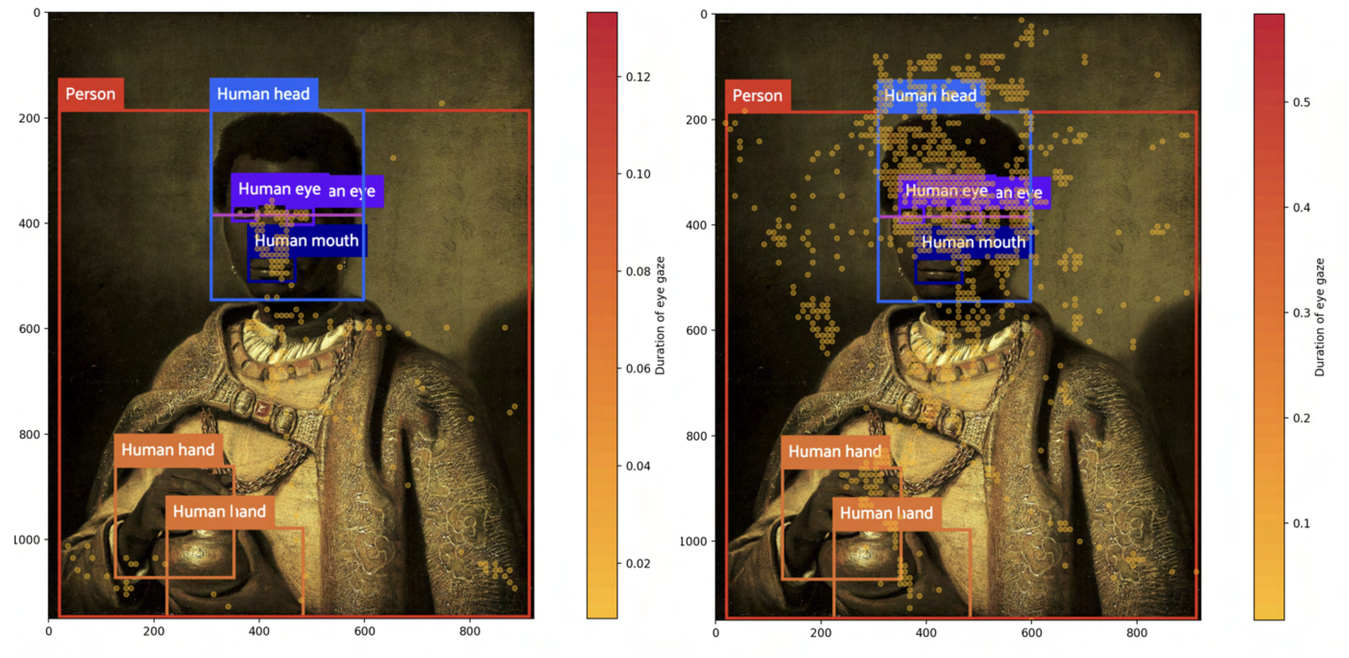


Figure 6: Examples of object model predictions versus ground truth. Predicted bounding boxes are dark blue color with the model's confidence scores in brackets.



(a) Participant with a lower interest in the painting

(b) Participant with a higher interest in the painting

Figure 7: Example of generated heatmaps and ground truth bounding boxes for two participants over one of the VR exhibition paintings

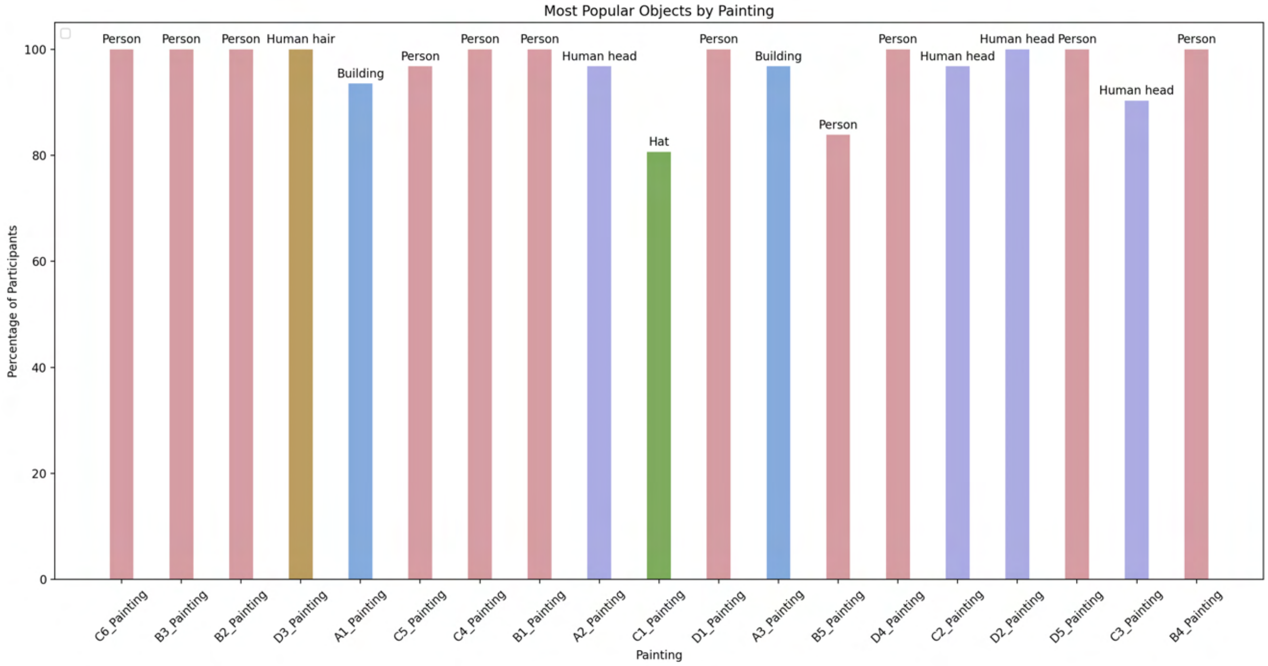


Figure 8: Most popular objects by paining based on average gaze duration. The height of the progress bar represents the percentage of participants who had this object as the most popular.