A. Dataset

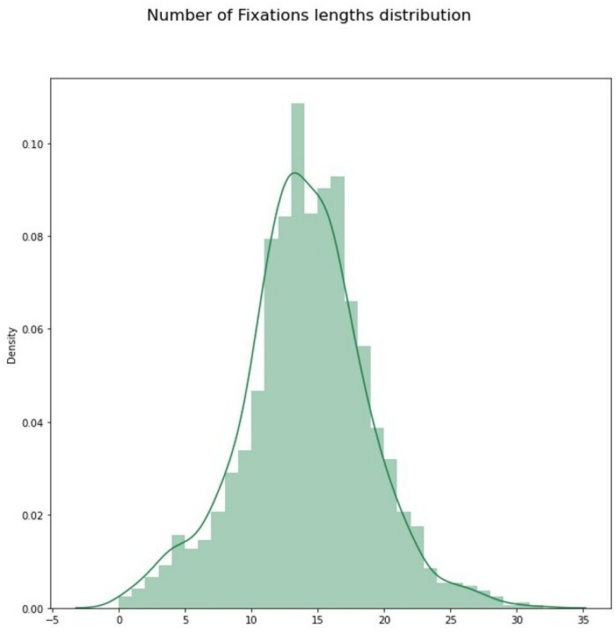

Figure 1. Fixation length density plot.

A.1. Artistic Styles

Style	Number of paintings
Abstract art	15
Baroque	12
China ink	10
classicism	7
cubism	14
Early Renaissance	10
Expressionism	9
High Renaissance	11
Impressionism	12
Luminism	11
Neo-Impressionism	10
Realism	10
Romanticism	10
Surrealism	5
Tonalism	11
Ukio-e	13
Total	168

Table 1. List of artistic styles and the number for paintings for each style.

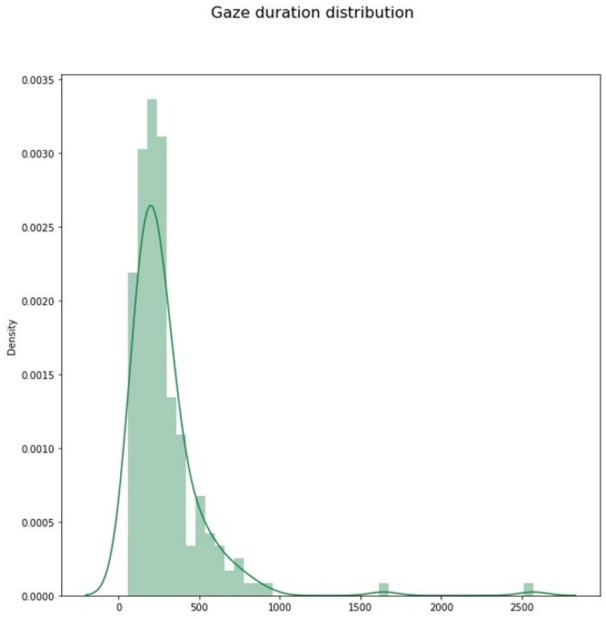

Figure 2. Duration density plot.

Our dataset is comprised of 168 painting images scrapped from online sources. They come from different artistic styles and movements. Some are new movements like Surrealism, Cubism and Tonalism which emerged in 20^{th} century. Others are older and belong to the 19^{th} like Romanticism Impressionism and Luminism. Older styles go back to the $15^{th} - 16^{th}$ century like Baroque, Early and High Renaissance. We also incorporated oriental painting styles like the Chinese ink style and the Japanese Ukio-e. Some paintings represent the world with high detail and fidelity, while others represent distorted views like in Cubism or completely abstract art. This diversity gives our dataset a better distributional representation of the artistic world compared to [1]. The dataset has an average of 10.5 painting per style. For some new styles like surrealism, it is harder to find non copyrighted images, while others like the ancient Ukio-e are majorly placed in the public domain.

A.2. Analysis of dataset

We present a further analysis of our proposed dataset in this section. Fig 1 and Fig. 2 present a clearer representation of Fig.3 in the paper and represent the density plots of the

length of fixation saccades and the duration of fixations as detailed in the paper. Due to Covid-19 restrictions, we were not able to use a chin-rest for our tests. This leads to slight variations of the head position of test subjects during the experiments. Fig 3 represents the distribution density plot of the aforementioned variation. We observe a distribution with a mean distance of $55cm$, yet the distances range from $45 - 65cm$ while ignoring some outliers.

Fig 8 shows the density distribution of left and right eye pupil diameter. Both take a bell curve shape with mean close to $3mm$ and vary from $0.75mm$ to $5mm$. Fig 6 also shows a joint distribution of the pupils diameters through a scatter plot. While we see a moderate amount of variation, we can clearly see a linear correlation between the two pupils diameter changes.

Fig 7 shows the distribution of fixation points from the whole dataset. We clearly observe the centrality of the distribution over both spatial axes. The histogram in red shows the Gaussian like distribution on the horizontal axis. The same observation can be made for the blue histogram representing the vertical space axis. These distributions can modeled through the central bias map represented in Fig. 4 as a saliency map and in Fig 5 as colored heatmap.

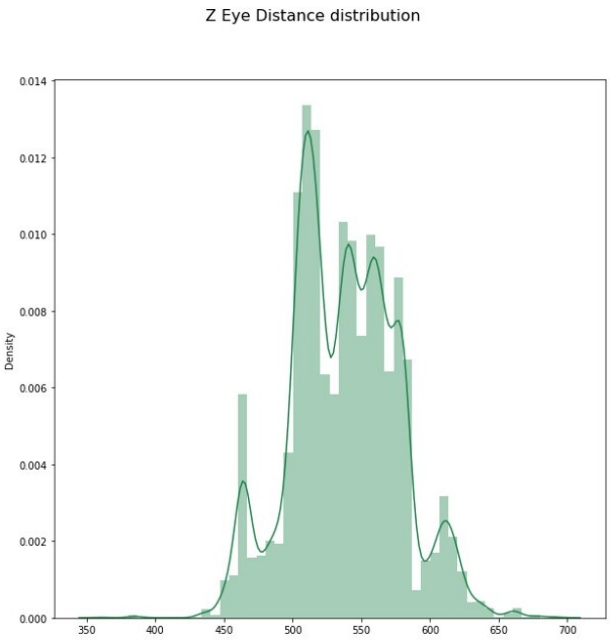

Figure 3. Distance from eye tracker distribution density plot.

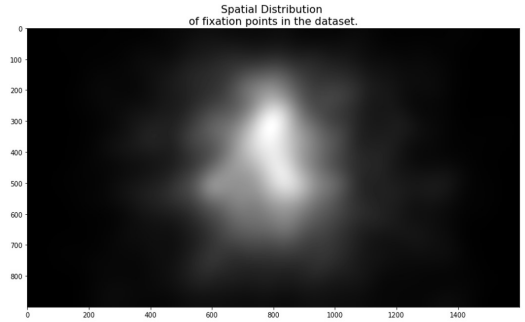

Figure 4. Central bias saliency map.

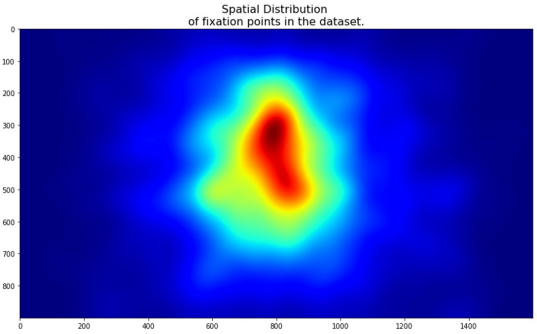

Figure 5. Central bias heatmap.

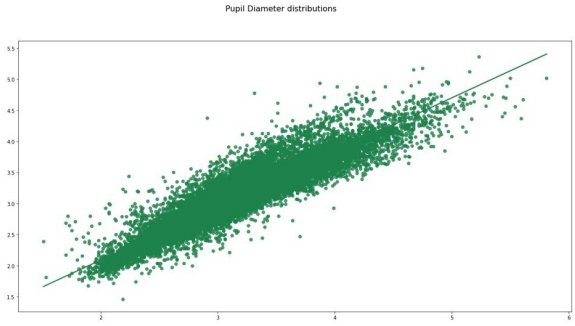

Figure 6. Left and right eye pupil diameter joint distribution.

B. Baseline Model

B.1. Gaussian priors

In order to free our model for modeling task related biases, we trained a module to learn biases from the datasets separately. The module represented the biases in a 16 feature maps representing 2D Gaussian distributions. In fig 9, we present the obtained feature maps. Some biases a locally distributed and represent a limit region of the image while others represent bias over only one of the two spatial axis.

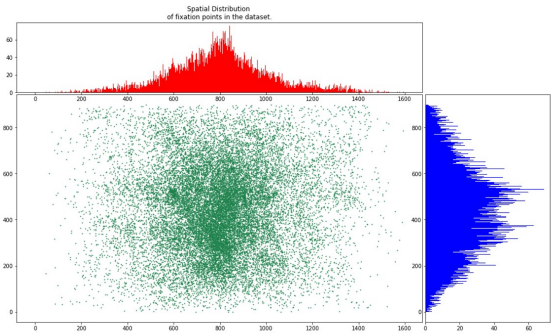

Figure 7. Central distribution of fixations.

B.2. Positional Encoding

Because of inductive bias of permutation invariance that the baseline model branches introduce, we used positional encoding to correct the inductive bias.

As transformers [2] face a similar issue, we set up our positional encoding in a similar manner as represented by the following equation:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \quad (1)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \quad (2)$$

where pos is the position, i is the dimension and d_{model} represents the dimensionality of the encoded vectors.

The positional encoding function for each dimension is represented by a sinusoidal function of a different frequency. The frequency is relative to the size i. These functions permit the encoding values to stay normalised with a range small enough to not override the semantic information contained in the feature vector. We combine the positional embedding and semantic vectors by applying the following function:

$$EV = PE + SV \quad (3)$$

where PE is the vector containing the positional embedding values. SV is the semantic vector resulting from the flattening process after the merging convolution, while EV represents the embedded vector passed to both branches.

C. Visualizations

Fig 10 represents samples from the dataset where we present different images accompanied by their saliency maps at the bottom of each stimuli and Fig 11 presents further results of the predictions of our self-supervised model.

References

[1] Olivier Le Meur, Tugdual Le Pen, and Rémi Cozot. Can we accurately predict where we look at paintings? *Plos one*, 15(10):e0239980, 2020. 1

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

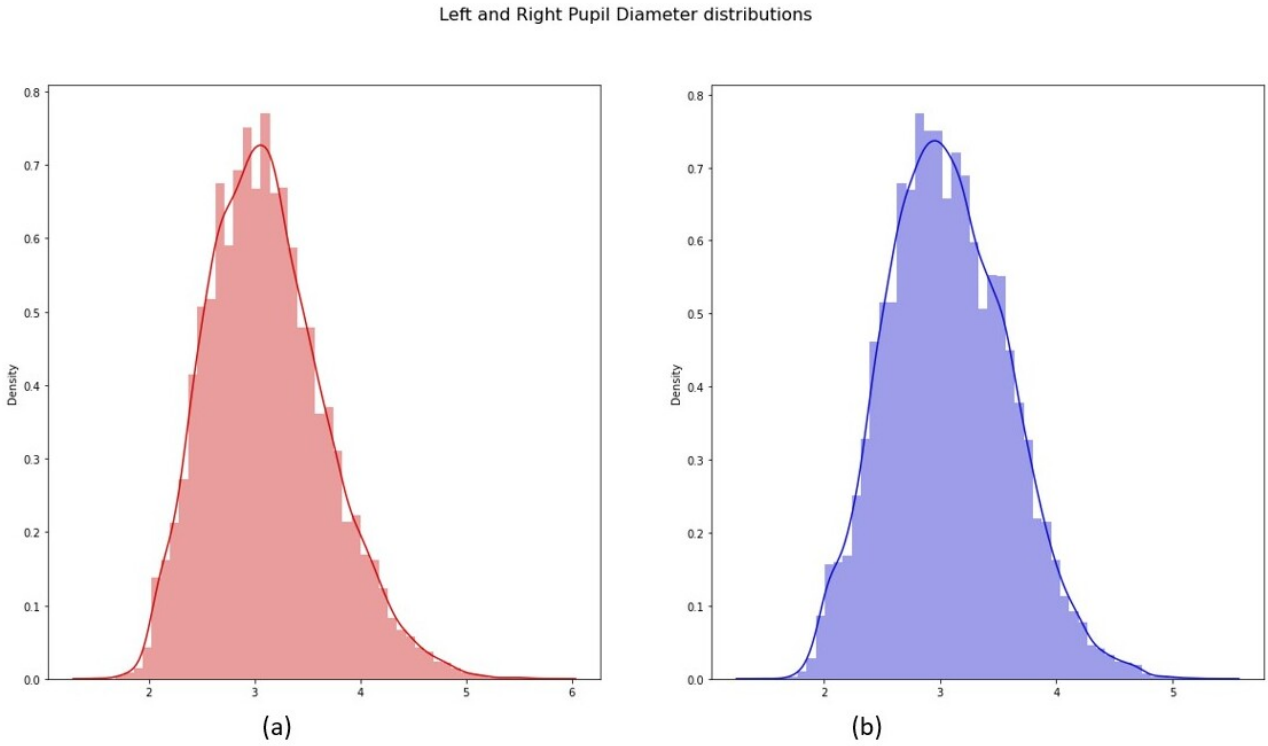

Figure 8. Density plot for (a) Left and (b) Right pupil diameter.

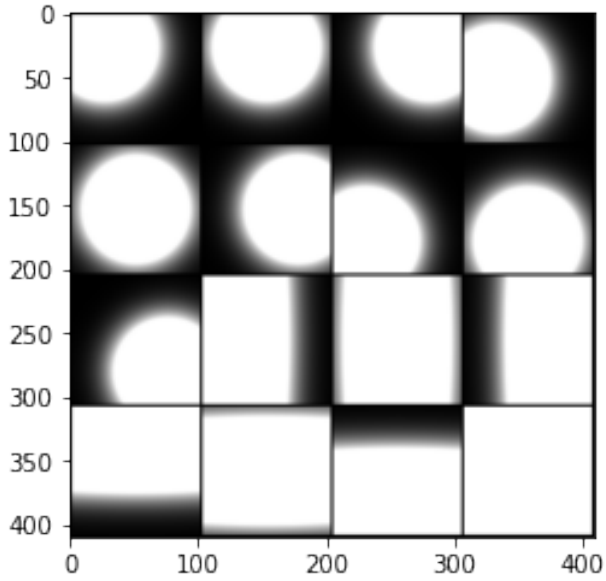

Figure 9. Learnable Gaussian distribution biases.

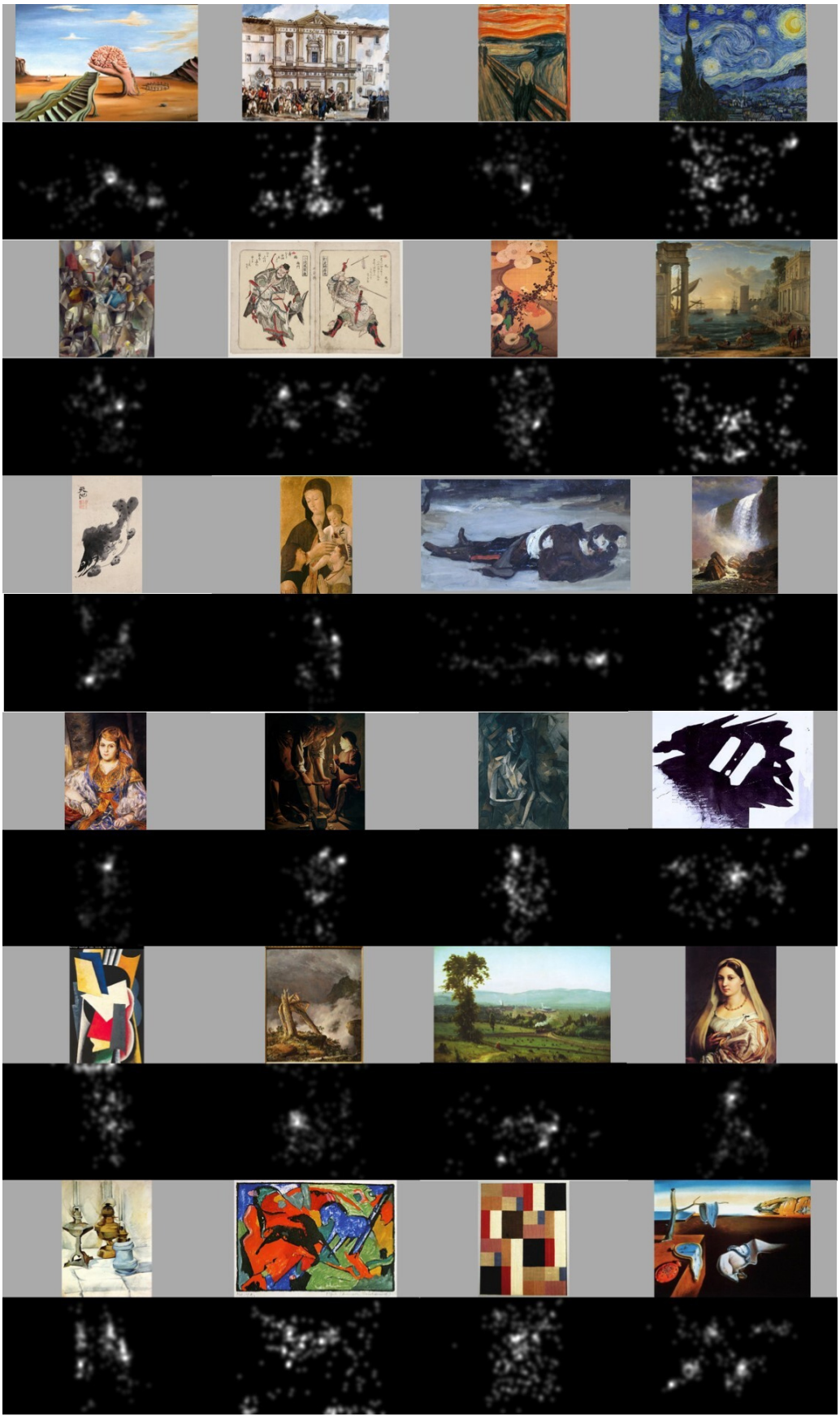

Figure 10. Visualization of our dataset.

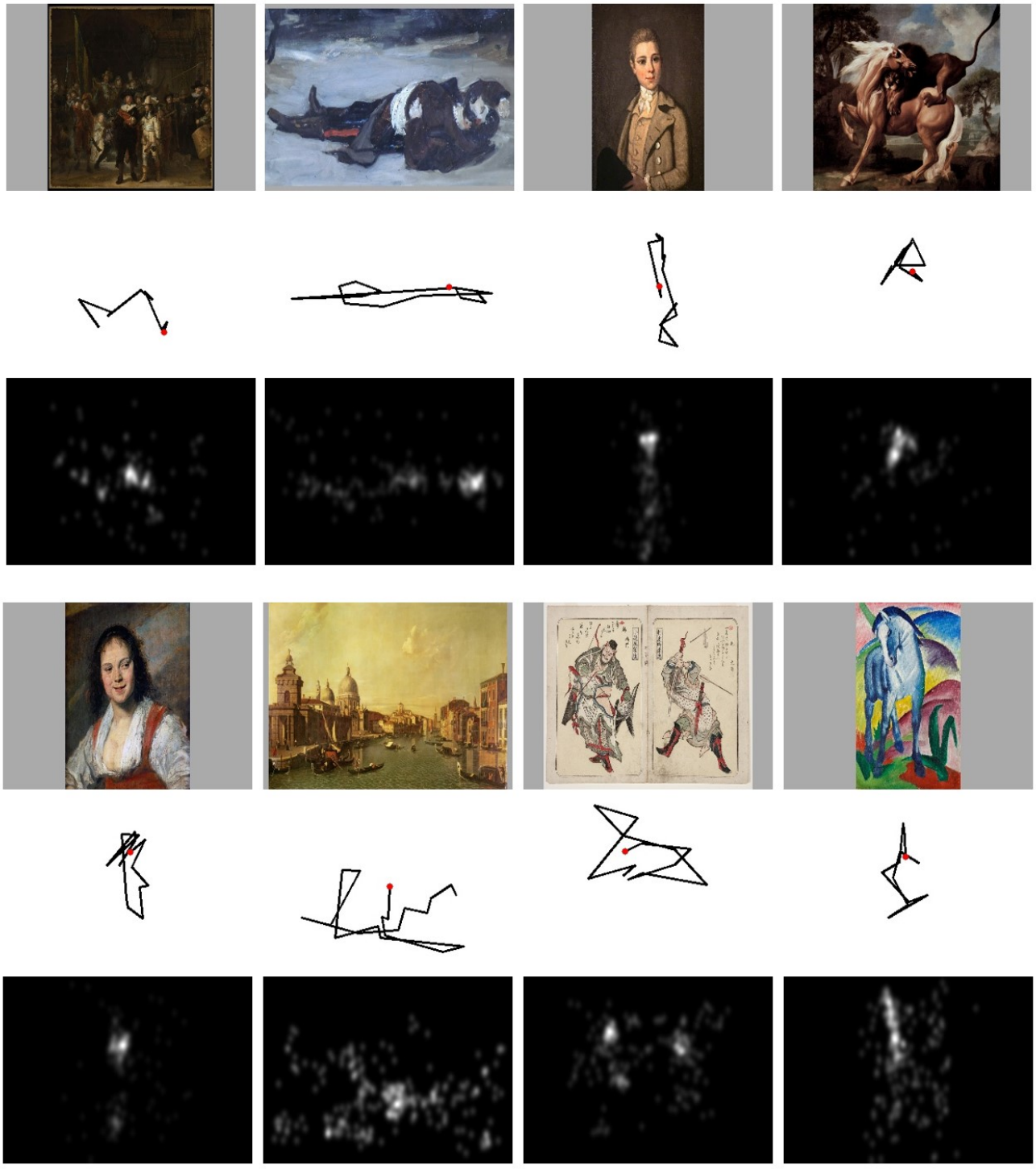

Figure 11. Visualization of scanpaths predicted by our model on our dataset.