Self Supervised Scanpath Prediction Framework for Painting Images

Marouane Tliba[1] *, Mohamed Amine KERKOURI[1*], Aladine Chetouani [1], Alessandro Bruno [2]
Université d'Orléans[1] , Bournemouth University [2]
{marouane.tliba,mohamed-amine.kerkouri,aladine.chetouani}@univ-orleans.fr, abruno@bournemouth.ac.uk

Abstract

In our paper, we propose a novel strategy to learn distortion invariant latent representation from painting pictures for visual attention modelling downstream task. In further detail, we design an unsupervised framework that jointly maximises the mutual information over different painting styles. To show the effectiveness of our approach, we firstly propose a lightweight scanpath baseline model and compare its performance to some state-of-the-art methods. Secondly, we train the encoder of our baseline model on large-scale painting images to study the efficiency of the proposed self-supervised strategy. The lightweight decoder proves effective in learning from the self-supervised pre-trained encoder with better performances than the end-to-end fine-tuned supervised baseline on two painting datasets, including a proposed new visual attention modelling dataset. [1] [2]

1. Introduction

Visual attention represents one of the most advanced and efficient perceptual mechanisms in human beings. It refers to the process used by the Human Visual System (HVS) to filter the $10^{10} bits$ received by the eye receptors each second. This enormous amount of detailed data needs to be filtered and reduced before reaching the visual cortex, where the signal is processed further to pass the relevant information to other regions [29]. Visual attention induces the observer to select specific regions from any visual stimuli to focus on.

Visual attention can be categorized according to the processing path: Bottom-Up attention and Top-down attention. The former is stimulus-driven and highly related to low-level image features such as colour, intensity, texture and so forth [46] [45]. The latter relates to higher-level semantic image features such as faces, text or objects. It refers to intentional, voluntary and task-dependent processes. A lot of works have tackled both attention types under different perspectives. In our paper, we focus on the task-agnostic Bottom-Up attention carried out during free-viewing experimental sessions. The modelling and prediction of saliency and scanpaths became a cornerstone task that improves the efficiency of many other computer vision applications like indoor localization [19], image quality [1], image watermarking [25], image compression [39],image search and retrieval [49] or image enhancement for people with CVD (Colour Vision Deficiency) [9].

An increasing interest in developing accurate attention prediction systems has been noticed as perception mechanisms have demonstrated effective in laying out intelligent systems. However, using supervised learning methods proved limits due to the unavailability of publicly available and manually annotated data. Self-Supervised Learning (SSL) allows to leverage the underlying data structure to extract supervisory signals by enabling the model to learn more relevant information from observing data structure. SSL methods can be divided into several categories: Generative, Discriminative and Generative-Discriminative (Adversarial). Generative approaches try to infer a lower-dimensional informative representation from learning a probability distribution that is similar to the distribution of the real data (e.g. Autoencoder [4]). Discriminative approaches use a more conditional representation method, as they use an intermediate step where they create a proxy discriminative task that learns from the embedded relationships in the data distribution. In this case, the acquired knowledge from the proxy task cloud be leveraged to any new downstream task. Finally, unlike generative models, Generative-Discriminative approaches use a discriminative model jointly with the generative one. For instance, GANs (Generative Adversarial Networks) use a generator to model by playing a Min-Max game with the discriminator working out a realistic representation of the data distribution [21].

Self-supervised discriminative methods can also be categorized according to their objective into **Similarity maximization** approaches (SM-SSL) and **Redundancy Reduction** approaches (RR-SSL). SM-SS uses multiple strategies as learning in contrastive (e.g., mutual information maximization, instance discrimination) [30] (SimCLR [13], and MoCo [26]), Clustering (SwAV [10]), and knowledge distillation (BYOL [22] and SimSiam [15]) with similarity measuring between the representation of the inputs replacing the

*Equal contribution; the order of first authors was randomly selected.

[1]Code and dataset will be availble here : https://github.com/kmamine/SSLArtScanpath

[2]Funded by the TIC-ART project, Regional fund (Region Centre-Val de Loire)

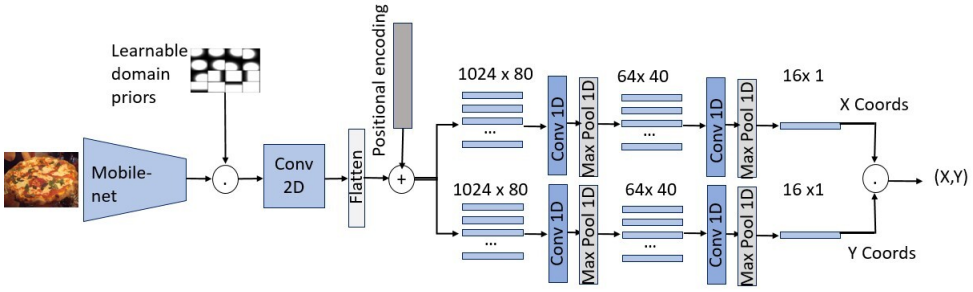

Figure 1. Baseline model architecture.

intermediate tasks. RR-SSL reduces redundant information inside the representation while maximizing the mutual information with the models generated from augmented views of the pristine input. In our context, we use an RR-SSL approach to train our model while exploiting various techniques and strategies introduced by multiple SM-SSL methods. Our method extends the recent advances on SSL that adopt information bottleneck inductive bias in further detail. As a result, we maximize the learning of different painting movements of the same artistic image while reducing the information related to the induced distortions. In other words, we set out a trade-off between two crucial points: having our model as informative as possible on different augmentation of stimuli semantic features and forcing the learned representation to be effectively invariant to style transformations.

To sum up, the contributions of this work are summarized as follows:

1. We propose a new self-supervised framework adapted to the characteristics of painting images.
2. We introduce a new publicly available dataset for art visual attention modelling, including saliency maps and scanpath sequences for paintings images from multiple artistic movements.
3. We present quantitative results on painting benchmarks to show the relevance of the proposed self-supervised representation learning for the related visual attention modelling downstream task.

The rest of the paper is organized as follows: In section 2, we review a part of the scientific literature concerning scanpath and previous contrastive learning approaches. In section 3, we detail the proposed SSL method to improve our model. Section 4 describes our new dataset and the protocols adopted to acquire and analyse the data. We present the evaluation results and protocols of our model in section 5. Finally, section 6 ends the paper.

2. Related Work

Scanpath prediction: Scanpath prediction is an important task that has gained popularity lately. For instance, the winner-take-all (WTA) module presented in [28] predicts scanpaths out of saliency maps peaks. In [37], a stochastic approach was proposed to generate scanpaths using a predicted saliency map and modelling the probabilities of several biases (i.e. saccade amplitudes and saccade orientations). In [50], the authors conceived the saliency map as a 2D gravitational field affecting the trajectory of a mass representing the gaze. In [44], the authors proposed a model that uses high-level features from CNN and Memory Bias, including short-term and long-term memory for scanpath prediction. In [3], the authors presented a deep model where saliency volumes are predicted, then sampled to generate scanpaths. The same authors introduced later PathGan [2] that uses a Long Short Term Memory (LSTM) network working together with a conditional GAN architecture to predict the scanpath of a stimulus. The so-called DCSM (Deep Convolutional Saccadic Model) [5] predicts the foveal saliency maps and temporal duration while modelling the Inhibition of Return (IoR). An end-to-end model which simultaneously predicts the scanpath and saliency map of an image was introduced in [33]. A similar approach was successively generalized for 360° images [34].

Self-Supervised learning: Self-supervised learning approaches aim to learn the underlying feature representation from unlabelled data. Nowadays, increasingly more complex scenarios demand accuracy and the ability to generalise information from multiple tasks. Restricted Boltzman Machines (RBMs) introduced in [42] represented an essential precursor to SSL. They were successively improved in [12]. Denoising Autoencoders [48] aim to decrease the distance between data points outside and inside the manifold. SSL can also trace its history to autoregressive models [6] where individual data point distributions are more accessible to the model than the data as a whole. In computer vision, the employment of Siamese Networks [8] for representation improvement started in the '90s with Self Organising Nets [7]. Several SSL approaches rely on Siamese networks as a fundamental structural architecture. At the same time, new loss objectives such as contrastive loss [16], triplet loss [43], NCE loss [24] and so forth were proposed. Most modern approaches use an augment view method to create positive pairs. For instance, SimCLR [14] employs two identical networks trained on a

contrastive loss. Moreover, they used large batches in their training to sample the negative instances. On the other side, in MoCo [26], the authors exploit a dynamic memory bank of representations as a FIFO queue for negative sampling. SwAV [11] introduced a non-contrastive method using the Sinkhorn-Knopp algorithm [17] to cluster the features and reduce the distance between distributions. BYOL [23] uses an augmented view to learn the similarity without using a contrastive approach. It does not employ negative samples for comparison. It minimises the L_2 loss between the online and target features like MoCo. Yet the momentum encoder parameters are optimised using an Exponential Moving Average (EMA). Barlow Twins [51] aim to decorrelate the features while reducing the distance between the representations of the same images. They first calculate an empirical cross-entropy matrix between the elements of the two parallel networks. They then minimise it by comparing it to an identity matrix.

3. Proposed Method

3.1. Baseline Model

This section provides our baseline model, a lightweight deep neural network for image scanpath prediction. The baseline model is a lightweight architecture and predicts a variable-length scanpath. In fig 1 the main architecture components are shown, respectively, a pre-trained MobileNet network encoding an image into a different representational space, learnable domain prior bias maps concatenated to the output of the feature extractor, representing the visual attention-related bias, a merging convolutional layer, which combines two concatenated features distributions. The 2D feature maps' output is flattened onto a 1D vector and merged with a positional encoding vector. Then, the encoded features feed two branches for predicting the feature extractor's vertical and horizontal coordinates. CNNs are labour intensive and applied pixel-wise, which makes the dimensional complexity of the visual stimuli high. MobileNet [27] was introduced for mobile and embedded platform operations. The employment of point-wise and depth-wise separable convolution makes the network lightweight. Here, we employ MobileNet as a feature extractor that successfully finds semantic information relevant to our task with a way far shorter training time than in [2] and a much lower data size than ImageNet's [18].

Experiments and findings [47] reveal that viewers tend to focus their attention on the central regions of a given scene in the visual field (centre bias). It also relates to standard photography practices, such as taking pictures by placing the cameras' field of view over subjects of interest. The latter aspect impacts observers' gazes towards the central regions, which empirically stand out from the spatial distribution of fixation points gathered with eye-tracking sessions. We also formulate "Central bias" [38] with a spatially Gaussian distribution having mean position in the image centre ($\mu_{xy} = (image_w/2, image_h/2)$), and the standard deviation (σ) adaptive of datasets. The dataset biases are represented with a set of Gaussian distributions with different means and standard deviations 2. Each distribution is expressed by Eq.1 and represented by a 2D heatmap, namely a "Prior map":

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp^{-((\frac{x-\mu_x}{2\sigma_x})+(\frac{y-\mu_y}{2\sigma_y}))} \tag{1}$$

$$S = \{G_1(x,y), G_2(x,y), ..., G_16(x,y)\} \tag{2}$$

where S is the set of Gaussian distributions, (x,y) represent the spatial coordinates of a point in the map. (μ_x, μ_y) and (σ_x, σ_y) are the corresponding mean and standard deviation of the distribution, respectively.

Here, the model learns 16 prior map distributions to embody them with the features extracted out of the encoder. The step generates features, including domain-specific and image attributes. The method frees the encoder from the bias modelling task to focus its performance on stimuli-specific properties. The two branches of our baseline model are inspired by PointNet [41]. As a result, an inductive bias for effective vector-wise features extraction is introduced, and by extension, the generated fixation coordinates become invariant against permutations. However, this property presents a hypothetical unsound argument related to scanpaths being of a sequential nature. Therefore, we tackle it by adding a positional encoding module to our architecture.

3.2. Training

At first, the model was trained on 9000 natural scene images from Salicon [31] dataset and validated on 1000 images. The coordinates of the scanpath were normalized before training, following the image's respective dimensions. Short scanpaths are padded to a specific shape of 16 fixations since 97.34% of the pictures does not surpass this number of points. We trained our model using the following loss function:

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N}\sqrt{{y_x}^2 - \hat{y_x}^2} + \frac{1}{N}\sum_{i=0}^{N}\sqrt{{y_y}^2 - \hat{y_y}^2} \tag{3}$$

where N represents the number of fixation points of the scanpath, y_x and y_y represent the coordinates of the ground truth scanpath, while $\hat{y_x}$ and $\hat{y_y}$ are the coordinates of the predicted scanpath.

Then, to test the model on our proposed dataset in Section 4, we further fine-tune it on the dataset in a supervised manner using Adam's optimizer with a learning rate of 10^{-4} and a batch size of 1. The model converged after 175 epochs. Finally, to further improve the results, and due to the particular nature of painting images, we retrained the encoder in a self-supervised manner, as shown in Section 3.3. After

that, we trained our model to conduct the same testing while freezing the encoder ensued from the Self-Supervised training, well-known as a linear evaluation of the model. Again, we used the same hyper-parameters for training as supervised learning. As a result, the SSL model converged in 65 epochs, while the supervised model converged after more than 175 epochs.

3.3. Self-Supervised Learning

Paintings images are different by nature from natural scene images. The diversity of styles and movements alongside techniques variate images' colors and textures compared to pictures taken by cameras. While these low-level features might be influential for the first few fixations, semantic features have a more significant influence in guiding the user's gaze as long as recognizable objects are depicted. That means that artistic low-level features influence should be decorrelated to high-level features influences for our encoders' representation. Thus, we tried reducing the variance related to the painting style. To this end, we propose a new self-supervised learning approach inspired by several past approaches [51] [23], and introduce the use of a new transformation primarily related to the painting images domain (see Fig. 2).

Neural Style Transfer is a technique used to merge two images: a content image containing semantic information and a style image providing visual features information like color and texture. The resulting image would be a composite that retains the content semantics while obtaining the same visual style as the other image, first introduced in [20]. That is usually done by extracting mid-level features from both images in a neural network and then aligning those features to obtain a better representation. The first loss introduced was the following equation:

$$L = \alpha MSE(R_C, R_X) + \beta MSE(R_{GS}, R'_{GX}) \tag{4}$$

Where R_c and R_x are representations extracted from a mid-level layer of the neural network for the content and stylized image. While R_{GS} and R'_{GX} are gram matrices constructed from features extracted multiple levels from the neural network for the style and stylized image, respectively. The first term is called content loss, and the second one is style loss. In our work, we used the model proposed in [20] for data augmentation described below.

General Description: As depicted in Fig 2, our method is a compound algorithm relying on two primary steps for training during each batch. The first step leverages the Barlow Twins method by training on several neural styles. The following step leverages the representations obtained during the previous one to reduce the distance between the augmented distributions and the original one.

Data augmentation: Like other approaches, our method relies on the joint embedding of distorted images. We use several image transformation techniques for this purpose. As above mentioned, the first transformation technique used is NST. We chose a group of paintings belonging to several art movements to construct the embedding styles transformations set $\mathcal{S}$ and a context image group of 50000 images C scrapped from several web sources (i.e. Wikiarts dataset, Wikipedia Communes, etc.). The latter represents the training set of our encoder. Several image augmentations ($\mathbb{S} = S_1, S_2, ..., S_N$) are obtained after applying the transformations $\mathcal{S}$. On top of these changes, we used another set of image distortion operations $\mathcal{T}$ where we overlay common modifications like cropping, grayscale conversion, blurring and others.

At each time-step of our algorithm, two style augmentations are randomly selected and applied to the input x before employing the random transform $\mathcal{T}$ then forwarding the results x_a, x_b to the following mechanisms as in the equation below:

$$x_a, x_b = \mathcal{T}(Rand\_pair(\mathcal{S}(x|\mathbb{S}))) \tag{5}$$

Modified Barlow Twins: The augmented batches x_a and x_b are then fed to function f which corresponds here to the untrained encoder network (i.e. MobileNet), which we want to optimize utilizing unlabeled data. The resulting representations R_{S_a} and R_{S_b} are in accordance with the original Barlow Twins (BT) algorithm. An empirical cross-correlation matrix C_r is calculated between the two aforementioned representations. A loss is calculated according to the following loss [51] :

$$\text{where} \qquad C_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2}\sqrt{\sum_b (z^B_{b,j})^2}} \tag{6}$$

This loss pushes the obtained cross-correlation matrix to match an identity matrix of the exact dimensions. It ensures the two representations' features are aligned independently from the transformations applied to the input x. At the same time, this pushes the function f to decorrelate the non-corresponding features. That stems from information theory, where we are trying to reduce redundancy and, by extension, maximize the mutual information between the semantic elements. So this step iterates N times, with N being the number of styles belonging to the set $\mathbb{S}$. As the encoder switches between different styles, we introduced another parameter update operation using the exponential moving average across time to stabilize it further. The procedure is described as follows:

$$\theta_T := (1 - \gamma)\theta_T + \theta_{T-1} \tag{7}$$

where θ_T are the parameters weights at the step T.θ_{T-1} are the weights of the previous step and γ represents a weight decay term as such as $\gamma \in [0, 1]$.

In each step, we also preserve a representation of one style to a style bank $\mathcal{M}$ for the second major step of our algorithm.

Global Style Update: After executing N iterations of the modified Barlow Twins, the algorithm executes another step update before passing to the next batch iteration. In this step,

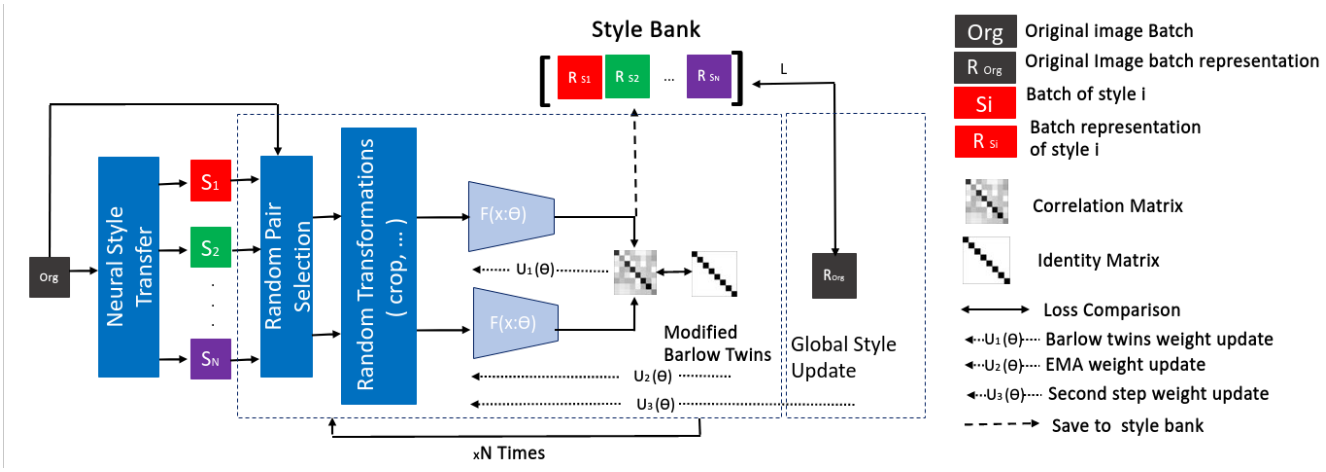

Figure 2. Flowchart demonstrating the proposed Self-supervised learning algorithm.

we use the style bank $\mathcal{M}$ where representations of different styles during different time steps are saved. In a fashion similar to [23], we use the L_2 loss (Eq. 8) to reduce the distance between the distribution of the original image and the distributions of other styles. This is done by forcing the function f to predict a representation R_{Org} similar to R_{S_i}, where $S_i \in \mathbb{S}$. This further helps stabilize the training moving R_{Org} closer to the style distributions' centroid.

$$\mathcal{L} = \|\bar{f}_\theta(R_{Org}) - \bar{R_{S_i}}\|^2 = 2 - 2 \cdot \frac{\langle f_\theta(R_{Org}), R_{S_i} \rangle}{\|f_\theta(R_{Org})\| \cdot \|R_{S_i}\|}. \quad (8)$$

Optimization: We follow the optimization method described above; we use the Adam optimizer for 450 epochs with a batch size of 128 on 50000 painting images, which are augmented to 5 neural styles. We use a learning rate of $5{\times}10^{-4}$ with a reduction on plateau scheduler with ten epochs waiting time. We used a search for the parameter λ of the loss function and found $\lambda = 0.5$. The training was performed using an Nvidia Quadro RTX 8000.

Algorithm 1 pseudocode .

```
dataloader = []
dataS =[]
StyleBank = []
for data in dataloader :
    dataS  = StyleTrasferTrasform(data)
    # Datastyles list of size N
    for i  in range(N):
        i, j = rand(0,N)
        #  Select 2 Styles at random
        S1 = RandomTrasform(data[i])
        S2 = RandomTrasform(data[j])
        z_a = f(S1) # passing S1
        z_b = f(S2) # passing S2
        z_a_norm = (z_a - z_a.mean(0)) / z_a.std(0) # NxD
        z_b_norm = (z_b - z_b.mean(0)) / z_b.std(0) # NxD
        # Saving params temporally
        f_temp.params = f.params
        # Saving representation to Stylebank
        StyleBank.append(z_a_norm)
        # Barlow Twins optimization
        loss = BTLoss( z_a_norm, z_b_norm)
        loss.backward()
        optimizer.step()
        # EMA
        f.params = gamma * f.params + \
            (1-gamma) * f_temp.params
    loss = []
    Sorg = f(data)
    # calc. L2 loss for each repesent.
    for i in range(len(StyleBank)) :
        loss.append( L2Loss(Sorg,StyleBank[i]))
    # Global Style Update optimization
    loss = loss.mean()
    loss.backward()
    optimizer.step()
```

4. Dataset

Building up a painting dataset is usually conditioned by copyright and intellectual property laws. Constructing a large painting dataset is a laborious and not trivial task, especially for the personal visual attention and perception tasks where the need to conduct experiments with human participants is essential. We constructed a new dataset consisting of 170 artistic paintings images and the corresponding eye-tracking data. The dataset is subject to expansion at several levels, including the number of images, subjects, tasks, etc. In this section, we provide detail and properties of the dataset.

Images: We collected 170 copyright-free painting images from Wikiart. The paintings belong to 16 different art movements and art schools. The dataset consists of landscape and portrait images. During the experimental campaign, it is necessary to consistently show participants the presented paintings. As such, we centred the paintings on a uniform grey image having 16/9 ratio (1600px x 900px). The paintings underwent resizing to fit the images while persevering their aspect ratio. The prominence of the grey stripes depends on the image's original aspect ratio.

Participants: There were 30 participants, from university undergraduates to university staff and faculty in an engineering school. All the participants are not experts in art subjects and have reported no previous experience or formal training. The cohort added up to 10 female and 20 male participants. Their age ranged from 18 and 62 years old. The mean age was 28 years old with a standard deviation of 11.5 years (

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan [2]	0.9608	0.5698	**0.9530**	0.8172	0.8252	-0.2904	0.0825
Le Meur [37]	0.9505	0.6231	0.9488	0.8605	0.8457	0.8780	0.4784
G-Eymol [50]	0.9338	0.6271	0.9521	**0.8967**	0,8524	0.8727	0.3449
SALYPATH [33]	**0.9659**	0.6275	0.9521	0.8965	**0,8605**	0.3472	0.4572
Our model	0.9552	**0.6466**	0.9509	0.8873	0.8600	**1.0062**	**0.5170**

Table 1. Results of scanpath prediction on Salicon.

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan [2]	0.9237	0.5630	0.8929	0.8124	0.7561	-0.2750	0.0209
DCSM (VGG) [5]	0.8720	0.6420	0.8730	0.8160	0,8007	-	-
DCSM (ResNet) [5]	0.8780	0.5890	0.8580	0.8220	0,7868	-	-
Le Meur [37]	0.9241	0.6378	**0.9171**	0.7749	0,8135	0.8508	**0.1974**
G-Eymol [50]	0.8885	0.5954	0.8580	0.7800	0,7805	**0.8700**	0.1105
SALYPATH [33]	**0.9363**	0.6507	0.9046	0.7983	0,8225	0.1595	0.0916
our model	0.9201	**0.6759**	0.9099	**0.8351**	**0.8352**	0.8186	0.1926

Table 2. Results of scanpath prediction on MIT1003.

Model	NSS	Congruency
Our model SSL	**0.9987**	**0.2673**
Our model SL	0.8554	0.2641

Table 3. Results of scanpath prediction on Le Meur paintings with and without SSL.

28 ± 11.5). All participants reported a normal or corrected vision and passed a sample Ishihara test with the three images representing 12,74, and 42 to check out any CVDs (Colour Vision Deficiencies). No participant exhibited any symptoms during the short examination.

Hardware and Experimental protocol: Participants were asked to watch the painting in the most natural manner possible for a free-viewing session, and no visual task was assigned. The experiments were held in compliance with the declaration of Helsinki, and all participants signed off a written consent. Observers sat down in front of a 25" screen at a distance of 55 cm. The screen covered 52° horizontally and 38° vertically. The screen resolution was 1920 x 1080 pixels, the visual stimulus size was 1600 x 900 pixels and covered 43° horizontally and 32° vertically. Each degree of visual angle corresponded to 37 pixels. A fixed Tobii X2-30 eye-tracker was used for the experiments with a frequency of 30Hz. Though the sampling frequency is low, it was enough to gather fixation points throughout the experimental campaign. Each image was shown to 15 participants for 15 seconds, with a 3-second blank grey screen being displayed between two consecutive pictures to refresh the viewers' retina and avoid any bias impacting the new upcoming visual stimulus perception. Afterwards, a further analysis step was carried out on the first 4 seconds of each image and the corresponding eye-tracking data. Fixation points related to later than the fourth second of observation will be used for other purposes. Each experimental session lasted for 15 minutes.

Dataset Analysis After gathering eye-tracking sessions data, we aggregated fixation maps from all participants' data for each image shown. We generated saliency maps out of fixation maps using a common smoothing filtering technique: a Gaussian kernel with the size of 1° of visual angle, which is formulated by the following equation:

$$\text{with } \text{Fixmap}_i(x,y) = \begin{cases} 1, & \text{if } (x,y) \text{ in } (X,Y)_i \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where x, y are the coordinates in the 2D space. G_σ represents the Gaussian kernel with a standard deviation σ. (X, Y) are the set of fixation point coordinates for all viewers of the image i.

The distribution of the fixation durations follows a skewed long-tailed distribution, as shown in Fig 3 (a). While the distribution of the number of fixation points follows a near symmetric bell curve distribution, as shown in Fig.3 (b). Fig 4 (a) shows a polar distribution histogram of the relative saccadic angles direction. The polar axis represents the angles of the distribution, and the radial axis is the statistical density. Fig. 4 (c) represents a saccade length distribution. Again, an asymmetric skewed heavy-tailed distribution is noticeable, with most of the saccades being short. That is also confirmed in Fig.4(b), which depicts the joint distribution of saccadic lengths and directions. The polar axis represents the saccade relative angles as the radial axis counts the saccades' lengths and distribution by color as detailed in the sidebar.

5. Experimental Protocol

We follow standard practices by evaluating our baseline architecture on natural scene images to ensure a fair comparison protocol for our model with others. For instance, we use 5000 images from Salicon [31] dataset, and then in a cross-dataset evaluation manner, we run a comparison without fine-tuning our model on MIT1003 [32] dataset. To that end, we use two hybrid metrics (i.e. NSS [40], Congruency [35]) to compare the predicted scanpaths with the ground-truth saliency maps. Furthermore, we adopted a vector-based metric (i.e. Multimatch) to match the predicted scanpath with

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
Our model SSL	**0.9454**	**0.7019**	**0.9423**	**0.8196**	**0.8523**	**0.8083**	**0.1641**
our model SL	0.9231	0.6952	0.9083	0.7800	0.8266	0.4179	0.1247

Table 4. Results of scanpath prediction on our paintings with and without SSL.

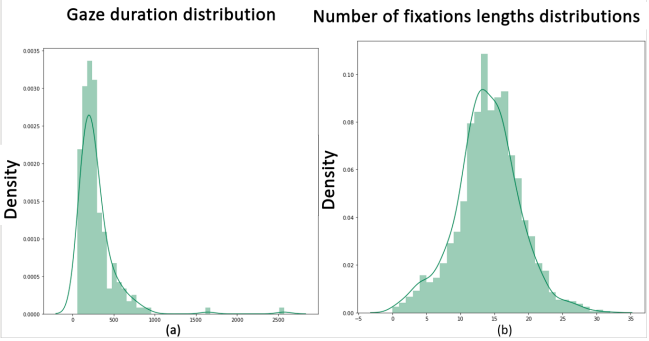

Figure 3. (a) Distributions of fixations duration. (b) Distributions of number of fixations per scanpath.

the ground-truths based on 5 characteristics (Shape, Direction, Length, Position and Duration). We use the first 4 for comparison here as our model does not predict the duration.

For fairness, we did not compare our model with other models because of the difficulty of fine-tuning them on our dataset. However, as our model shows competitiveness with other models on natural scene images benchmarks, we evaluated it after fine-tuning it on our proposed painting dataset. To assess our SSL method on the painting dataset, we compare the model resulting from SSL with the fine-tuned SL model.

5.1. Results on natural scene datasets

The results of the comparison of our baseline model on natural scene images are described in Tables 1 and 2. The results on Salicon [31] dataset show that our baseline model scored the highest outcomes for the hybrid metrics NSS and congruency. Especially on NSS, where it surpasses the runner-up by a wide margin. The results for the MultiMatch metric were very competitive as our baseline model was overcome by SALYPATH [33] with a very negligible margin on the mean score while demonstrating abilities in learning the distribution of the directions. On MIT1003, Our baseline model achieved competitive results on the NSS while being surpassed by a very negligible margin on congruency. It also proved a better ability to generalize the MultiMatch mean (MM Mean) score while maintaining the highest direction score and achieving the same position score.

5.2. Results on paintings

As the comparing models cannot be fine-tuned on our proposed dataset, we present the results of our proposed model and the model ensuing from the linear evaluation of the SSL training. We tested the model on our proposed and Le Meur datasets using only hybrid metrics since only aggregated saliency maps are provided (i.e. there are no scanpaths).

Table 3 shows results obtained from testing our fine-tuned model denoted by "Our model SL" and the model resulting from the self-supervised training as "Our model SSL". Our models showed very high performance on the NSS metric. It is mainly due to the content of the chosen paintings in the Le Meur painting dataset, which was in a cross-dataset validation fashion. Furthermore, most images and styles depict semantics close to reality; this realistic art style shows slight deviation from natural scene images, thus explaining the high score. The same consideration goes for the congruency where the SSL model surpasses the fine-tuned one. The testing on our dataset was done using 50 images. In addition to comparing the hybrid metrics, we also assessed the vector-based metric of MultiMatch. Table 4 clearly shows that the SSL method surpasses the fine-tuned model on all metrics.

Finally, Fig. 5 presents some visualization examples of predicted scanpaths from both the supervised model and the self-supervised model. We notice clearly that the supervised model fixations occupy a centric location in the image with mostly a long tail afterwards towards the upper left corner of the image. As for self-supervised model prediction, we can notice that they spread evenly over the salient regions present in the ground truth.

5.3. Discussion

The presented model proved its effectiveness and competitiveness compared to other similar state-of-the-art models for natural scene images. We extended then the model to painting images. The results obtained from fine-tuning showed room for improvement. Predicting scanpaths is mainly a supervised task, which calls for the use of large amounts of data. Our dataset is the only public one that presents scanpath ground-truths for painting images domain to the best of our knowledge. We proposed a new SSL training method to improve the results that capitalize on previous state-of-the-art methods. Our approach is the first to employ Neural Style Transfer as a data augmentation strategy and the first that uses SSL for visual scanpath prediction. The results indicate that SSL methods suit well our visual attention tasks. Free-viewing is basically a bottom-up attention task, meaning visual attention lies in intrinsic image features. It seems to resonate well with the redundancy reduction SSL methods. The goal is to limit the representation to diminish the amount of information to a minimal size. At the same time represent the most critical information from the scene.

Our SSL method also showed a high ability to converge faster compared to other methods trained for more than 1000

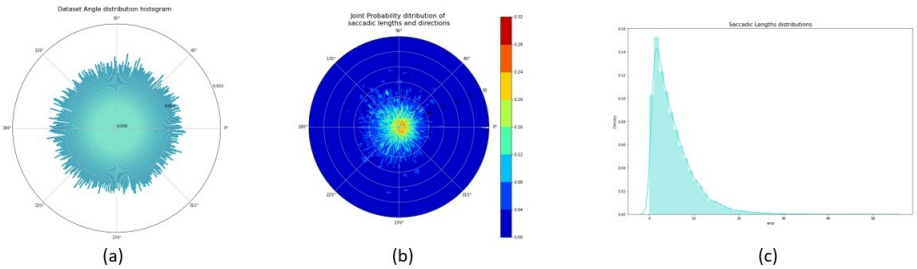

Figure 4. (a) Distributions relative angles of scanpaths. (b) Joint Distributions of relative angles and saccades lengths. (c) Distributions lengths of saccades.

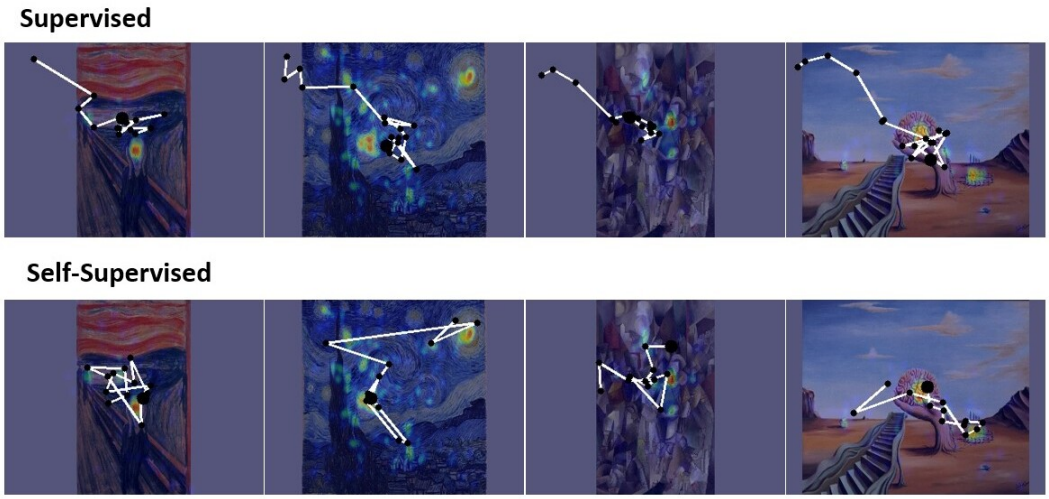

Figure 5. Visualization of predicted scanpaths on painting images overlaid with Ground-Truth saliency maps.

epochs. It also seems to inherit the characteristic of the no need for extensive data from Barlow Twins. The fast convergence is due to the stabilizing parameter weight updates using an exponential moving average at the end of each Barlow Twins update to smooth the transaction between different Styles. The "Global Style Update" step optimizes weights at the end of each batch also proves beneficial to reduce the distance between the styles representations. The use of 128 as a batch size is mainly due to hardware limitations. We also noticed a particular disparity between the results of our dataset and Le Meur paintings [36]. The model's performance indeed has dropped. That is due to the greater diversity in schools, movements and techniques depicted by our dataset, which contains many hard-to-predict images belonging to abstract paintings, oriental ink paintings and cubism. Nevertheless, the competitively empirical and qualitative results show the efficiency of our self-supervised method in extracting relevant features from the paintings.

6. Conclusion

We introduced in this paper a baseline scanpath prediction model that proved competitive against state-of-the-art methods and overcame them in many instances over multiple datasets and metrics. We also introduced a visual attention dataset for painting images. Our dataset is the first publicly available that provides aggregated saliency maps and scanpaths of individual observers. However, the dataset is more diverse in terms of images and thus harder to predict. We also proposed a novel self-supervised learning approach that builds over previous methods. We adopted Neural Style Transfer as a data augmentation technique for self-supervised learning in our work. Furthermore, its performance turned out very well in learning adequate representations for paintings. The above-depicted scenario opens a long way to SSL methods in visual perception tasks related to art and other use cases like painting style, era identification, and aesthetics quality assessment.

References

[1] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. No-reference mesh visual quality assessment via ensemble of convolutional neural networks and compact multi-linear pooling. *Pattern Recognition*, 100:107174, 2020. 1

[2] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 3, 6

[3] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2331–2338, 2017. 2

[4] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *ICML Unsupervised and Transfer Learning*, 2012. 1

[5] Wentao Bao and Zhenzhong Chen. Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 2020. 2, 6

[6] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966. 2

[7] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992. 2

[8] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 2

[9] Alessandro Bruno, Francesco Gugliuzza, Edoardo Ardizzone, Calogero Carlo Giunta, and Roberto Pirrone. Image content enhancement through salient regions segmentation for people with color vision deficiencies. *i-Perception*, 10(3):2041669519841073, 2019. 1

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3

[12] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[15] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[16] Sumit Chopra, Raia Hadsell, and Yann Lecun. Learning a similarity metric discriminatively, with application to face verification. volume 1, pages 539– 546 vol. 1, 07 2005. 2

[17] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 3

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[19] W. Elloumi, K. Guissous, A. Chetouani, and S. Treuillet. Improving a vision indoor localization system by a saliency-guided detection. In *2014 IEEE Visual Communications and Image Processing Conference*, pages 149–152, 2014. 1

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 4

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 1

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3, 4, 5

[24] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[25] Mohamed Hamidi, Aladine Chetouani, Mohamed El Haziti, Mohammed El Hassouni, and Hocine Cherifi. Blind robust 3d mesh watermarking based on mesh saliency and wavelet transform for copyright protection. *Information*, 10(2), 2019. 1

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 3

[27] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[28] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 2

[29] Laurent Itti, Geraint Rees, and John K Tsotsos. *Neurobiology of attention*. Elsevier, 2005. 1

[30] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021. 1

[31] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080. IEEE Computer Society, 2015. 3, 6, 7

[32] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 6

[33] Mohamed A. Kerkouri, Marouane Tliba, Aladine Chetouani, and Rachid Harba. Salypath: A deep-based architecture for visual attention prediction. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1464–1468, 2021. 2, 6, 7

[34] Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, and Mohamed Sayeh. Salypath360: Saliency and scanpath prediction framework for omnidirectional images. *Proc. IST Int'l. Symp. on Electronic Imaging: Human Vision and Electronic Imaging*, pages 168–1 – 168–7, 2022. 2

[35] Olivier Le Meur, Thierry Baccino, and Aline Roumy. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 373–382, 2011. 6

[36] Olivier Le Meur, Tugdual Le Pen, and Rémi Cozot. Can we accurately predict where we look at paintings? *Plos one*, 15(10):e0239980, 2020. 8

[37] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164, 2015. 2, 6

[38] Sabira K Mannan, Keith H Ruddock, and David S Wooding. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, 1996. 3

[39] Yash Patel, Srikar Appalaraju, and R Manmatha. Saliency driven perceptual image compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 227–236, 2021. 1

[40] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005. 6

[41] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[42] David E. Rumelhart and James L. McClelland. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, pages 194–281. 1987. 2

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 2

[44] Xuan Shao, Ye Luo, Dandan Zhu, Shuqin Li, Laurent Itti, and Jianwei Lu. Scanpath prediction based on high-level features and memory bias. In *International Conference on Neural Information Processing*, pages 3–13. Springer, 2017. 2

[45] Marouane Tliba, Mohamed A. Kerkouri, Bashir Ghariba, Aladine Chetouani, Arzu Çöltekin, Mohamed Shehata, and Alessandro Bruno. Satsal: A multi-level self-attention based architecture for visual saliency prediction. *IEEE Access*, pages 1–1, 2022. 1

[46] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. volume 12, pages 97–136. Elsevier, 1980. 1

[47] Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7):4–4, 2009. 3

[48] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[49] Haoxiang Wang, Zhihui Li, Yang Li, Brij B Gupta, and Chang Choi. Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, 130:64–72, 2020. 1

[50] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 6

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 3, 4