

AN INTER-OBSERVER CONSISTENT DEEP ADVERSARIAL TRAINING FOR VISUAL SCANPATH PREDICTION

Mohamed Amine Kerkouri^{1*}, Marouane Tliba^{1*}, Aladine Chetouani¹ and Alessandro Bruno²

¹Laboratoire PRISME, Université d’Orléans, Orléans, France

²IULM University, Milan, Italy

ABSTRACT

The visual scanpath represents the fundamental concept upon which visual attention research is based. As a result, the ability to predict them has emerged as a crucial task in recent years. It is represented as a sequence of points through which the human gaze moves while exploring a scene. In this paper, we propose an inter-observer consistent adversarial training approach for scanpath prediction through a lightweight deep neural network. The proposed method employs a discriminative neural network as a dynamic loss that better models the natural stochastic phenomenon while maintaining consistency between the distributions related to the subjective nature of scanpaths traversed by different observers. The competitiveness of our approach against state-of-the-art methods is shown through a testing phase.

Index Terms— Visual Attention, scanpath prediction, adversarial training, inter-observer consistency.

1. INTRODUCTION

The human retina receives around 10^{10} bits/sec of visual information. Most of this information represents high-definition receptors located in the fovea, which covers approximately 1° of the visual field. This gigantic amount of information is further reduced to 3×10^6 bits/sec before traveling through the optical nerve, and further reduced afterward while traveling through the visual cortex [1]. The mechanism called “Visual attention” is ruled by the previously mentioned anatomical constraints and other further neurological and psychological ones. The observer is induced to only pay attention to some scene’s specific regions. This phenomenon is manifested through saccadic eye movements, representing the gaze shifting from one region to another for a visual stimulus. As eye movements focus on an area, the gaze fixates on specific points, namely “fixation points”. The latter can be collected with eye trackers, allowing the projection of fixation points from multiple observers onto a binary map, better known as a “fixation map”. On top of that, a “saliency map” is generally obtained with smoothing filters to give a blob-shaped spatial

distribution of fixation points over visual stimuli. Saliency maps are usually represented as normalized heatmaps, with each pixel value representing the probability of the pixel catching viewers’ attention. The above-depicted mechanism provides the human visual system with outstanding efficiency. The prediction of scanpaths/saliency map is helpful to a lot of computer vision applications like indoor localization [2], quality assessment [3, 4, 5], image watermarking [6], image compression [7], perception [8], and retrieval [9], CVD detection [10].

The interest of the scientific community in saliency [11, 12] and scanpath prediction has risen lately. For instance, the winner-take-all (WTA) principle was used by Itti et al. [13] in their first work, where the scanpath is extracted from the most salient regions. In [14], the authors generated scanpaths from a saliency map through statistical features derived from several datasets. In [15], LSTM layers and the VGG model were employed with adversarial training. In [16], the saliency map was modeled as a gravity field where the gaze mass travels using physical laws. A foveated saliency map was used along with inhibition of return maps to predict scanpaths in [17]. The authors of [18] presented an end-to-end model to simultaneously predict the scanpath and the saliency map of an image [18], later generalized for 360° images [12].

The authors of [19] proposed a self-supervised training approach to train the model for painting image scanpath prediction. While in [20] they used a domain adaptation approach to generalize the predictive ability from natural scenes to paintings.

Through these previous works, we found that this task still presents some fundamental and interesting challenges. The stochastic nature of scanpaths is a function of the subjectivity of observers, and modeling this inter-observer distribution in a consistent manner proves to not be an evident task. At the same time, modeling multiple observers induces difficulty for neural networks in generating results that emulate the qualitative properties of the real data. So, the main concern manifests itself in how to train a neural network to predict scanpaths while maintaining consistency between the subjectivities of multiple observers.

The proposed method presents the following contribution to solving the aforementioned challenges:

* The first two authors have equal contributions.

This research was funded by the regional TIC-ART Project from Centre-Val de Loire.

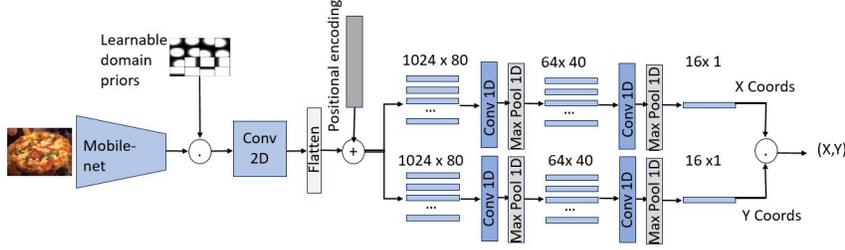


Fig. 1. Generator model architecture.

- We employ an adversarial training approach with a min-max game. This dynamic method helps better emphasize the complex nature of scanpaths.
- We condition the learning on the probabilistic distribution of all users, forcing the network to distill the subjective properties of observer population.
- We prove the validity and competitiveness of the proposed method through testing of our model on 2 large datasets.

In the rest of this paper, Section 2 describes the proposed method in detail along with training details. In Section 3, we present the experimental protocol as well as discuss the quantitative and qualitative results obtained. Section 4 ends the paper with conclusions.

2. PROPOSED METHOD

To solve the challenges mentioned in Section 1, we designed an adversarial training architecture with a thought-out fully convolutional generator model and a discriminator model that is used as a dynamic progressive loss. This later refines the predictive ability of the generator during training by improving its own discriminative ability. This section presents the proposed models (i.e. generator and discriminator) and the training strategies applied.

2.1. Generator Architecture

The proposed model utilizes lightweight components to predict scanpaths with variable lengths. Fig. 1 illustrates the overall architecture of our generator model, which takes an input image and generates a scanpath. To encode the input into a different representational space, we use a pre-trained MobileNet network as a lightweight feature extractor. To enhance the representational ability of our model related to our downstream task, we introduce the use of domain-specific priors through a learnable set of spatially Gaussian distributions, which is a generalization of the "Central bias" theory for visual attention [21]. We model these priors using Eqs. 1 and 2, where $\mu_{x,y}$, $\sigma_{x,y}$ and S represent the mean of the distribution, the standard deviation and the set of Gaussian priors, respectively.

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

$$S = \{G_1(x, y), G_2(x, y), \dots, G_{16}(x, y)\} \quad (2)$$

In this study, we modeled 16 different Gaussian priors, each with two parameters. The information contained in the set S is then integrated with the features resulting from MobileNet through concatenation followed by a 2D convolution. As we can consider the scanpaths as an ordinal sequence, we added a positional encoding feature and use a 1D convolutional-based architecture to predict the succession of fixations. More precisely, we used 2 branches of 1D convolutions in order to disentangle the representations of the multi-variable sequence (i.e. the two spatial dimensions).

2.2. Discriminator Architecture

The second component of the architectural setup is the discriminator network, illustrated in Fig. 2. Its purpose is to discriminate between the distributions of the ground-truth scanpaths and generated ones. During the training, this model enhances its ability to represent the ground-truth scanpath distribution, acting thus as a gradually improving and dynamic loss function for the scanpath generator model. Inspired by the generator performance, we separated the sequences representing the coordinates into two different spatial dimensions, enabling the disentanglement of the features of the two dimensions. Each branch of the discriminator model consists of a succession of 1D convolutions activated by a Leaky ReLU function with a slope of 0.2. The extracted features are gradually increased in proportion to the depth of the network. At the end of each branch, a global Max Pooling Layer is employed on each of the feature vectors. The resulting vectors are then concatenated to build a global representation of the scanpath. Finally, we employed three fully connected layers for discriminating the features and thus classifying the scanpaths.

2.3. Adversarial Training

In order to model with greater accuracy and maintain consistency of the predicted scanpath with multiple users, we opted to use the min-max game between the classifier discriminator

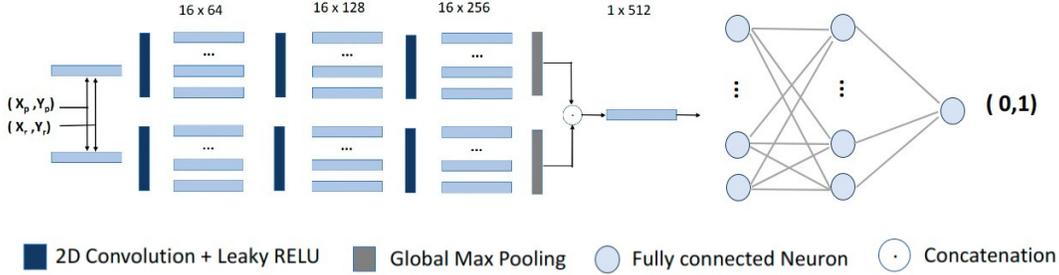


Fig. 2. Discriminator model architecture.

and the predictive generator networks, represented through the following equation:

$$\min_G \max_D V(G, D) = \mathbb{E}_{\hat{y} \sim p_{\text{data}}(y)} [\log D(\hat{y}|p(y))] + \mathbb{E}_{x \sim p_x(x)} [1 - \log D(G(x|p(y)))] \quad (3)$$

where x represents the input image, $p(y)$ is the distribution of different random ground truth scanpaths from multiple users changed periodically during the training, and \hat{y} represents the predicted scanpath. G and D are the generator and discriminator models, respectively.

This training approach aims to reduce the distance between the predicted scanpath and the whole set of users, allowing to the integration of the cognitive biases of multiple viewers while maintaining good qualitative shapes for the scanpath. Therefore, this forces the network to learn a non-user-specific representation of the perceptual function. The model was trained for 246 epochs with a learning rate equal to 10^{-5} .

3. EVALUATION

3.1. Datasets

We evaluated our method on two widely-used datasets, namely **Salicon** [22] and **MIT1003** [23]. **Salicon** [22] consists of 9000 images for training, 1000 images for validation, and 5000 images for testing with the corresponding saliency maps and scanpaths data for all users.

MIT1003 [23] is usually presented along with the MIT300 dataset [24] benchmark. It consists of 1003 natural scene images with the corresponding saliency maps and fixation points gathered throughout eye-tracking sessions. Each image has 15 observers, resulting in 15045 scanpaths.

3.2. Experimental Protocol

In our work, we test on the 5000 images on Salicon dataset with approximately 250000 scanpaths, which ensures the empirical soundness of the results. It is worth noting that in this study our model was trained only on the training set of this dataset. We then used the entire MIT1003 dataset for testing in a cross-dataset evaluation manner without fine-tuning our

model on this dataset. Similarly to the first dataset, the significant number of scanpaths ensures empirically sound results.

To evaluate the performance of our method, we employed three commonly used metrics: MultiMatch, NSS, and Congruency. *MultiMatch* (MM) metric [25] compares the similarity of two vectors using five characteristics (Shape, Direction, Length, Position, and Duration). Since the model predicts only spatial coordinates, we use only the first four characteristics and measure the overall performance with their mean value. Two hybrid metrics that compare the predicted scanpaths with a general saliency map for a given image are also employed: *NSS* and *Congruency*. *NSS* calculates the mean saliency value of the scanpath fixation locations over the ground truth saliency map, while *Congruency* computes the ratio of the predicted fixation points which are in the salient regions after thresholding and binarizing the ground truth saliency map. The hybrid metrics allow to measure the accordance and consistency between the predicted scanpath and the users.

3.3. Quantitative Results

Tables 1 and 2 show the results obtained after testing our model according to the protocol described in Section 3.2 on the Salicon and MIT1003 datasets, respectively. The performance reached by our method is compared to state-of-the-art methods.

The results achieved on Salicon (see Table 1) show that our model outperformed state-of-the-art methods on the shape and length components of the *multimatch* metric. More precisely, we notice a significant improvement in the shape and length components, while the overall mean *multimatch* shows an improvement compared with the other models. We also achieved the top results for the *congruency* metric, which indicates that predicted fixations are mostly located in salient regions, maintaining thus a certain consistency with the distribution of users on the Salicon dataset. This is further emphasized and supported by the state-of-the-art results obtained on other metrics.

As we tested the model on the MIT1003[23] dataset without any kind of fine-tuning, the results obtained in Table 2 show a natural decrease compared to those obtained on Salicon. Nonetheless, the results achieved are still quite high

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan[15]	0.9608	0.5698	0.9530	0.8172	0.8252	-0.2904	0.0825
Le Meur[14]	0.9505	0.6231	0.9488	0.8605	0.8457	0.8780	0.4784
G-Eymol[16]	0.9338	0.6271	0.9521	0.8967	0.8524	0.8727	0.3449
SALYPATH [18]	0.9659	0.6275	0.9521	0.8965	0.8605	0.3472	0.4572
our model (Adversarial)	0.9745	0.6246	0.9642	0.8892	0.8631	0.9762	0.5226

Table 1. Results of scanpath prediction on Salicon.

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan[15]	0.9237	0.5630	0.8929	0.8124	0.7561	-0.2750	0.0209
DCSM (VGG)[17]	0.8720	0.6420	0.8730	0.8160	0.8007	-	-
DCSM (ResNet)[17]	0.8780	0.5890	0.8580	0.8220	0.7868	-	-
Le Meur[14]	0.9241	0.6378	0.9171	0.7749	0.8135	0.8508	0.1974
G-Eymol[16]	0.8885	0.5954	0.8580	0.7800	0.7805	0.8700	0.1105
SALYPATH [18]	0.9363	0.6507	0.9046	0.7983	0.8225	0.1595	0.0916
our model	0.9614	0.6529	0.9423	0.7862	0.8357	0.7523	0.1797

Table 2. Cross dataset evaluation: Results of scanpath prediction on MIT1003.

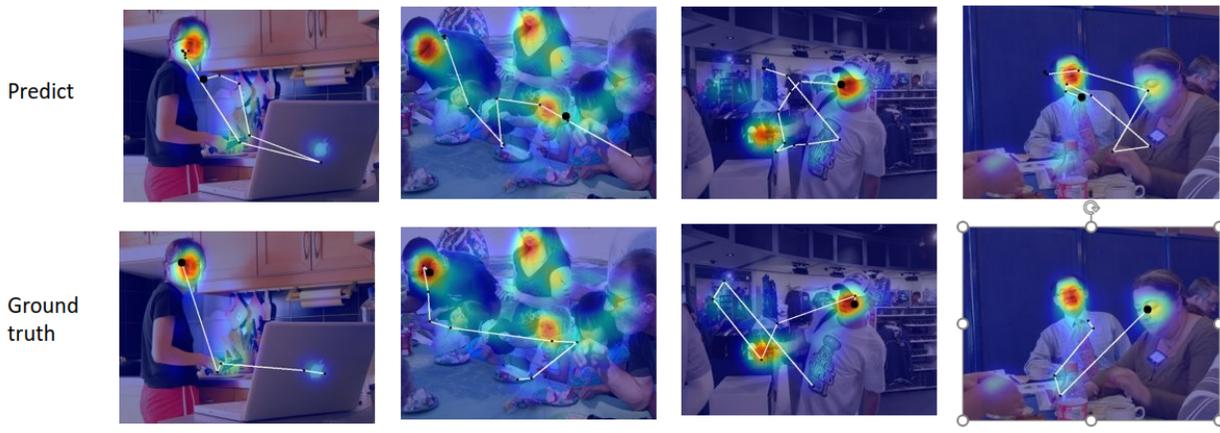


Fig. 3. Visualisation results.

and competitive with the state-of-the-art. This shows the generalization ability of our approach to data distributions coming from different sources. This is especially true, knowing that some of the comparative models trained on the MIT1003 dataset like DCSM [17] and Le Meur[14]. The results show a large improvement in the shape and length components of *multimatch* compared to other models, while maintaining competitive results for direction. The overall result is competitive compared to the other models. The results on the hybrid metrics (i.e. *NSS* and *congruency*) maintain close margins with the state-of-the-art since our model was not train on any subset of MIT1003, which has a different distribution of observers. We can also notice that Le Meur[14] and G-Eymol[16] models were able to maintain a slightly better performance because they rely on a saliency map generation step before sampling the scanpath.

3.4. Qualitative results

Fig. 3 depicts some qualitative results of predicted scanpaths (i.e. in the center) compared to ground truth scanpaths (i.e.

on both sides). Our model’s predictions show high fidelity to the original scanpaths while maintaining consistency between scanpaths originating from different users, making each predicted scanpath highly plausible.

4. CONCLUSION

In this paper, we introduced an adversarial training method that uses a discriminative network as a dynamic loss for gradually improving the representative ability of our model, while maintaining inter-observer consistency originating from the subjective nature of scanpaths. We tested our model on the two most used datasets for visual attention modeling and achieved outstanding competitive results on several hybrid and vector-based metrics. The qualitative results showed that our method succeeded to emulate scanpath obtained in the real world. This confirms that substituting traditional loss functions with adversarial training methods would yield better results for complex tasks of perception and attention.

5. REFERENCES

- [1] Laurent Itti, Geraint Rees, and John K Tsotsos, *Neurobiology of attention*, Elsevier, 2005.
- [2] 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*, 2014, pp. 149–152.
- [3] S. Jia and Y. Zhang, “Saliency-based deep convolutional neural network for no-reference image quality assessment,” *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14859–14872, 2018.
- [4] El Hassouni M. et al. AbouelazizI., Chetouani A., “v,” in *Neural Comput & Applic*, 2020, vol. 32, p. 16589–16603.
- [5] Aladine Chetouani, “Convolutional neural network and saliency selection for blind image quality assessment,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2835–2839.
- [6] 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*, vol. 10, no. 2, 2019.
- [7] 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*, 2021, pp. 227–236.
- [8] Marouane Tliba, Mohamed A. Kerkouri, Bashir Ghariba, Aladine Chetouani, Arzu Çöltekin, Mohamed Sami Shehata, and Alessandro Bruno, “Satsal: A multi-level self-attention based architecture for visual saliency prediction,” *IEEE Access*, vol. 10, pp. 20701–20713, 2022.
- [9] Haoxiang Wang, Zhihui Li, Yang Li, Brij B Gupta, and Chang Choi, “Visual saliency guided complex image retrieval,” *Pattern Recognition Letters*, vol. 130, pp. 64–72, 2020.
- [10] Alessandro Bruno, Marouane Tliba, Mohamed Amine Kerkouri, Aladine Chetouani, Carlo Calogero Giunta, and Arzu Çöltekin, “Detecting colour vision deficiencies via webcam-based eye-tracking: A case study,” in *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2023, ETRA ’23, Association for Computing Machinery.
- [11] 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*, pp. 1–1, 2022.
- [12] Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, and Mohamed Sayeh, “Salypath360: Saliency and scanpath prediction framework for omnidirectional images,” *Electronic Imaging*, vol. 34, no. 11, pp. 168–1–168–1, 2022.
- [13] Laurent Itti and Christof Koch, “Computational modelling of visual attention,” *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
- [14] Olivier Le Meur and Zhi Liu, “Saccadic model of eye movements for free-viewing condition,” *Vision research*, vol. 116, pp. 152–164, 2015.
- [15] 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)*, 2018, pp. 0–0.
- [16] Dario Zanca, Stefano Melacci, and Marco Gori, “Gravitational laws of focus of attention,” *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [17] Wentao Bao and Zhenzhong Chen, “Human scanpath prediction based on deep convolutional saccadic model,” *Neurocomputing*, 2020.
- [18] 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)*, 2021, pp. 1464–1468.
- [19] Marouane Tliba, Mohamed Amine Kerkouri, Aladine Chetouani, and Alessandro Bruno, “Self supervised scanpath prediction framework for painting images,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1538–1547.
- [20] Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, and Alessandro Bruno, “A domain adaptive deep learning solution for scanpath prediction of paintings,” in *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, New York, NY, USA, 2022, CBMI ’22, p. 57–63, Association for Computing Machinery.
- [21] 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.
- [22] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, “Salicon: Saliency in context,” in *CVPR*. 2015, pp. 1072–1080, IEEE Computer Society.
- [23] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, “Learning to predict where humans look,” in *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [24] Tilke Judd, Frédo Durand, and Antonio Torralba, “A benchmark of computational models of saliency to predict human fixations,” in *MIT Technical Report*, 2012.
- [25] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist, “It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach,” *Behavior research methods*, vol. 44, no. 4, pp. 1079–1100, 2012.