PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks

Marc Assens[1(✉)], Xavier Giro-i-Nieto[2(✉)], Kevin McGuinness[1(✉)], and Noel E. O'Connor[1(✉)]

[1] Dublin City University, Glasnevin, Whitehall, Dublin 9, Ireland
kevin.mcguinness@insight-centre.org
[2] Universitat Politecnica de Catalunya, 08034 Barcelona, Catalonia, Spain
xavier.giro@upc.edu

Abstract. We introduce PathGAN, a deep neural network for visual scanpath prediction trained on adversarial examples. A visual scanpath is defined as the sequence of fixation points over an image defined by a human observer with its gaze. PathGAN is composed of two parts, the generator and the discriminator. Both parts extract features from images using off-the-shelf networks, and train recurrent layers to generate or discriminate scanpaths accordingly. In scanpath prediction, the stochastic nature of the data makes it very difficult to generate realistic predictions using supervised learning strategies, but we adopt adversarial training as a suitable alternative. Our experiments prove how PathGAN improves the state of the art of visual scanpath prediction on the iSUN and Salient360! datasets.

Keywords: Saliency · Scanpath · Adversarial training · GAN · cGAN

1 Introduction

When a human observer looks at an image, he spends most of his time looking at specific regions [1,2]. He starts directing his gaze at a specific point and explores the image creating a sequence of fixation points that covers the salient areas of the image. This process can be seen as a resource allocation problem; our visual system decides where to direct its attention, in which order, and how much time will be spent in each location given an image.

Visual saliency prediction is the field of computer vision that focuses on estimating the image regions that attract human attention. The understanding of this process can provide clues on human image understanding, and has applications in domains such as image and video compression, transmission, and rendering. In order to train and evaluate saliency prediction models, there exist scientific datasets containing fixation points generated by human observers when exploring an image without any specific task in mind. They are traditionally captured with eye-trackers [3], mouse clicks [4], and webcams [5].

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11133, pp. 406–422, 2019.
https://doi.org/10.1007/978-3-030-11021-5_25

These fixation points have an important characteristic: stochasticity [6]. Different human observers can produce very different fixation points. Thus, researchers in the field of saliency prediction have traditionally aggregated fixations of multiple observers to generate a consistent representation called saliency map [7]. A saliency map is a single channel image obtained by convolving a Gaussian kernel with each fixation. The result is a gray-scale heatmap that represents the probability of each pixel in an image being fixated by a human, and it is usually used as a soft-attention guide for other computer vision tasks.

Because fixations are aggregated over the temporal dimension, the saliency map representation loses all the temporal information. Thus, information like *the parts of an image that are being fixated first* is not retained. Recent studies have shown some of the limitations of saliency maps and have raised the need for a representation that is also temporally-aware [8]. In some situations saliency maps fail to represent the relative importance of the different parts of an image, giving more relevance to small regions with text where humans spend a long time reading. We believe that the regions where a human first fixates might be more relevant, therefore they should have more weight in a soft-attention representation. Another argument that favors temporally-aware saliency representations is the recent explosion of Virtual Reality technologies. It has brought new challenges regarding the usage of omni directional images (360-degree images), and it seems that solutions will depend on the use of temporal information.

Thus, there is an increasing demand for temporally-aware saliency representations such as scanpaths, and algorithms that are capable of working with them. Scanpaths as a temporally-aware saliency representation have received recent attention [9,10] and different datasets are available today.

Previous work on scanpath prediction shows that there are difficulties when working with very stochastic data [6]. One of the problems that has been found is that supervised learning algorithms using the MSE loss do not perform well for this task because the final prediction tends to be the average of all the possible predictions [11]. When predicting scanpaths, the average prediction tends to be always in the center. Recently, Goodfellow et al. [12] proposed a framework to create generative functions via an adversarial process, in which two models are trained simultaneously: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample comes from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This process allows models to generate realistic predictions even when the data has very complicated distributions. This framework seems a suitable technique for the generation of realistic scan paths.

This paper explores an end-to-end solution for omni directional scanpath prediction using conditional adversarial training. We show that this framework is suitable for this task and it significantly improves the performance. Our results achieve state-of-the-art performance using a convolutional-recurrent architecture, whose parameters are refined with a discriminator.

This paper is structured as follows. Section 2 reviews the state-of-the-art models for visual saliency prediction and recent advances on conditional adversarial networks. Section 3 presents PathGAN, our deep convolutional-recurrent neural network, as well as the discriminator network used during the adversarial training. Section 4 describes the training procedure and the loss functions used. Section 5 includes the experiments and results of the described techniques. Finally, Sect. 6 discusses the main conclusions and future work. Our results can be reproduced with the source code and trained models available at https://github.com/imatge-upc/pathgan.

2 Related Work

2.1 Visual Saliency Prediction

Saliency Maps. Saliency prediction has received interest by the research community for many years. Thus seminal works by Itti et al. [7] proposed considering low-level features at multiple scales and combining them to form a two-dimensional saliency map. Harel et al. [13], also starting from low-level feature maps, introduced a graph-based saliency model that defines Markov chains over various image maps, and treat the equilibrium distribution over map locations as activation and saliency values. Judd et al. in [14] presented a bottom-up, top-down model of saliency based not only on low but mid and high-level image features. Borji [15] combined low-level features saliency maps of previous best bottom-up models with top-down cognitive visual features and learned a direct mapping from those features to eye fixations.

As in many other fields in computer vision, a number of deep learning solutions have very recently been proposed that significantly improve the performance. For example, the Ensemble of Deep Networks (eDN) [16] represented an early architecture that automatically learns the representations for saliency prediction, blending feature maps from different layers. Their network might be consider a shallow network given the number of layers. In [17] shallow and deeper networks were compared. DCNN have shown better results even when pre-trained with datasets build for other purposes. DeepGaze [18] provided a deeper network using the well-know AlexNet [19], with pre-trained weights on Imagenet [20] and with a readout network on top whose inputs consisted of some layer outputs of AlexNet. The output of the network is blurred, center biased and converted to a probability distribution using a softmax. Huang et al. [21], in the so call SALICON net, obtained better results by using VGG rather than AlexNet or GoogleNet [22]. In their proposal they considered two networks with fine and coarse inputs, whose feature maps outputs are concatenated.

Li et al. [23] proposed a multi resolution convolutional neural network that is trained from image regions centered on fixation and non-fixation locations over multiple resolutions. Diverse top-down visual features can be learned in higher layers and bottom-up visual saliency can also be inferred by combining information over multiple resolutions. Cornia et al. [24] proposed an architecture that combines features extracted at different levels of a DCNN. They introduced a

loss function inspired by three objectives: to measure similarity with the ground truth, to keep invariance of predictive maps to their maximum and to give importance to pixels with high ground truth fixation probability. In fact choosing an appropriate loss function has become an issue that can lead to improved results. Thus, another interesting contribution of Huang et al. [21] lies on minimizing loss functions based on metrics that are differentiable, such as NSS, CC, SIM and KL divergence to train the network (see [25,26] for the definition of these metrics. A thorough comparison of metrics can be found in [27]). In Huang's work [21] KL divergence gave the best results. Jetley et al. [28] also tested loss functions based on probability distances, being the Bhattacharyya distance the one that provided the best results.

Scanpaths. The literature on the related task of scanpath prediction is much smaller, but has received recent attention caused by the rise of VR and AR technologies [29]. In [9], Cerf et al. concluded that human observers – when not instructed to look for anything in particular – tend to fixate on a human face within the first two fixations with a probability over 80%. Moreover, the consistency of scanpaths increases when faces are present. Hu et al. [30] introduced a model that predicts relevant areas of a 360-degree video and decides in which direction a human observer should look for each frame. Some authors have also focused on omni-directional images [29,31,32].

SalTiNet [6] proposed a deep learning approach that proposes a novel three-dimensional representation of saliency maps: the *saliency volumes*. This data structure captured the temporal location of the fixation across an additional temporal axis added to the classic saliency maps. The final scanpath are generated by sampling fixation points from this saliency volumes and finally introducing a post-filtering stage. PathGAN also uses a deep neural model, but provides a fully end-to-end solution where the model directly generates a scanpath, with no need of any sampling nor post-processing.

2.2 Generative Adversarial Networks

The generation of a sequence of fixation points over an image with a Recurrent Neural Network (RNN) had been previously attempted in [6]. The authors trained a RNN to minimize the L^2 loss between predicted and ground truth scanpaths, but the resulting model tended to predict output fixations always in the center, as this is the best option on average for that loss function. Similar problems have been observed in other image prediction problems (e.g. *pix2pix*), where blurred images where output as a result [11,33,34].

The generation of diverse and realistic new data samples has received a lot of interest thanks to the work of Ian Goodfellow et al. on Generative Adversarial Networks (GANs) [12]. In this framework, two models are trained iteratively. First, the generative model G tries to capture the data distribution. Second, the discriminator model D estimates the probability that a given sample is synthesized or real. During training, G tries to maximize the probability of fooling D. This process can also be seen as if GANs learn a loss function to tell if a sample

is real or fake. Generated samples that are not realistic (e.g. blurry images, or scanpaths with all the fixations in the center) will not be tolerated.

A popular variation of GANs are the Conditional Adversarial Networks (cGANs) [35], where G does not output a sample purely from a noise vector, but it is also conditioned on a given input vector. In this setting, D needs to observe the conditioning vector to decide about the nature of the sample to be classified into synthesized or real. There have been multiple variations around the cGAN paradigm. Isola et al. [36] proposed cGANs as a general purpose solution for image-to-image translation tasks using a *U-Net* [37] architecture for the generator, and a convolutional *PatchGAN* [38] architecture for the discriminator. Reed et al. bridge recent advances in the image and text fields and propose a GAN architecture that is capable of generating plausible images of birds and flowers from detailed text descriptions [39]. Mirza et al. conditioned GANs to discrete labels in order to generate MNIST digits conditioned on class labels [40]. Gauthier et al. generates faces with specific attributes by varying the conditional information provided to the network [41].

In our work, we adopt the cGAN paradigm to overcome the limitation reported in [6] when trying to use a RNN for visual scanpath prediction. This way, PathGAN proposes to train a RNN following an adversarial approach, in such a way that the resulting generator produces realistic and diverse scanpaths conditioned to the input image.

3 Architecture

The overall architecture of PathGAN is depicted in Fig. 1. It is composed by two deep neural networks, the generator and the discriminator, whose combined efforts aim at predicting a realistic scanpath from a given image. The model is trained following the cGAN framework to allow the predictions to be conditioned to an input image, encoded by a pre-trained convolutional neural network. This section provides details about the structure of both networks and the considered loss functions.

3.1 Objective

GANs are generative functions that learn a transformation from random noise vectors z to output vectors y, $G : z \rightarrow y$ [12]. Conditional GANs learn a transformation from a given input vector x and random noise vector z, to y, $G : x, z \rightarrow y$. Therefore, the objective function of cGANs can be expressed as:

$$L_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))], \qquad (1)$$

where the generator tries to minimize the loss, while the discriminator tries to maximize it.

Multi-objective Loss Functions. Previous works has found useful to mix the GAN's loss function with another traditional loss such as the Euclidean

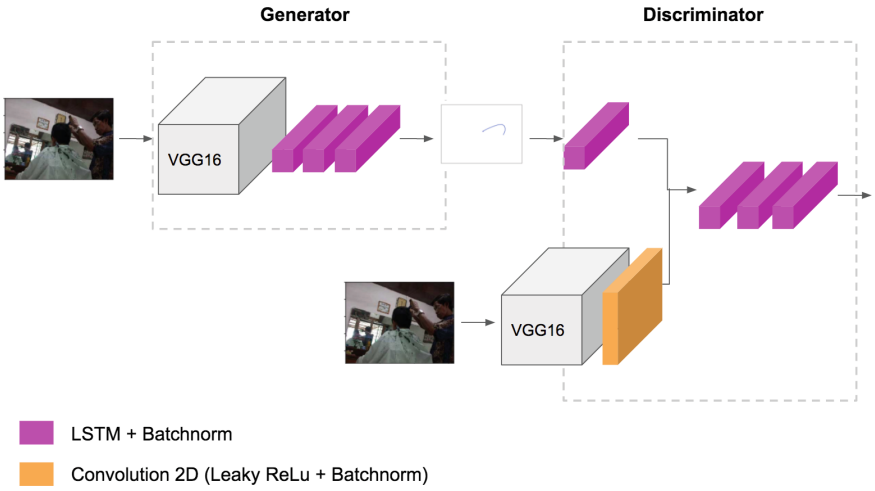


Fig. 1. Overall architecture of the proposed convolutional-recurrent model

distance [33]. In this case, the task of the discriminator remains unchanged, but the generator is forced to output samples that are close to the ground truth (in terms of L^2 distance). We found that this setting improved stability and convergence rate of the adversarial training. As it will be explained in the next section, each prediction of our model contains four dimensions. The L^2 distance is computed using all four dimensions. We called this parameter *content loss*, and it is defined as:

$$L_{L^2}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|^2]. \tag{2}$$

The final formulation of the loss function for the generator during adversarial training is:

$$L = L_{\text{cGAN}}(G,D) + \alpha L_{L^2}(G). \tag{3}$$

In Eq. 1, $(1-D(G(x,z))$ represents the probability of the generator fooling the discriminator. Thus, we expect the loss to decrease as the chances of fooling the discriminator increase. In our experiments we used the hyperparameter $\alpha = 0.05$. It is also important to note that z plays an important role making the output of the generator non-deterministic [42]. During the training of the discriminator the content loss is not used.

3.2 Generator

The generator reads images as input and outputs a variable length sequence of predicted fixation points. In addition to the coordinates of the fixation points, our model has an end-of-sequence (EOS) neuron to encode the scanpath variable length behavior. This neuron has values between $[0, 1]$ and represents the probability of having reached the end of the sequence. Thus, each prediction

of our model contains a fixation point (composed by a spatial coordinate and a timestamp) and an EOS parameter $[x, y, t, \text{EOS}]$. At training time, we train on fixations of a scanpath until we reach the EOS, and at test time we predict scanpath fixations until we reach the EOS.

We propose a convolutional-recurrent architecture that learns its filter parameters to predict scanpaths. Figure 1 illustrates the architecture of the model, composed of 49 million free parameters. The generator is composed of two parts. First, high-level image features are extracted using a convolutional neural network for image recognition named VGG16 [43] pre-trained on the ImageNet dataset [20]. Then, resampling of the VGG16 activations is performed with an Average Pooling layer to a fixed size representation. This allows the usage of this model with different image sizes and different types of datasets. Finally, a recurrent module composed of 3 fully connected LSTMs with tanh activation and 1,000 hidden units is used to generate a variable length scanpath. Batch normalization layers are placed after each recurrent layer to improve convergence and accelerate learning.

3.3 Discriminator

Figure 1 also shows the architecture and layer configuration of the discriminator. This network predicts if a given scanpath is synthesized or not, and this decision is conditioned to the associated image.

It is clear that knowledge of the image that a scanpath corresponds to is essential to evaluate quality. Moreover, previous work has shown that conditioning the discriminator function to the input significantly increases the performance, sometimes preventing the generation from collapse [36]. In our architecture, the discriminator has two input branches; a branch where a scanpath is read, and a branch where the image is read. This allows discriminating whether a scanpath is realistic for a given image. The features of the two branches are concatenated.

Briefly, the discriminator function is based on a recurrent architecture where the scanpath fixations are read sequentially. The network is composed of a VGG16 module that extracts image features, and three recurrent layers interspersed with batch normalization layers. The recurrent layers contain 1000 hidden units and a tanh activation. Similarly to the generator, the VGG16 activations are resampled with an Average Pooling layer to a fixed size representation. The recurrent layers all use *tanh* activations, with the exception of the final layer, which makes use of a sigmoid activation.

4 Training

The weights of the model have been learned with an objective function that combines an adversarial loss and a content loss [36]. The content loss follows a simple approach in which the generated and ground truth fixation points are compared using the L^2 norm (or mean square error). The adversarial loss depends on the probability of the generator fooling the discriminator.

We trained the PathGAN architecture on two datasets. First, the network was trained using the iSUN dataset, which contains 6,000 training images. Then, the filter weights were fine-tuned on omni directional images using the Salient360 dataset, which has 40 training images. For validation purposes, we split the training data into 80% for training and the rest for validation. Notice that for each gradient update a single scanpath is used.

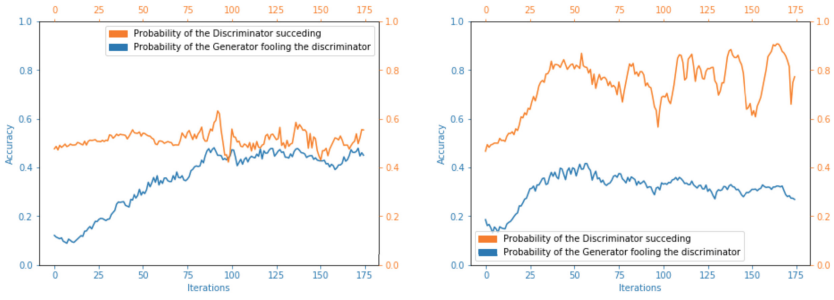
The spatial positions of the fixations were normalized to $[0, 1]$. Moreover, when training on the Salient360 dataset, input images were downsampled to fit the dimensions of 300×600 prior to training. We also subtracted the mean pixel value of the training set from the image's pixels to zero center them.

The architecture was trained using the *RMSprop* optimizer with the following settings: $lr = 10^{-4}$, $\rho = 0.9$, $\epsilon = 10^{-8}$ and without decay.

Our network took approximately 72 h to train on six NVIDIA Tesla K80 GPU using the Keras framework with Tensorflow backend. At test time it generates approximately 4 scanpaths per second. Figure 2 shows the evolution of the validation set accuracy during the adversarial training.

Our networks train on a minibatch size of $m = 100$, and after trying various combinations, we settled on the generator doing 8 gradient updates, while the discriminator does 16 for each iteration. At train time, the generator is first bootstrapped by training only on the content loss for a duration of 5 epoch. Then, the adversarial training begins.

This architecture was designed considering the amount of training data available, and multiple strategies were introduced to prevent overfitting. In the first place, the convolutional modules initialized from the VGG16 model were not fine-tunned, decreasing the number of training parameters. Second, the input images were resized to a smaller dimension, and dropout noise was introduced at training time. We also used dropout noise ($p = 0.1$) on the recurrent layers. With the objective of increasing variance of the generated scanpaths, Gaussian noise ($\sigma = 3, \mu = 0$) was added to the input images at prediction time. The added noise caused a very small perceptual difference on the images.



(a) Training with adversarial and content loss

(b) Training only with adversarial loss

Fig. 2. iSUN validation set accuracies for training with GAN+MSE vs GAN on varying number of epochs.

5 Experiments

PathGAN was assessed and compared from different perspectives. First, we evaluated the performance on traditional images using the iSUN dataset. Second, we show quantitative performance results on omni directional images using the Salient360 dataset.

5.1 Datasets

The network was initially trained on the iSUN dataset [44] that contains 6,000 training images, and its performance is benchmarked in Sect. 5.3. Then, the network was fine-tuned to predict scanpaths on omni directional images using the Salient360 dataset, which contains 60 training images with data obtained from head and eye movements from the human observers.

It is worth noticing that our use of omni directional images in this network implies an important simplification. We assume that omni-directional images are similar to traditional flat images, just with a bigger size. This presents advantages like being able to reuse the same architecture, and easily fine-tune it, and this strategy has been previously successful [6]. Nevertheless, it neglects the characteristic of omni directional images where points that are close to opposite corners are spatially close.

5.2 Metrics

The similarity metric used in the experiments is the Jarodzka algorithm [45]. This metric presents different advantages over other common metrics like the Levenshtein distance or correlating attention maps. In the first place, it preserves the overall shape, direction and amplitude of the saccades, the position and duration of the fixations. Second, it provides more detailed information on the type of similarity between two vectors. This metric has been recently used in the Salient360, scanpath prediction challenge at ICME 2017 [10]. The implementation of the metric for omni directional images was released by the University of Nantes [46]. This code was adapted to compute the Jarodzka metric for conventional images on the iSUN dataset.

The ground truth and predicted scanpaths are then matched 1-to-1 using the Hungarian algorithm to obtain the obtain the minimum cost. The presented results compare the similarity of 40 generated scanpaths with scanpaths in the ground truth.

5.3 Results

Comparison with State-of-the-Art. PathGAN is compared using the iSUN and Salient360! datasets. Table 1 compares the performance on omni directional images using the Jarodzka metric, against other solutions presented at the Salient360! Challenge [10], which took place at the IEEE ICME 2017 conference

in Hong Kong. The results of the participants were calculated by the organization, on a test set whose ground truth was not public at the time. Although at the time of writing this test set is public, our model has only been trained on the training set. These results indicate the superior performance of PathGAN with respect to the participants.

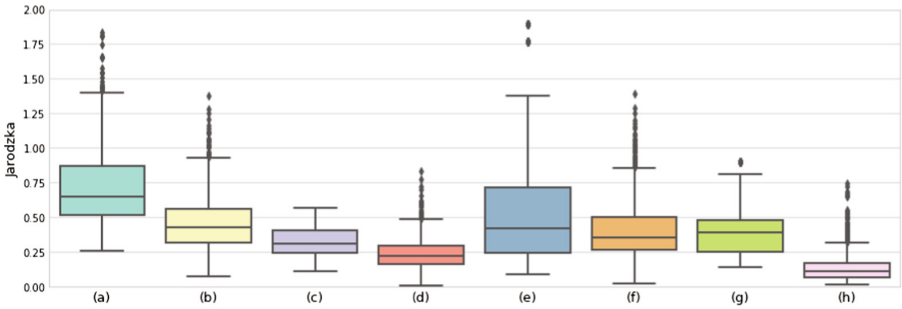
Table 1. Comparison with the best submissions to the ICME 2017 Salient360! Lower values are better.

	Wuhan University	SJTU	SaltiNet	**PathGAN**
Jarodzka ↓	5.9517	4.6565	2.8697	**0.74**

Figure 3 compares the performance of PathGAN with different baselines and another state-of-the-art model on the iSUN dataset. To accurately test the performance of the best scanpath prediction model of the Salient360! Challenge 2017 on the iSUN dataset, we fine-tuned it. Figure 4 illustrates how the Jarodzka performance of PathGAN evolves during training.

id		Jarodzka↓
a	Random positions and number of fixations	0.71
b	Random positions and GT number of fixations	0.45
c	Sampling ground truth saliency maps	0.31
d	Interchanging scanpaths across images	0.23
e	SalTiNet	0.69
f	PathGAN without content loss	0.42
g	SalTiNet (fine-tuned on iSUN)	0.40
h	**PathGAN**	**0.13**

(a) Mean performance on iSUN with the Jarodzka metric



(b) Distribution of results obtained for each model

Fig. 3. Comparison on iSUN between the state-of-the-art and baselines. The distribution of results and the mean performance are depicted. Lower values are better.

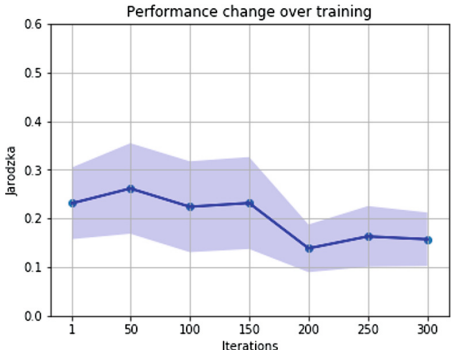


Fig. 4. iSUN validation set Jarodzka evaluation on varying number of mini-batches.

Content-Loss Gain. The performance gain that comes with the use of a content-loss based on MSE was analyzed from different perspectives. Figure 2 shows that the *content loss* (mentioned in Sect. 3.1) significantly improves convergence. In our experiments, we have not been able to achieve convergence without using the MSE loss. Figure 3 illustrates that these improvements are also reflected in the Jarodzka metric.

Qualitative Results. Our model's performance has also been explored from a qualitative perspective by observing the generated scanpaths on the iSUN dataset and on the Salient360! dataset (Figs. 6 and 7). Notice the diversity of results given the generative nature of the model, based on the drop out ratio in the LSTM.

Another way of assessing the behaviour of our model is by comparing the distributions of generated and ground truth fixations. Figure 5 compares the distribution of spatial locations where the model fixates on the iSUN's validation dataset. We observe that the model correctly finds a center-bias.

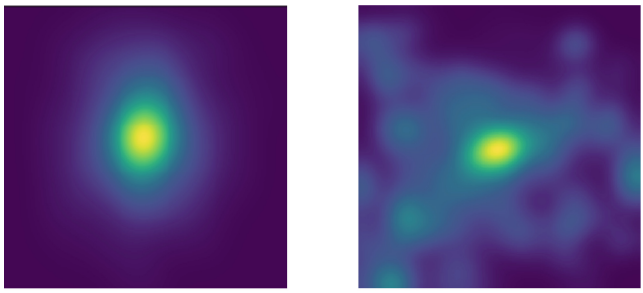


Fig. 5. Comparison of generated and ground truth spatial distribution of fixations

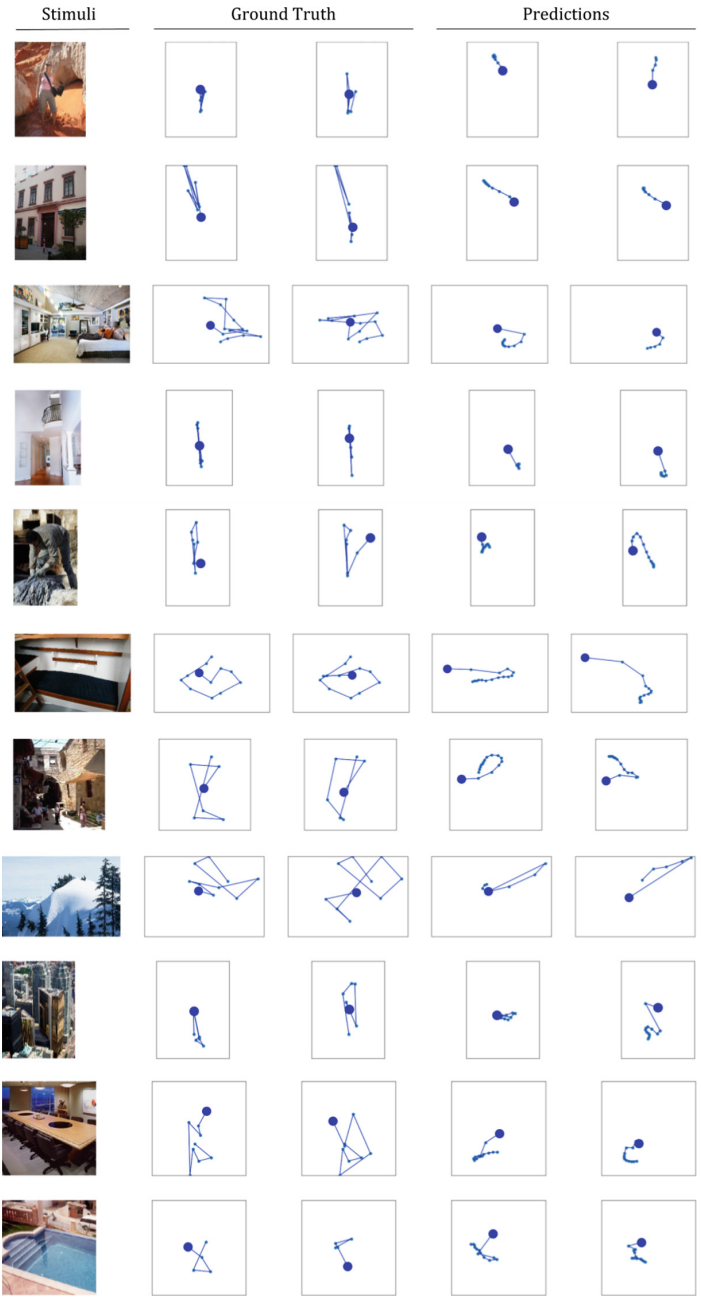


Fig. 6. Examples of predictions and ground truth on the iSUN dataset.

Fig. 7. Examples of predictions on the Salient360! dataset. The stimuli has the ground truth annotated.

6 Conclusions

Most of the work that has been done in the field of saliency estimation focuses on aggregating fixations from multiple observers and the prediction of saliency maps. Thus, it does not pay any attention to the temporal dimension of saliency estimation. This paper addressed a task that is closer to what a human does

when observing an image: *scan path prediction*. This task presents several challenges, such as the complicated distribution of the data, and we address them accordingly.

We presented PathGAN, an end-to-end model capable of predicting scanpaths on ordinary and omni-directional images using the framework of conditional adversarial networks. Our experiments show that this architecture achieves state-of-the-art results on both scenarios. Moreover, this model has the following desirable characteristics: (1) the probability of a fixation is conditioned to previous fixations; and (2) the length of the scanpath, the duration of each fixation, and the spatial position of the fixations are treated as conditioned random variables.

Finally, we want to note that the use of this model with omni-directional images assumes the simplification that an omni-directional image is similar to a traditional image, but with a larger size. While this presents advantages, it also has a drawback: it neglects the characteristic of omni-directional images where points that are close to opposite corners are spatially close.

Future work could aim to solve the issue mentioned above, or could try to include top-down task specific information during training. Our results can be reproduced with the source code and trained models available at https://github.com/imatge-upc/pathgan.

Acknowledgement. This research was partially supported by the Spanish Ministry of Economy and Competitivity and the European Regional Development Fund (ERDF) under contract TEC2016-75976-R. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/15/SIRG/3283. We acknowledge the support of NVIDIA Corporation for the donation of GPUs.

References

1. Porter, G., Troscianko, T., Gilchrist, I.D.: Effort during visual search and counting: insights from pupillometry. Q. J. Exp. Psychol. **60**, 211–229 (2007)
2. Amor, T.A., Reis, S.D., Campos, D., Herrmann, H.J., Andrade Jr., J.S.: Persistence in eye movement during visual search. Sci. Rep. **6**, 20815 (2016)
3. Wilming, N., et al.: An extensive dataset of eye movements during viewing of complex images. Sci. Data **4**, 160126 (2017)
4. Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1072–1080. IEEE (2015)
5. Krafka, K., et al.: Eye tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Assens, M., Giro-i Nieto, X., McGuinness, K., OConnor, N.E.: SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 2331–2338. IEEE (2017)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **20**(11), 1254–1259 (1998)

8. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 809–824. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_49
9. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Advances in Neural Information Processing Systems, pp. 241–248(2008)
10. University of Nantes, Technicolor: Salient360: Visual attention modeling for 360° images grand challenge (2017)
11. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
12. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
13. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Neural Information Processing Systems (NIPS) (2006)
14. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
15. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
16. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
17. Pan, J., Sayrol, E., Giró-i Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
18. Kümmerer, M., Theis, L., Bethge, M.: DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet. In: International Conference on Learning Representations (ICLR) (2015)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
21. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks. In: IEEE International Conference on Computer Vision (ICCV) (2015)
22. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
23. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
24. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: International Conference on Pattern Recognition (ICPR) (2016)
25. Riche, N.M.D., Mancas, M., Gosselin, B., Dutoit, T.: Saliency and human fixations. State-of-the-art and study comparison metrics. In: IEEE International Conference on Computer Vision (ICCV) (2013)
26. Kümmerer, M., Theis, L., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proc. Natl. Acad. Sci. (PNAS) **112**(52), 16054–16059 (2015)

27. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1610.01563 (2016)
28. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
29. Rai, Y., Le Callet, P., Guillotel, P.: Which saliency weighting for omni directional image quality assessment? In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2017)
30. Hu, H.N., Lin, Y.C., Liu, M.Y., Cheng, H.T., Chang, Y.J., Sun, M.: Deep 360 pilot: learning a deep agent for piloting through 360 sports video. In: CVPR, vol. 1, p. 3 (2017)
31. Zhu, Y., Zhai, G., Min, X.: The prediction of head and eye movement for 360 degree images. Signal Process. Image Commun. (2018)
32. Ling, J., Zhang, K., Zhang, Y., Yang, D., Chen, Z.: A saliency prediction model on 360 degree images using color dictionary based sparse representation. Signal Process. Image Commun. (2018)
33. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
34. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126 (2016)
35. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
36. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
37. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
38. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
39. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)
40. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
41. Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. Cl. Proj. Stanf. CS231N Convolutional Neural Netw. Vis. Recognit. Winter Semester **2014**(5), 2 (2014)
42. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 318–335. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_20
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)

45. Jarodzka, H., Holmqvist, K., Nyström, M.: A vector-based, multidimensional scanpath similarity measure. In: Proceedings of the 2010 Symposium on Eye-tracking Research & Applications, pp. 211–218. ACM (2010)
46. Gutiérrez, J., David, E., Rai, Y., Le Callet, P.: Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360° still images. Signal Process. Image Commun. (2018)