SaltiNet: Scan-path Prediction on 360 Degree Images using Saliency Volumes

Marc Assens*and Xavier Giro-i-Nieto
Image Processing Group
Universitat Politecnica de Catalunya (UPC)
Barcelona, Catalonia/Spain

xavier.giro@upc.edu

Kevin McGuinness and Noel E. O'Connor
Insight Center for Data Analytics
Dublin City University
Dublin, Ireland

kevin.mcguinness@insight-centre.org

Abstract

We introduce SaltiNet, a deep neural network for scan-path prediction trained on 360-degree images. The model is based on a temporal-aware novel representation of saliency information named the saliency volume. The first part of the network consists of a model trained to generate saliency volumes, whose parameters are fit by back-propagation computed from a binary cross entropy (BCE) loss over downsampled versions of the saliency volumes. Sampling strategies over these volumes are used to generate scan-paths over the 360-degree images. Our experiments show the advantages of using saliency volumes, and how they can be used for related tasks. Our source code and trained models available at https://github.com/massens/saliency-360salient-2017.

1. Motivation

Visual saliency prediction is a field in computer vision that aims to estimate the areas of an image that attract the attention of humans. This information can provide important clues to human image understanding. The data collected for this purpose are fixation points in an image, produced by a human observer that explores the image for a few seconds, and are traditionally captured with eye-trackers [30], mouse clicks [13] and webcams [15]. The fixations are usually aggregated and represented with a saliency map, 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.

Traditionally, saliency maps have only described fixation information with respect to the spatial layout of an image. This type of representations only encode the probability of
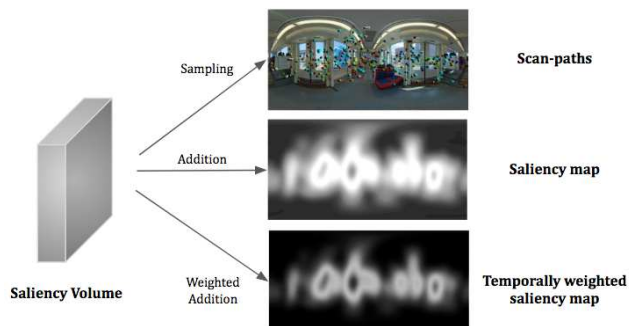

Figure 1: Scan-paths, saliency maps and temporally weighted saliency maps can be generated from a saliency volume.

each image pixel capture the visual attention of the user, but with no information regarding the order in which these pixels may be scanned or the duration of the fixation. Recent studies have raised the need for a representation that is also temporal-aware [3]. We address the temporal challenge for the particular case of 360° images, which contain the complete scene around the capture point and allow the viewer to choose the observation angle. Predicting the pattern that humans follow in 360° images is a topic of special interest for VR/AR applications, as it facilitates an efficient encoding and rendering on the display devices.

The main contributions of this paper are the following:

- the introduction of *saliency volumes* to capture the temporal nature of eye-gaze scan-paths;

- the SaltiNet architecture to generate scan-paths from a deep neural network that predicts saliency volumes and a sampling strategy over them;

- this work has been awarded as the best scanpath solution at the Salient360! challenge from the IEEE International Conference on Multimedia and Expo (ICME) 2017 [29].

---

*Work developed while Marc Assens was a visiting student at Insight Center for Data Analytics.

This paper is structured as follows. Section 2 reviews the related literature in saliency prediction for eye fixations and presents our work with respect to them. Section 3 presents the whole architecture of the system, and Section 4 describes how the deep neural network was trained. Section 5 describes the experiments and results to assess the performance of the model, while Section 6 draws the conclusions and future work.

2. Related Work

2.1. Saliency prediction

The first models for saliency prediction were biologically inspired and based on a bottom-up computational model that extracted low-level visual features such as intensity, color, orientation, texture and motion at multiple scales. Itti et al. [11] proposed a model that combines multiscale low-level features to create a saliency map. Harel et al. [8] presented a graph-based alternative that starts from low-level feature maps and creates Markov chains over various image maps, treating the equilibrium distribution over map locations as activation and saliency values.

Though this models did well qualitatively, the models had limited use because they frequently did not match actual human saccades from eye-tracking data. It seemed that humans not only base their attention on low-level features, but also on high-level semantics [3] (e.g., faces, humans, cars, etc.). Judd et al. introduced in [14] an approach that used low, mid and high-level image features to define salient locations. This features where used in combination with a linear support vector machine to train a saliency model. Borji [1] also combined low-level features with top-down cognitive visual features and learned a direct mapping to eye fixations using Regression, SVM and AdaBoost calssifiers.

Recently, the field of saliency prediction has made great progress due to advance of deep learning and its applications on the task of image classification [16] [32]. The advances suggest that these models are able to capture high-level features. As stated in [3], in March of 2016 there where six deep learning models among the top 10 results in the MIT300 saliency Benchmark [2].

The enormous amount of training data necessary to train these networks makes them difficult to train directly for saliency prediction. With the objective of allowing saliency models to capture this high-level features, some authors have adapted well-known models with good performance in the task of Image Recognition. DeepGaze [17] achieved state of the art performance by reusing the well-known AlexNet [16] pretrained on ImageNet [7] with a network on top that reads activations from the different layers of AlexNet. The output of the network is then blurred, center biased and converted to a probability distribution using a softmax. A second version called DeepGaze 2 [19] used features from VGG-19 [27] trained for image recognition. In this case, they did not fine-tune the network. Rather, some readout layers were trained on top of the VGG features to predict saliency with the SALICON dataset [13]. This results corroborated that deep features trained on object recognition provide a versatile feature space for performing related visual tasks. A complete new architecture designed and trained for saliency prediction was proposed in [24], but the same work also observed the benefits of using deeper pre-trained models for image classification as a basis. Other advances in deep learning such as generative adversarial training (GANs) and attentive mechanisms have also been applied to saliency prediction: SalGAN [23] is a deep network for saliency prediction that measured the gain in performance when using a universal adversarial training in opposite to optimizing for a specific loss function. The Saliency Attentive Model (SAM) [6] includes a Convolutional LSTM that focuses on the most salient regions of the image to iteratively refine the predicted saliency map.

In [28], Torralba et al. studied how the scene modules visual attention and discovered that the same objects recieve different attention depending on the scene where they appear (i.e. pedestrians are the most salient object in only 10% of the outdoor scene images, being less salient than many other objects. Tables and chairs are among the most salient objects in indoor scenes). With this insight, Liu et al. proposed DSCLRCN [21], a model based on CNNs that also incorporates global context and scene context using RNNs. Their experiments have obtained outstanding results in the MIT Saliency Benchmark.

Recently, there has been interest in finding appropiate loss functions. Huang et al. [10] made an interesting contribution by introducing loss functions based on metrics that are differentiable, such as NSS, CC, SIM and KL divergence to train a network (see [26] and [18]).

2.2. Scanpath prediction

Unlike with the related task of saliency map prediciton, there has not been much progress in the task of scanpath prediciton over the last years. Cerf et al. [4] discovered that observers, even when not instructed to look for anything particular, fixate on a human face with a probability of over 80% within their first two fixations. Furthermore, they exhibit more similar scanpaths when faces are present. Recently, Hu et al. [9] have introduced a model capable of selecting relevant areas of a 360° video and deciding in which direction should a human observer look at each frame. An object detector is used to propose candidate objects of interest and a RNN selects the main object at each frame.
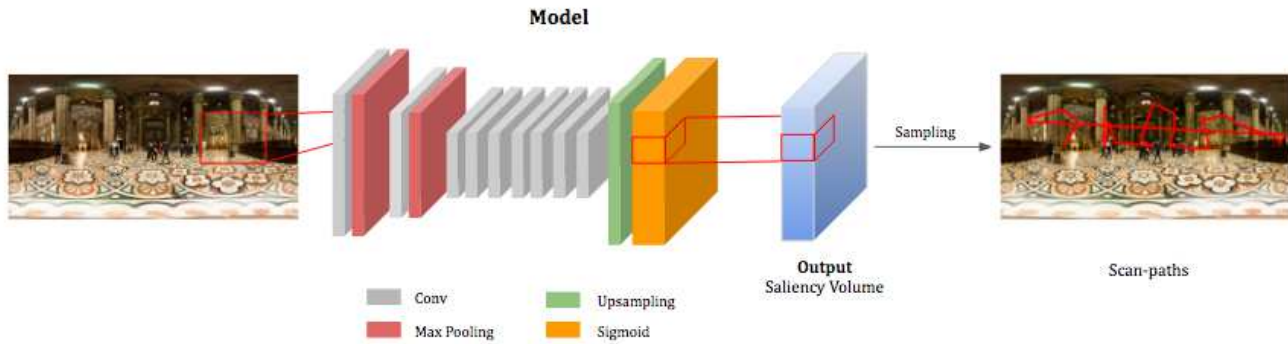
Figure 2: Overall architecture of the proposed scanpath estimation system.

3. Architecture

The central element in the architecture of SaltiNet is a deep convolutional neural network (DCNN) that predicts a saliency volume for a given input image. This section provides detail on the structure of the network, the loss function, and the strategy used to generate scan-paths from saliency volumes.

3.1. Saliency Volumes

Saliency volumes aim to be a suitable representation of spatial and temporal saliency information for images. They have three axes that represent the width and height of the image, and the temporal dimension.

Saliency volumes are generated from information already available in current fixation datasets. First, the timestamps of the fixations are quantized. The length of the time axis is determined by the longest timestamp and the quantization step. Second, a binary volume is created by placing a '1' at fixation points and a '0' on the remaining positions. Third, a multivariate Gaussian kernel is convolved with the volume to generate the saliency volume. The values of each temporal slice are normalized, converting the slice into a probability map that represents the probability of each pixel being fixated by a user at each timestep.

Figure 1 shows how saliency volumes are a meta-representation of saliency information and how other saliency representations can be extracted from them. Saliency maps can be generated by performing an addition operation across all the temporal slices of the volume, and normalizing the values to ensure they add to one. A similar representation are *temporally weighted saliency maps*, which are generated by performing a weighted addition operation of all the temporal slices. Finally, scan-paths can also be extracted by sampling fixation points from the temporal slices. Sampling strategies that aim to generate realistic scan-paths are will be discussed in Section 5.3.

3.2. Convolutional Neural Network

We propose a convolutional neural network (CNN) that adapts the filters learned to predict flat saliency maps to predict saliency volumes. Figure 2 illustrates the architecture of the convolutional neural network, composed of 10 layers and a total of 25.8 million parameters. Each convolutional layer is followed by a rectified linear unit non-linearity (ReLU). Excluding the last layer, the architecture follows the proposal of SalNet [24], whose first three layers are initialized from the VGG-16 model [5] trained for image classification.

Our network was designed considering the amount of training data available. Different strategies where introduced to prevent overfitting. First, the model was previously trained on the similar task of saliency map prediction, and the obtained weights were fine-tunned for the task of saliency volume prediction. Second, the input images where resized to $[300 \times 600]$, a much smaller dimension than their initial size $[3000 \times 6000]$. The last layer of the network outputs a volume of size $[12 \times 300 \times 600]$, with three axis that represent time, height, and width of the image.

3.3. Scan-path sampling

We take a stochastic approach to scan-path sampling[1]. The generation of scan-paths from the saliency volumes requires determining: 1) number of fixations of each scan-path; 2) the duration in seconds of each fixation; and 3) the location of each fixation point. The first two values were sampled from their probability distributions learned from the training data. The location of each fixation point was also generated by sampling, this time from the corresponding temporal slice from the predicted saliency volume. Dif-

---

[1]We also experimented with using an LSTM to directly predict scan-paths from the training data. However, we found that this resulted in the model regressing to the image center [22]. Future work will consider using adversarial training to address this.
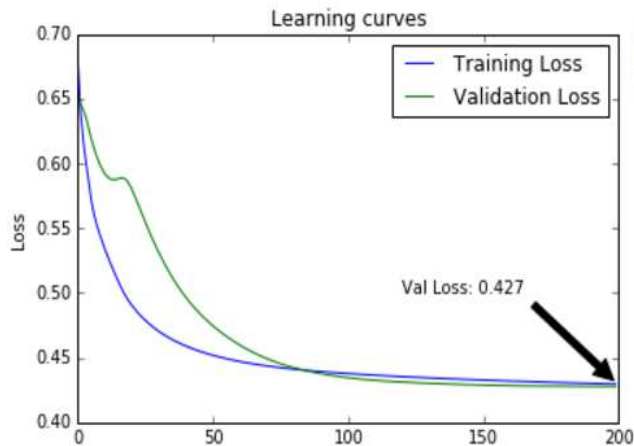
Figure 3: Training curves for our model (binary cross entropy loss.)

ferent strategies were explored for this purpose, presented together with their performance in Section 5.

4. Training

We trained the CNN in SaltiNet over 36 images of the 40 training images from the Salient360 dataset [29], leaving aside 4 images for validation. We normalized the values of the saliency volumes to be in the interval of [0, 1]. Both the input images and the saliency volumes were downsampled to 600×300 prior to training. The saliency volumes where generated from fixations using a multivariate Gaussian kernel with bandwiths $[4, 20, 20]$ (time, height, width).

The CNN was trained using stochastic gradient descent with cross entropy loss using a batch size of 1 image during 90 epoch. The binary cross entropy loss is defined as L_{BCE} in Eq. (1), where S_j and $\hat{S}_j$ correspond to the ground truth and predicted values of the saliency map.

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{j=1}^{N} S_j \log(\hat{S}_j) + (1 - S_j) \log(1 - \hat{S}_j).$$

(1)

During training, results on the validation set were tracked to monitor convergence and overfitting problems. The L^2 weight regularizer (weight decay) was used to avoid overfitting. Our network took approximately two hours to train on a NVIDIA GTX Titan X GPU using the Keras framework with the Theano backend. The learning rate was set to $\alpha = 0.001$ throughout training. Figure 3 shows the learning curves.

5. Experiments

SaltiNet was assessed and compared from different perspectives. First, we assess the impact of different sampling strategies to generate scan-paths from saliency volumes. Second, we show quantitative performance results of the model.

5.1. Datasets

Due to the small size of the training dataset, we performed transfer learning to initialize the weights of the network using related tasks. First, the network was trained to predict saliency maps using the SALICON dataset [10] using the same architecture of SalNet [24]. Then, the network was trained to predict saliency volumes generated from the iSUN dataset [31] that contains 6000 training images. The network was fine-tuned using the 60 images of the dataset of head and eye movements provided by the University of Nantes [25]. This dataset was acquired based on the images displayed on the head mounted display (HMD) Oculus-DK2. Eye gaze data was captured from a Sensomotoric Instruments (SMI) sensor in the HMD, which transmitted eye-tracking data binocularly at 60Hz. There were 40-42 observers, who could freely observe the scene with no task instructed. Each 360 images were shown for 25 seconds and there was a 5 second gray screen between two images.

5.2. Metric

The similarity metric used is a variation of the Jarodzka algorithm [12] proposed by the authors of the 360 saliency dataset [25]. The standard similarity criteria was slightly modified to use equirectangular distances in 360 instead of Euclidean distances. The generated and ground truth scan-paths are matched 1 to 1 using the Hungarian algorithm to obtain the minimum possible final cost. The presented results compare the similarity of 40 generated scan-paths with the scan-paths in the ground truth.

5.3. Sampling strategies

Figure 4 shows the distribution of the number of fixations and the duration of each fixations for the training set. During scan path generation, we sample the number of fixations and their duration from these empirical distributions.

Regarding the spatial location of the fixation points, three different strategies were explored. The simplest approach (1) consists of taking one fixation for each temporal slice of the saliency volume. Through qualitative observation we noticed that scan-paths generated in this way were unrealistic, as the probability of each fixation is not conditioned on previous fixations. A more elaborated sampling strategy (2) consists of forcing fixations to be closer to their respective previous fixation. This is accomplished by multiplying a temporal slice (probability map) of the saliency
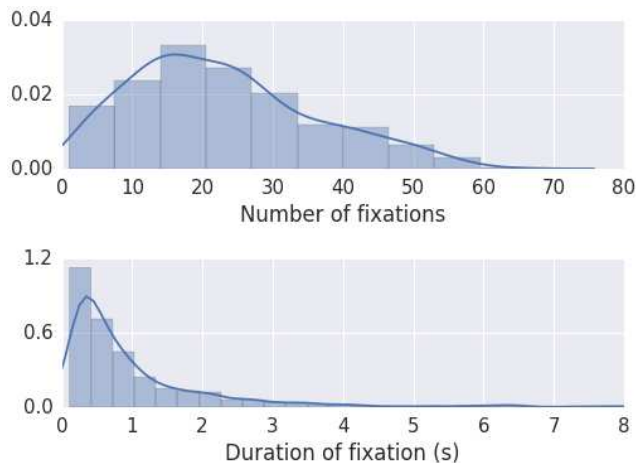
Figure 4: Empirical distributions of the number of fixations per scan-paths (top) and duration of each fixation (bottom).

	Jarodzka↓
Random scan-paths	4.94
(1) Naive sampling strategy	3.45
(3) Avoiding fixating on same places	2.82
(2) Limiting distance between fixations	**2.27**
Sampling ground truth saliency map	1.89
Sampling ground truth saliency volume	1.79
Ground truth scan-paths	1.2e-8

Table 1: Comparison between the three considered spatial sampling strategies. Lower values are better.

volume with a Gaussian kernel centered at the previous fixation point. This suppresses the probability of positions that are far from the previous fixation point. The third sampling strategy (3) we assessed consisted of suppressing the area around all previous fixations using Gaussian kernels. As shown in Table 1, we found that the best performing model was the one using sampling strategy (2).

5.4. Results

Scan-path prediction evaluation has received attention lately and it is a very active field of research [20][12].

Table 1 presents the impact of different sampling strategies over the saliency volume. We have compared our results with the accuracy that would be obtained by a model that outputs random fixations, and a model that outputs the ground truth fixations.

Table 2 compares our system with two other solutions presented at the Salient360! Challenge [29] held at the 2017 IEEE ICME conference in Hong Kong. These figures were provided by the organizers of the challenge. Results clearly indicate the superior performance of our system with re-

	Jarodzka↓
SaltiNet (Ours)	**2.8697**
SJTU	4.6565
Wuhan University	5.9517

Table 2: Comparison between three submissions to the Salient360! Challenge. Lower values are better.

spect to the two other participants.

The performance of our model has also been explored from a qualitative perspective by observing the generated saliency volumes and scan-paths. Figure 5 compares a generated scan-path with a ground truth scan-path. Figure 6 shows two examples of ground truth and generated saliency volumes.

6. Conclusions

In this work we have presented SaltiNet, a model capable of predicting scan-paths on 360° images. We have also introduced a novel temporal-aware saliency representation that is able to generate other standard representations such as scanpaths, saliency maps or temporally weighted saliency maps. Our experiments show that it is possible to obtain realistic scanpaths by sampling from saliency volumes, and the accuracy greatly depends on the sampling strategy.

We have also found the following limitations to the generation of scanpaths from saliency volumes: 1) the probability of a fixation is not conditioned to previous fixations; 2) the length of the scanpaths and the duration of each fixation are treated as independent random variables. We have tried to address the first problem by using more complex sampling strategies. Nevertheless, this three parameters are not independently distributed and therefore our model is not able to accurately represent this relationship. Future work will aim at training a fully end-to-end neural network capable of prediction the scan-paths with no need of the sampling module.

Our results can be reproduced with the source code and trained models available at `https://github.com/massens/saliency-360salient-2017`.

7. Acknowledgments
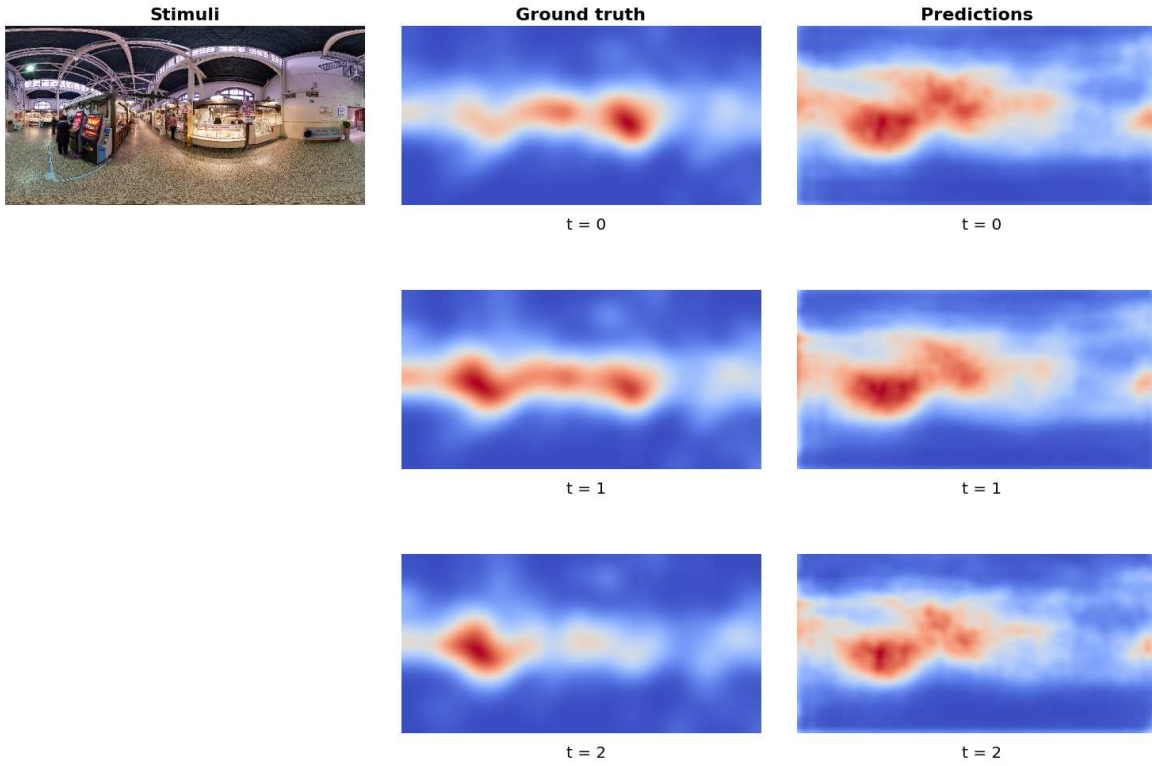
(a) Example of predicted scan-path
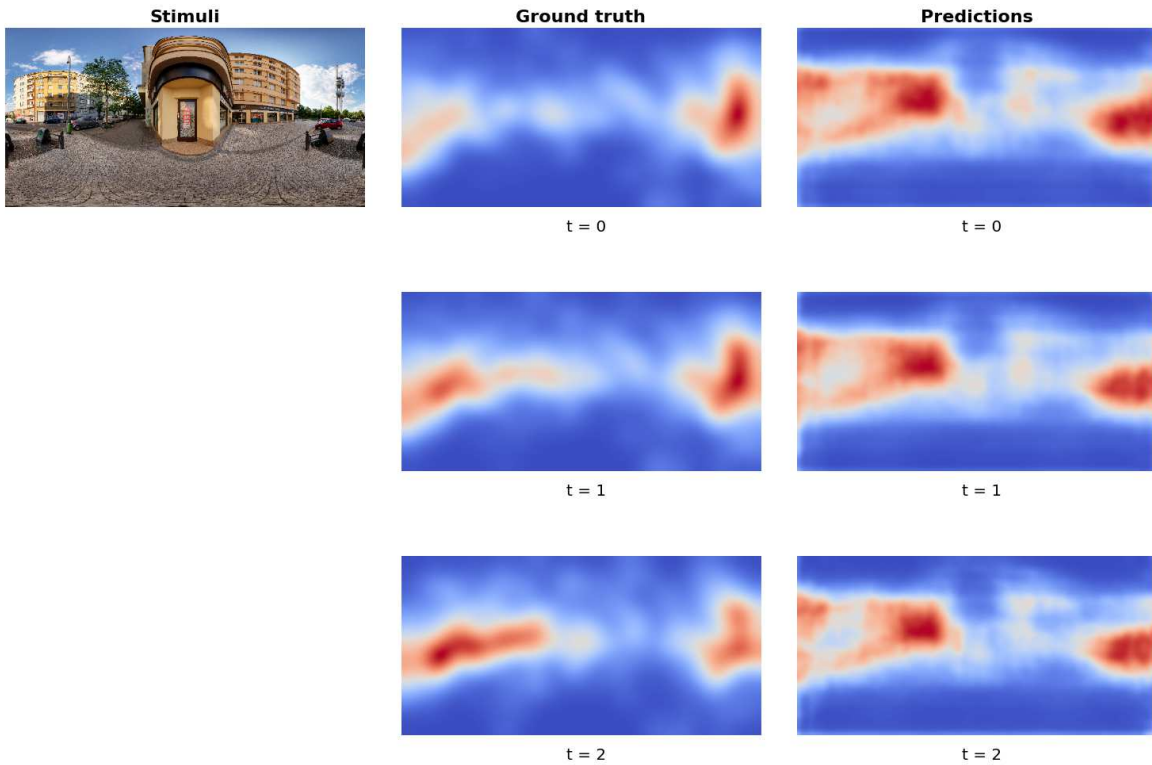

(b) Example of ground truth scan-path

Figure 5: The top image shows a predicted scanpath, sampled from a predicted saliency volume. The image at the bottom shows a single ground truth scanpath.

References

[1] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 438–445. IEEE, 2012.

[2] Z. Bylinskii, T. Judd, A. Ali Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT saliency benchmark. http://saliency.mit.edu/.

[3] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016.

[4] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, pages 241–248, 2008.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*, 2016.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[9] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 {\deg} sports video. *arXiv preprint arXiv:1705.01759*, 2017.

[10] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[12] H. Jarodzka, K. Holmqvist, and M. Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 211–218. ACM, 2010.

[13] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[15] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[18] M. Kümmerer, T. S. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015.

[19] M. Kümmerer, T. S. Wallis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv preprint:1610.01563*, 2016.

[20] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[21] N. Liu and J. Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *arXiv preprint arXiv:1610.01708*, 2016.

Stimuli **Ground truth** **Predictions**

t = 0 t = 0

t = 1 t = 1

t = 2 t = 2

(a) Indoor example

Stimuli **Ground truth** **Predictions**

t = 0 t = 0

t = 1 t = 1

t = 2 t = 2

(b) Outdoor example

Figure 6: The images above show the predicted and ground truth saliency volumes for a given stimulus. For each saliency volume, three temporal slices are shown.

[22] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[23] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[24] J. Pan, E. Sayrol, X. Giró-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 205–210. ACM, 2017.

[26] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160, 2013.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[29] T. University of Nantes. Salient360: Visual attention modeling for 360 images grand challenge, 2017.

[30] N. Wilming, S. Onat, J. P. Ossandón, A. Açık, T. C. Kietzmann, K. Kaspar, R. R. Gameiro, A. Vormberg, and P. König. An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, 4:160126, 2017.

[31] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.