Adversarial consistent training guided omnidirectional scanpath prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract
Emerging concepts like the Metaverse introduced a revolution in research related to virtual reality. Most actors in technology are paying more attention to the study of visual attention which has seen quite the development since the explosive advent of deep learning in the last decade. In this paper, we introduce a simple neural architecture and use a conditional adversarial consistent training approach to guide our neural network. The approach anchors one ground truth scanpath and varies another to improve the consistency of the discriminative network, and thus the predicted scanpath by extension. We also introduce partial connections for the multi-layer perceptron to implement an ordinality inductive bias for the sequences. Our method achieved great quantitative and qualitative results when compared to other state-of-the-art methods.

1 Introduction
Virtual and Augmented Reality (VR/AR) are becoming a center of attention in the vision community, especially as they are promising a fully immersive experience for the users. This has brought media types like the 360° omnidirectional images to the focal sphere of interest. This type of image covers the entirety of the space surrounding the viewing point (360° × 180°). This enables the viewer to observe any direction in the surrounding space. We define the angle of the field of view displayed by the headset as the viewport and the direction is represented by the viewport center. The ability to predict the attended viewport that corresponds to the orientations of the head movements beforehand helps to optimize the delivery process. Thus providing a better Quality of Experience (QoE) to the user. This can be achieved through the prediction of human visual attention, we qualify the most attractive regions in the scene as salient regions. The phenomenon is usually illustrated by using heatmaps (i.e. saliency maps), where the value of each pixel represents the probability and attractiveness of the corresponding pixel on the original scene to the gaze of the viewer. The saliency map is constructed by aggregating the gaze paths taken by several viewers and representing their distribution. These paths are called scanpaths. The points that designate the scanpath are called fixation points. They represent the regions most attractive to one of the viewers.

In this paper, we:
1. Present a deep fully convolutional architecture to predict the visual attention for omnidirectional images.
2. Employ a consistent conditional adversarial training approach to guide the training of our architecture.

The code for this work will be released publicly shortly.

3. Introduce architectural restriction to add an ordinality inductive bias to Multi-Layer Perceptrons (MLP).

2 Proposed Method

![Figure 1: Baseline model architecture.](image)

In this section, we detail the baseline model we used for scanpath prediction as well as our adversarial training method.

2.1 Model

Figure 1 presents the main simple architecture components, respectively, we used a VGG-16 [13] as a feature extractor to encode the images into a different representational space and extract the spatial and semantic information relevant to our subjective task. The encoded features are then passed through an adaptive average pooling layer with a $(1 \times 2)$ kernel. This would help extract the most relevant features from the encoded representation maps. These features are then transformed through a 2D convolutional block composed of 5 layers. We use a point-wise convolution (i.e. $(1 \times 1)$ kernel) to transform our feature vectors to an appropriate scanpath vector, in each layer we reduce the dimensionality in order to have 100 fixation points in the predicted scanpath. The choice of having 100 points was made in accordance with statistical analysis of the lengths of scanpaths in our training dataset.

2.2 Adversarial Training

The architecture of our baseline model introduces an inductive bias of invariance to ordinal information. However, this property presents a hypothetical misconception in relation to the nature of scanpaths being of an ordinal sequence of fixations. Therefore, we employed an adversarial training approach by employing a new discriminator network that factors in the ordinality of its inputs. This relies on a min-max game between our baseline generator model and the discriminator model, where the discriminator tries to differentiate the distribution of the generated scanpaths $\hat{Y}$ from the ground truth $Y$. This makes the discriminator network behave like a gradually changing loss function that eventually models distribution of the ground-truth data. To achieve that we employ the architectural composition presented in Figure 2 for the discriminator network.

This later, takes in at a first time the generated scanpath $\hat{y}$ and a ground truth anchor scanpath $y_a$ at the same time to classify $\hat{y}$. This helps to condition any generated scanpath $\hat{y}$ to an anchor ground truth scanpath $y_a$. A second pass to classify the real scanpath uses two ground truth scanpaths, the first is the anchor scanpath $y_a$, while the other is randomly selected each time from other viewers, we define it as $y$. We used this mechanism to introduce consistency between the generated scanpath and all the scanpaths of viewers, as the goal is not only to generate scanpaths in accordance with properties from multiple viewers.

The modeling of the training would be represented by the following equation:

$$
\min_{G} \max_{D} V(G, D) = \mathbb{E}_{\hat{y} \sim p_{\text{data}}(y, y_a)}[\log D(\hat{y}|y, y_a)] + \mathbb{E}_{x \sim p_{\text{z}}(x)}[1 - \log D(G(x|y, y_a))] \quad (1)
$$

2
Where $x$ represents the input image, $y$ represents a random ground truth scanpath, $y_a$ represents the anchor ground truth scanpath, and $\hat{y}$ represents the predicted scanpath. $G$ represents our baseline model and $D$ represents the discriminator network.

The sequences of the 2 spacial dimensions are fed to the network separately, firstly, in order to decorrelate, and on another hand due to the limitation of perceptrons to operate on multi-variate vectors. The input perceptron layers had their connection restricted where each neuron connects only to the neuron of the same ordinal rank and after from the next layer. This restriction introduces a sequentiality inductive bias to the architecture and represents an important feature in design to mitigate the ordinal invariance of the baseline model. Features extracted from the predicted scanpath $\hat{y}$ and the ground-truth scanpath are then concatenated and passed through 4 fully connected layers activated by a $\text{LeakyRelu}$ function with a slope of 0.2.

The model was trained on 70 images from the Salient360! dataset \cite{11} \cite{12} for 300 epochs on Nvidia Quadro RTX 8000. We used Adam \cite{7} optimizer and a learning rate of $10^{-4}$ for the generator and $10^{-5}$ for the discriminator.

3 Experimental protocol

3.1 Dataset

To our knowledge the Salient360! dataset \cite{11} \cite{12} is the only one provide scanpath labels for its images. The dataset was proposed as part of the 2018 Salient360! challenge. It is composed of 85 omnidirectional images provided with their corresponding scanpaths and saliency maps. There are about 36 scanpaths for each image and each scanpath is composed of nearly 100 fixation points designated by their coordinates.

We split the dataset into train-test sets where the former is composed of 70 images while the latter is composed of 15 images. This designates an 82 - 18% split. We chose this split in compliance with practices in other papers \cite{6} \cite{14}. It should be noted that while the testing was done on 15 images which may seem lacking, the comparison of the predicted scanpaths was done with all the ground truth scanpaths, totaling more than 540 scanpaths.

3.2 Quantitative evaluation

In order to evaluate our model, we employ several vector-based and hybrid metrics. We employed the vector-based $\text{Jarodzka}$ metric \cite{4} which compares two scanpaths vectors on shape, direction,
and time. As our model does not predict the time variable we omit its results for calculating the final results. We used the same code provided in the competition toolbox to calculate this metric [5].

The other hybrid metrics are NSS [10] which calculates the normalized mean saliency value of the scanpath fixations, and congruency [9] which calculates the ratio of fixation points present in salient regions.

Table 1 presents the results obtained by our model on these metrics and compares them to the state-of-the-art models commonly used in literature. We can observe that we obtained results similar to state-of-the-art results on the Jarodzka metric, while the results on the NSS show a considerable improvement, we also achieved meaningful results on the congruency metric which calculates the ratio of fixations located in the most salient regions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Jarodzka ↓</th>
<th>NSS ↑</th>
<th>Congruency ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>PathGan[1]</td>
<td>0.1777</td>
<td>-0.1518</td>
<td>-</td>
</tr>
<tr>
<td>SaltiNet[2]</td>
<td>0.2621</td>
<td>0.0834</td>
<td>-</td>
</tr>
<tr>
<td>SalyPath360[6]</td>
<td>0.1363</td>
<td>0.2896</td>
<td>-</td>
</tr>
<tr>
<td>Our model</td>
<td>0.1361</td>
<td>0.3867</td>
<td>0.3265</td>
</tr>
</tbody>
</table>

Table 1: Scanpath prediction comparison. Best results are highlighted in bold.

3.3 Qualitative Evaluation

Figure 3 represents a visualization of the obtained scanpaths from our model. For each image, we depict the predicted scanpath with the white segments and the fixation point with the black point. In order to be able to visually evaluate the scanpaths, we overlaid the image with the ground-truth saliency heatmap.

We can clearly observe that the scanpath adheres to the equator bias theory, which states that most fixations in an omnidirectional image are distributed around the equator. It is also worth noting that fixation points are largely located in highly salient regions. These results indicate the successful and appropriate prediction of relevant scanpaths.

4 conclusion

In this paper, we introduce a new architecture used for scanpath prediction on omnidirectional images. We opted for a fully convolutional architecture. Due to the inductive biases introduced by the simplicity of the architecture we used an adversarial training approach where we integrated a sequentiality mechanism to the discriminator that we employ as a progressive learnable loss function. We added a conditional consistent element to the architecture by anchoring one ground truth scanpath and varying the second throughout the training. We also achieved state-of-the-art results for the quantitative results on several metrics and demonstrated the validity of our model’s prediction through outstanding qualitative results. As the model does not predict the time component, it presents a fine limitation that could be overcome in the future.
References


Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default [TODO] to [Yes], [No], or [N/A]. You are strongly encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section ??.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] See the results at Section 3.
   
   (b) Did you describe the limitations of your work? [Yes] Refer to the conclusion section
   
   (c) Did you discuss any potential negative societal impacts of your work? [No] While the work has some technological impact it does not have any direct social impact.
   
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] While the data used are publicly available, the code will soon be open sourced.
   
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Refer to Section 3.1 and 2.2 for the details.
   
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Refer to section 2.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   
   (b) Did you mention the license of the assets? [No] The data used were part of a public challenge at ICME2017.
   
   (c) Did you include any new assets either in the supplemental material or as a URL? [TODO]
   
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] As include in the appendix.
   
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
A Appendix - Additional Qualitative Data

Figure 4 represents some additional visualizations for the results obtained on the dataset.

B Appendix - Qualitative progression of the training

In figure 5, the central image represents the evolution of scanpath shape predicted on an image during training. To the sides are examples of ground truth scanpaths. We can notice that the scanpath shape is slowly approaching that of the ground truths. This process was rendered to a video where the changes that happen during all the epochs can be visualized on the following link: https://youtu.be/myac0WGD-Y8
Figure 5: Qualitative evolution of scanpath prediction during training.