ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images

Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belen Masia

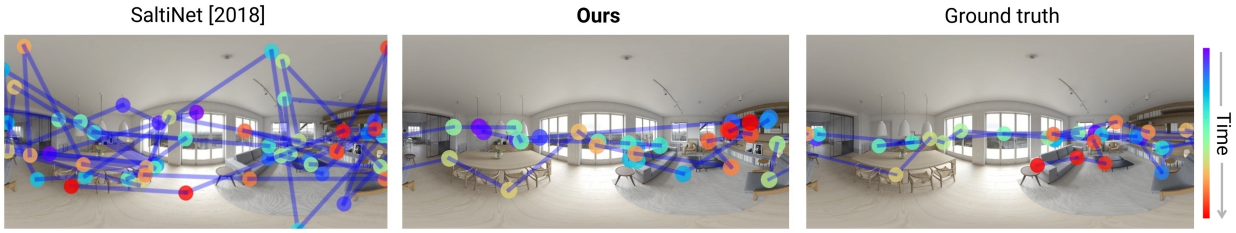

Fig. 1. We present ScanGAN360, a generative adversarial approach to scanpath generation for 360° images. For a given 360° scene, ScanGAN360 generates realistic scanpaths (*center*), outperforming state-of-the-art methods (*left*) and mimicking the human baseline (*right*).

Abstract— Understanding and modeling the dynamics of human gaze behavior in 360° environments is crucial for creating, improving, and developing emerging virtual reality applications. However, recruiting human observers and acquiring enough data to analyze their behavior when exploring virtual environments requires complex hardware and software setups, and can be time-consuming. Being able to generate *virtual observers* can help overcome this limitation, and thus stands as an open problem in this medium. Particularly, generative adversarial approaches could alleviate this challenge by generating a large number of scanpaths that reproduce human behavior when observing new scenes, essentially mimicking virtual observers. However, existing methods for scanpath generation do not adequately predict realistic scanpaths for 360° images. We present ScanGAN360, a new generative adversarial approach to address this problem. We propose a novel loss function based on dynamic time warping and tailor our network to the specifics of 360° images. The quality of our generated scanpaths outperforms competing approaches by a large margin, and is almost on par with the human baseline. ScanGAN360 allows fast simulation of large numbers of virtual observers, whose behavior mimics real users, enabling a better understanding of gaze behavior, facilitating experimentation, and aiding novel applications in virtual reality and beyond.

Index Terms—Scanpath generation, 360° images, virtual reality, generative adversarial models, saliency, human behavior.

1 Introduction

Virtual reality (VR) is an emerging medium that unlocks unprecedented user experiences. To optimize these experiences, it is crucial to understand how people explore immersive virtual environments [19]. However, this depends on gathering large amounts of data from many observers exploring multiple virtual environments, which requires complex hardware and software setups and is a time-consuming task. This burden could be alleviated by modeling time-dependent visual exploration of a given 360° scene, where gaze behavior predictions from the model serve as *virtual observers* of the scene. These models are important for many applications in VR, including designing and editing VR content [49], generating realistic gaze trajectories of digital avatars [18], understanding dynamic visual attention and visual search behavior [67], or developing new rendering, display, and compression algorithms, among others.

Current approaches that model how people explore virtual environments often leverage saliency prediction [4, 38, 50]. While this is useful for some applications, the fixation points or regions predicted by these approaches do not account for the time-dependent visual behavior of the user, making it difficult to predict the order of fixations, or give insight into how people explore an environment over time. To handle this temporality, some recent work has explored scanpath prediction in 360° images [4–6, 70]. However, these algorithms do not adequately model how people explore immersive virtual environments, resulting in erratic or non-plausible scanpaths.

Scanpath prediction has also been explored for conventional 2D images [7, 14, 54]. However, virtual environments are inherently different from traditional images: they offer a larger space to interact with, and users fully control the camera or viewpoint (i.e., they decide where to look at), seeing only a part of a larger scene at each moment. For instance, a region of interest in a traditional image can be seen regardless of the starting point, and therefore many users could direct their attention to it. However, in 360° content, a region of interest may or may not fall in the view of an observer. This adds another degree of freedom to the way in which users explore a scene, allowing them the freedom to fixate on other parts of the scene. This generally leads to more complex and varied gaze patterns between users [50], hence hampering the precision of existing 2D techniques, which are trained on controlled conditions with the whole stimuli visible to the observer for a shorter period of time. Moreover, traditional methods are trained on sets of data that present certain biases, such as the center bias [31], where most of the gaze information falls in the center of the image. However, these biases vary in virtual environments, where, for instance, gaze has been shown to be directed to the whole equator, rather than just the center of the image. The patterns employed by conventional 2D models or present in data from conventional 2D images are not representative of viewing behavior in virtual environments [50]) and could therefore reduce their applicability on them.

In this work, we present ScanGAN360, a novel framework for scanpath generation for 360° images (Figure 1). Our model builds on a conditional generative adversarial network (cGAN) architecture, for

- *Correspondence to: danims@unizar.es*
- *Daniel Martin, Ana Serrano, and Belen Masia are with Universidad de Zaragoza, I3A.*
- *Alexander W. Bergman and Gordon Wetzstein are with Stanford University.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

which we discuss and validate two important insights that we show are necessary for realistic scanpath generation. First, we propose a novel loss function based on a spherical adaptation of dynamic time warping (DTW); DTW is a metric for measuring similarity between two time series, such as scanpaths, which to our knowledge has not been used so far in this context. Second, we present a parameterization of the scanpaths specifically tailored to 360° content. These insights allow us to demonstrate state-of-the-art results for scanpath generation in 360° images, close to the human baseline and far surpassing the performance of existing methods. Further, our approach is the first to enable robust scanpath prediction over long time periods of up to 30 seconds, and, unlike previous work, our model does not rely on saliency, which is typically not available as ground truth.

Our model produces about 1,000 scanpaths per second, which enables fast simulation of large numbers of *virtual observers*, whose behavior mimics that reported for real users [50], without requiring complex setups or time-consuming experiments. Using ScanGAN360, we also explore applications in virtual scene design, which is useful in video games, interior design, cinematography, and tourism, and scanpath-driven video thumbnail generation of 360° images, which provides previews of VR content for social media platforms. Beyond these applications, we propose to use ScanGAN360 for applications such as gaze behavior simulation for virtual avatars or gaze-contingent rendering.

Our contributions can be summarized as follows:

- We present a generative adversarial approach with an architecture and loss function specifically designed for scanpath generation in 360° images and a parameterization of the scanpaths tailored to this kind of content, and show that it is able to outperform existing approaches both qualitatively and quantitatively using metrics specifically designed for scanpath evaluation.
- Our model is able to generate around 1,000 scanpaths per second, whose behavioral aspects and exploratory patterns closely mimic those of captured, real human scanpaths [50].
- We show how our model can also be useful for tackling different open problems in VR, including assisting virtual scene design, or generating scanpath-driven video thumbnails of static 360° images.

Our code and pre-trained model are publicly available at: `https://webdiis.unizar.es/~danims/projects/ScanGAN360`.

2 Related work

Modeling and predicting attention The multimodal nature of attention [37], together with the complexity of human gaze behavior, make this a very challenging task. Many works devoted to it have relied on representations such as saliency, which is a convenient representation for indicating the regions of an image more likely to attract attention. Early strategies for saliency modeling have focused on either creating hand-crafted features representative of saliency [8, 23, 24, 36, 59, 69], or directly learning data-driven features [29, 57]. With the proliferation of extensive datasets of human attention [9, 24, 46, 50, 66], deep learning–based methods for saliency prediction have been successfully applied, yielding impressive results [14, 43, 44, 58, 60, 61, 64].

However, saliency models do not take into account the dynamic nature of human gaze behavior, and therefore, they are unable to model or predict time-varying aspects of attention. Being able to model and predict dynamic exploration patterns has been proven to be useful, for example, for avatar gaze control [12, 48], video rendering in virtual reality [33], or for directing users' attention over time in many contexts [10, 45]. Scanpath models aim to predict visual patterns of exploration that an observer would perform when presented with an image. In contrast to saliency models, scanpath models typically focus on predicting plausible scanpaths, i.e., they do not predict a unique scanpath and instead they try to mimic human behavior when exploring an image, taking into account the variability between different observers. Ellis and Smith [16] were pioneers in this field: They proposed a general framework for generating scanpaths based on Markov stochastic processes. Several approaches have followed this work, incorporating behavioral biases in the process in order to produce more plausible scanpaths [31, 34, 55, 56, 68].

In recent years, different deep learning approaches have been proposed to predict human scanpaths. Some of them have resorted to saliency prediction as a proxy for gaze estimation, either with deep convolutional models [7, 21, 26, 29, 60], or with iterative representation learning [63]. Further, some works have leveraged the benefits of recurrent neural networks to inherently model the temporal nature of scanpaths, either based on region-of-interest and inhibition-of-return strategies [54] or attentive modules [14]. Different from these works, we tackle the problem of scanpath prediction in 360° images from a generative perspective, and without resorting to saliency, which is not usually available as ground truth. We refer the reader to the state of the art review by Kümmerer et al. [28] for an exhaustive comparison of previous approaches on scanpath prediction.

Attention in 360° images Predicting plausible scanpaths in 360° rather than 2D imagery is a more complex task: Observers do not only scan a given image with their gaze, but they can now also turn their head or body, effectively changing their viewport over time. Several works have been proposed for modeling saliency in 360° images [11, 38, 40, 50, 51, 65]. However, scanpath prediction has received less attention. In their recent work, Assens et al. [5] generalize their 2D model to 360° images, but their loss function is unable to reproduce the behavior of ground truth scanpaths (see Figure 4, third column). A few works have focused on predicting short-term sequential gaze points based on users' previous history for 360° videos [21, 32, 42, 62, 65], but they are limited to small temporal windows (from one to ten seconds). For the case of still images, a number of recent methods focus on developing improved saliency models and principled methods to sample from them [4, 6, 70], or rely on additional inputs such as head orientation at each time step [22] or historical and task-related data [20].

Instead, we directly learn dynamic aspects of attention from ground truth scanpaths by training a generative model in an adversarial manner, with an architecture and loss function specifically designed for scanpaths in 360° images. This allows us to (i) effectively mimic human behavior when exploring scenes, bypassing the saliency generation and sampling steps, and (ii) optimize our network to stochastically generate 360° scanpaths, taking into account observer variability.

3 Our Model

We adopt a generative adversarial approach, specifically designed for 360° content in which the model learns to generate a plausible scanpath, given the 360° image as a condition. In the following, we describe the parameterization employed for the scanpaths, the design of our loss function for the generator, and the particularities of our conditional GAN architecture, ending with details about the training process.

3.1 Scanpath Parameterization

Scanpaths are commonly provided as a sequence of two-dimensional values corresponding to the coordinates (i, j) of each gaze point in the image. When dealing with 360° images in equirectangular projections, gaze points are also often represented by their latitude and longitude (ϕ, λ), $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\lambda \in [-\pi, \pi]$. However, these parameterizations either suffer from discontinuities at the borders of a 360° image, or result in periodic, ambiguous values. The same point of the scene can have two different representations in these parameterizations: any longitude λ, $\lambda + 2\pi$ represents the same meridian. This includes the leftmost (λ=-180°) and rightmost (λ=180°) borders of the 360º image. This ambiguity may hinder the learning process. We therefore resort to a three-dimensional parameterization of our scanpaths, where each gaze point $p = (\phi, \lambda)$ is transformed into its three-dimensional representation $P = (x, y, z)$ such that:

$$x = \cos(\phi)\cos(\lambda); \; y = \cos(\phi)\sin(\lambda); \; z = \sin(\phi).$$

This transformation assumes, without loss of generality, that the panorama is projected over a unit sphere. We use this parameterization

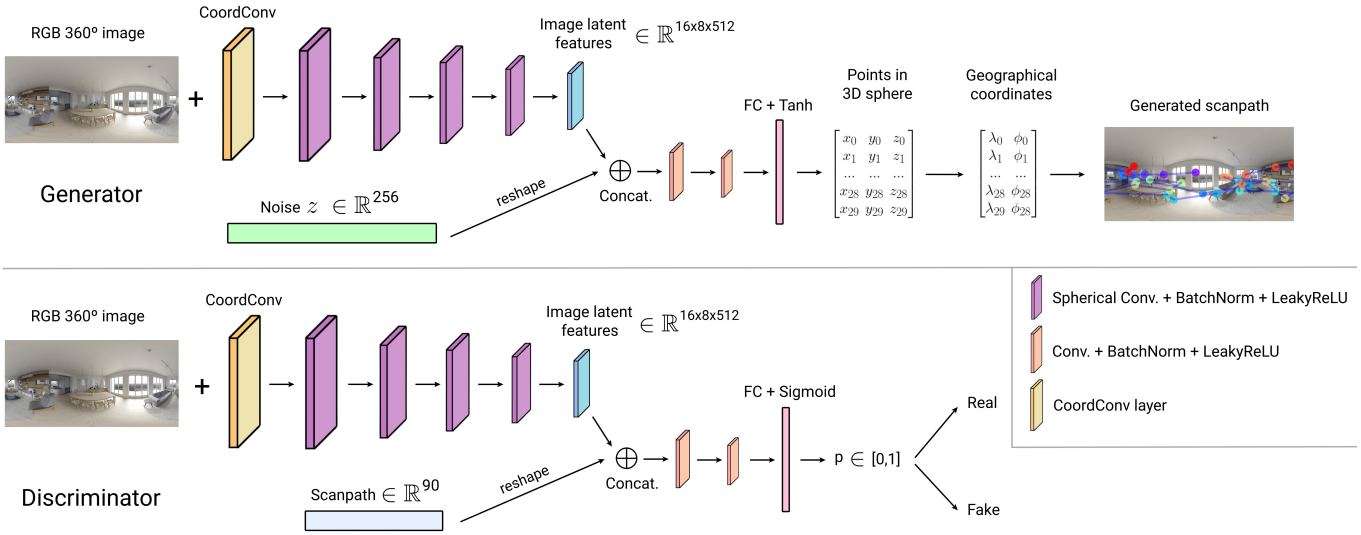

Fig. 2. Illustration of our generator and discriminator networks. Both networks have a two-branch structure: Features extracted from the 360° image with the aid of a CoordConv layer and an encoder-like network are concatenated with the input vector for further processing. The generator learns to transform this input vector, conditioned by the image, into a plausible scanpath. The discriminator takes as input vector a scanpath (either captured or synthesized by the generator), as well as the corresponding image, and determines the probability of this scanpath being real (or fake). We train them end to end in an adversarial manner, following a conditional GAN scheme. Please refer to the text for details on the loss functions and architecture.

for our model, which learns a scanpath $\mathbf{P}$ as a set of three-dimensional points over time. Specifically, given a number of samples T over time, $\mathbf{P} = (P_1, ..., P_T) \in \mathbb{R}^{3\times T}$. The results of the model are then converted back to a two-dimensional parameterization in terms of latitude ($\phi = \text{atan2}(z, \sqrt{x^2+y^2})$) and longitude ($\lambda = \text{atan2}(y,x)$) for display and evaluation purposes.

3.2 Overview of the Model

Our model is a conditional GAN, where the condition is the RGB 360° image for which we wish to estimate a scanpath. The generator G is trained to generate a scanpath from a latent code z (drawn randomly from a uniform distribution, $\mathscr{U}(-1,1)$), conditioned by the RGB 360° image y. The discriminator D takes as input a potential scanpath (x or $G(z,y)$), as well as the condition y (the RGB 360° image), and outputs the probability of the scanpath being real (or fake). The architecture of both networks, generator and discriminator, can be seen in Figure 2, and further details related to the architecture are described in Section 3.4.

3.3 Loss Function

The objective function of a conventional conditional GAN is inspired by a minimax objective from game theory, with an objective [39]:

$$\min_G \max_D V(D,G) = \mathbb{E}_x[\log D(x,y)] + \mathbb{E}_z[\log(1 - D(G(z,y),y)]. \tag{1}$$

We can separate this into two losses, one for the generator, $\mathscr{L}_G$, and one for the discriminator, $\mathscr{L}_D$:

$$\mathscr{L}_G = \mathbb{E}_z[\log(1 - D(G(z,y),y))], \tag{2}$$

$$\mathscr{L}_D = \mathbb{E}_x[\log D(x,y)] + \mathbb{E}_z[\log(1 - D(G(z,y),y))]. \tag{3}$$

While this objective function suffices in certain cases, as the complexity of the problem increases, the generator may not be able to learn the transformation from the input distribution into the target one. One can resort to adding a loss term to $\mathscr{L}_G$, and in particular one that enforces similarity to the scanpath ground truth data. However, using a conventional data term, such as MSE, does not yield good results (Section 4.3 includes an evaluation of this). To address this issue, we introduce a novel term in $\mathscr{L}_G$ specifically targeted to our problem, and based on dynamic time warping [41].

Dynamic time warping (DTW) measures the similarity between two temporal sequences, considering both the shape and the order of the elements of a sequence, without forcing a one-to-one correspondence between elements of the time series. For this purpose, it takes into account all the possible alignments of two time series $\mathbf{r}$ and $\mathbf{s}$, and computes the one that yields the minimal distance between them. Specifically, the DTW loss function between two time series $\mathbf{r} \in \mathbb{R}^{k\times n}$ and $\mathbf{s} \in \mathbb{R}^{k\times m}$ can be expressed as [15]:

$$\text{DTW}(\mathbf{r},\mathbf{s}) = \min_A \langle A, \Delta(\mathbf{r},\mathbf{s}) \rangle, \tag{4}$$

where $\Delta(\mathbf{r},\mathbf{s}) = [\delta(r_i, s_j)]_{ij} \in \mathbb{R}^{n\times m}$ is a matrix containing the distances $\delta(\cdot,\cdot)$ between each pair of points in $\mathbf{r}$ and $\mathbf{s}$, A is a binary matrix that accounts for the alignment (or correspondence) between $\mathbf{r}$ and $\mathbf{s}$, and $\langle \cdot,\cdot \rangle$ is the inner product between both matrices.

In our case, $\mathbf{r} = (r_1, ..., r_T) \in \mathbb{R}^{3\times T}$ and $\mathbf{s} = (s_1, ..., s_T) \in \mathbb{R}^{3\times T}$ are two scanpaths that we wish to compare. While the Euclidean distance between each pair of points is usually employed when computing $\delta(r_i, s_j)$ for Equation 4, in our scenario that would yield erroneous distances derived from the projection of the 360° image (both if done in 2D over the image, or in 3D with the parameterization described in Section 3.1). We instead use the distance over the surface of a sphere, or spherical distance, and define $\Delta_{sph}(\mathbf{r},\mathbf{s}) = [\delta_{sph}(r_i, s_j)]_{ij} \in \mathbb{R}^{n\times m}$ such that:

$$\delta_{sph}(r_i, s_j) = 2\arcsin\left(\frac{1}{2}\sqrt{(r_i^x - s_j^x)^2 + (r_i^y - s_j^y)^2 + (r_i^z - s_j^z)^2}\right), \tag{5}$$

leading to our spherical DTW:

$$\text{DTW}_{sph}(\mathbf{r},\mathbf{s}) = \min_A \langle A, \Delta_{sph}(\mathbf{r},\mathbf{s}) \rangle. \tag{6}$$

We incorporate the spherical DTW to the loss function of the generator ($\mathscr{L}_G$, Equation 2), yielding our final generator loss function $\mathscr{L}_G^*$:

$$\mathscr{L}_G^* = \mathscr{L}_G + \lambda \cdot \mathbb{E}_z[\text{DTW}_{sph}(G(z,y),\rho)], \tag{7}$$

where ρ is a ground truth scanpath for the conditioning image y, and the weight λ is empirically set to 0.1.

While a loss function incorporating DTW (or spherical DTW) is not differentiable, a differentiable version, soft-DTW, has been proposed. We use this soft-DTW in our model; details on it can be found in Section S1 in the supplementary material or in the original publication [15].

3.4 Model Architecture

Both our generator and discriminator are based on a two-branch structure (see Figure 2), with one branch for the conditioning image y and the other for the input vector (z in the generator, and x or $G(z,y)$ in the discriminator). The image branch extracts features from the 360° image, yielding a set of latent features that will be concatenated with the input vector for further processing. Due to the distortion inherent to equirectangular projections, traditional convolutional feature extraction strategies are not well suited for 360° images: They use a kernel window where neighboring relations are established uniformly around a pixel. Instead, we extract features using panoramic (or spherical) convolutions [13], which have been shown to perform better for equirectangular content [65]. Spherical convolutions are a type of dilated convolutions where the relations between elements in the image are not established in image space, but in a gnomonic, non-distorted space. These spherical convolutions can represent kernels as patches tangent to a sphere where the 360° is reprojected, and therefore allow the network to learn spatial relations of 360° content, such as longitudinal continuities or spherical distortions. Once the main features are extracted (see Figure 2), we resort to traditional convolutions to process the remaining latent information.

In our problem of scanpath generation, the location of the features in the image is of particular importance, hence we train our model to learn a mapping from the image domain to coordinates. To facilitate the spatial learning of the network, as well as to help stabilize the training process, we use the recently presented CoordConv strategy [35], which gives convolutions access to their own input coordinates by adding extra coordinate channels. We do this by concatenating a CoordConv layer to the input 360° image (see Figure 2). Without CoordConv, the network would be forced not only to learn the salient image features, but also their explicit coordinate location in the input image. Since convolutional neural network (CNN) architectures are designed to be shift invariant, using explicit coordinates as input eases the learning process. An ablation study in Section 4.3 shows the effectiveness of using CoordConv.

Although our model has no explicit module to handle time-dependence, our generator implicitly learns the sequentiality from the training data by optimizing the DTW-based loss function, which is specifically tailored for handling temporal sequences. This optimization process enforces our generator to generate scanpaths whose temporal component is coherent with that of the ground-truth scanpaths.

3.5 Dataset and Training Details

We train our model using Sitzmann et al.'s [50] dataset, composed of 22 different 360° images and a total of 1,980 scanpaths from 169 different users. Each scanpath contains gaze information captured during 30 seconds with a binocular eye tracking recorder at 120 Hz. We sample these captured scanpaths at 1 Hz (i.e., $T = 30$), and reparameterize them (Section 3.1), so that each scanpath is a sequence $\mathbf{P} = (P_0, ..., P_{29}) \in \mathbb{R}^{3\times T}$. Given the relatively small size of the dataset, we perform data augmentation by longitudinally shifting the 360° images (and adjusting their scanpaths accordingly); specifically, for each image we generate six different variations with random longitudinal shifting. We use 19 of the 22 images in this dataset for training, and reserve three to be part of our test set (more details on the full test set are described in Section 4). With the data augmentation process, this yields 114 images in the training set.

During our training process we use the Adam optimizer [27], with constant learning rates $l_G = 10^{-4}$ for the generator, and $l_D = 10^{-5}$ for the discriminator, both of them with momentum = $(0.5, 0.99)$. Further training and implementation details can be found in the supplementary material.

4 Results and Analysis

We evaluate the quality of the generated scanpaths with respect to the captured, ground truth scanpaths, as well as to other approaches. We also analyze behavioral aspects of our scanpaths, and compare them to those reported for real human observers by previous work. We additionally ablate our model to illustrate the contribution of the different design choices.

We evaluate or model on two different test sets. First, using the three images from Sitzmann et al.'s dataset [50] left out of the training (Section 3.5): *room*, *chess* and *robots*. To ensure our model has an ability to generalize, we also evaluate it with a different dataset from Rai et al. [46]. This dataset consists of 60 scenes watched by 40 to 42 observers for 25 seconds. Thus, when comparing to their ground truth, we cut our 30-second scanpaths to the maximum length of their data. Please also refer to the supplementary material for more details on the test set, as well as further evaluation and results.

Qualitative results of our model can be seen in Figure 3, which, for scenes with different layouts, shows: the scene, a sample ground truth scanpath, and three of our generated scanpaths sampled from the generator. Our model is able to produce plausible, coherent scanpaths that focus on relevant parts of the scene. In the generated scanpaths we observe regions where the user focuses (points of a similar color clustered together), as well as more exploratory behavior. The generated scanpaths are diverse but plausible, as one would expect if different users watched the scene (the supplementary material contains more ground truth scanpaths, showing this diversity). Further, our model is not affected by the inherent distortions of the 360° image. This is apparent, for example, in the *market* scene: The central corridor, narrow and seemingly featureless, is observed by generated *virtual observers*.

Scanpath similarity metrics Evaluating scanpath similarity quantitatively is not a trivial task, and a number of metrics have been proposed and used in the literature [2,7,17,54], each focused on a different context or aspect of gaze behavior. Proposed metrics can be roughly categorized into: (i) direct measures based on Euclidean distance; (ii) string-based measures based on string alignment techniques (such as the Levenshtein distance, LEV); (iii) curve similarity methods; (iv) metrics from time-series analysis (like DTW, on which our loss function is based); and (v) metrics from recurrence analysis (e.g., recurrence measure REC and determinism measure DET). We refer the reader to supplementary material and the review by Fahimi and Bruce [17] for an in-depth explanation and comparison of existing metrics. Here, we include a subset of metrics that take into account both the position and the ordering of the points (namely LEV and DTW), and two metrics from recurrence analysis (REC and DET), which have been reported to be discriminative in revealing viewing behaviors and patterns when comparing scanpaths. We nevertheless compute our evaluation for the full set of metrics reviewed by Fahimi and Bruce [17] in our supplementary material.

Since for each image we have a number of ground truth scanpaths, and a set of generated scanpaths, we compute each similarity metric for all possible pairwise comparisons (each generated scanpath against each of the ground truth scanpaths), and average the result. In order to provide an upper baseline for each metric, we also compute the human baseline (*Human BL*) [63], which is obtained by comparing each ground truth scanpath against all the other ground truth ones, and averaging the results. In a similar fashion, we compute a lower baseline based on sampling gaze points randomly over the image (*Random BL*).

Quantitative results in Table 1 further show that our generated scanpaths are close to the human baseline (*Human BL*), both in the test set from Sitzmann et al.'s dataset, and over Rai et al.'s dataset. A value close to *Human BL* indicates that the generated scanpaths are as valid or as plausible as the captured, ground truth ones. Note that obtaining a value lower than *Human BL* is possible, if the generated scanpaths are on average closer to the ground truth ones, and exhibit less variance.

4.1 Comparison to Other Methods

We compare ScanGAN360 to three methods devoted to scanpath prediction in 360° images: SaltiNet-based scanpath prediction [4, 6] (we

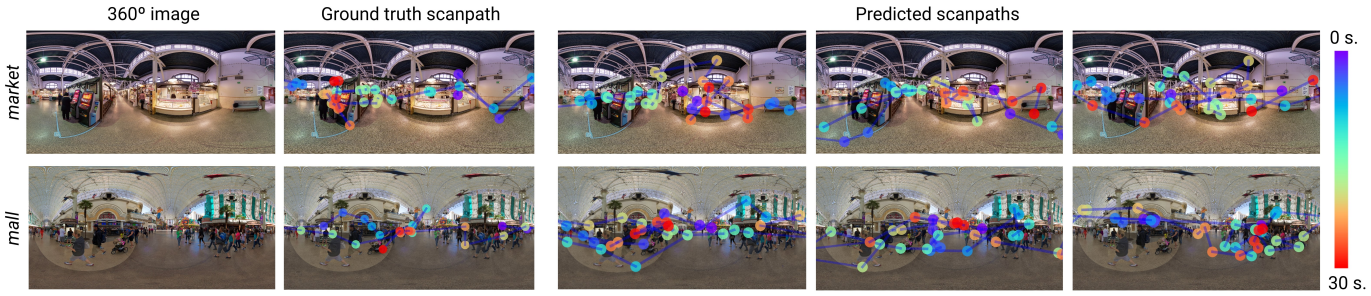

Fig. 3. Results of our model for two different scenes: *market* and *mall* from Rai et al.'s dataset [46]. *From left to right*: 360° image, ground truth sample scanpath, and three scanpaths generated by our model. The generated scanpaths are plausible and focus on relevant parts of the scene, yet they exhibit the diversity expected among different human observers. Please refer to the supplementary material for a larger set of results.

Table 1. Quantitative comparisons of our model against SaltiNet [6] and Zhu et al. [70]. We also compare against upper (human baseline, *Human BL*) and lower (randomly sampling over the image, *Random BL*) baselines. Arrows indicate whether higher or lower is better, and boldface highlights the best result for each metric (excluding the ground truth *Human BL*). *SaltiNet is trained with Rai et al.'s dataset; we include it for completeness.

Dataset	Method	LEV ↓	DTW ↓	REC ↑	DET ↑
Test set from Sitzmann et al.	Random BL	52.33	2370.56	0.47	0.93
	SaltiNet	48.00	1928.85	1.45	1.78
	ScanGAN360 (ours)	**46.15**	**1921.95**	**4.82**	**2.32**
	Human BL	43.11	1843.72	7.81	4.07
Rai et al.'s dataset	Random BL	43.11	1659.75	0.21	0.94
	SaltiNet*	48.07	1928.41	1.43	1.81
	Zhu et al.	43.55	1744.20	1.64	1.50
	ScanGAN360 (ours)	**40.99**	**1549.59**	**1.72**	**1.87**
	Human BL	39.59	1495.55	2.33	2.31

will refer to it as SaltiNet in the following), PathGAN [5] and Zhu et al.'s method [70]. For comparisons to SaltiNet we use the public implementation of the authors, while the authors of Zhu et al. kindly provided us with the results of their method for the images from Rai et al.'s dataset (but not for Sitzmann et al.'s); we therefore have both qualitative (Figure 4) and quantitative (Table 1) comparisons to these two methods. In the case of PathGAN, no model or implementation could be obtained, so we compare qualitatively to the results extracted from their paper (Figure 4, third column).

Since our model is generative, it can generate as many scanpaths as needed and model different potential observers. In order to perform a fair comparison, we carry out our evaluations on a random set of 100 scanpaths generated by our model, to match the number of generated scanpaths available for competing methods. Nevertheless, we have also analyzed the stability of our generative model by computing our evaluation metrics for a variable number of generated scanpaths (see Section 4.3).

Table 1 shows that our model consistently provides results closer to the ground truth scanpaths than Zhu et al.'s and SaltiNet. The latter is based on a saliency-sampling strategy, and thus these results indicate that indeed the temporal information learnt by our model is relevant for the final result. Our model, as expected, also amply surpasses the random baseline. In Figure 4 we see how PathGAN scanpaths fail to focus on the relevant parts of the scene (see, e.g., *snow* or *square*), while SaltiNet exhibits a somewhat erratic behavior, with large displacements and scarce areas of focus (*train*, *snow* or *square* show this). Finally, Zhu et al.'s approach tends to place gaze points at high contrast borders (see, e.g., *square* or *resort*).

4.2 Behavioral Evaluation

While the previous subsection employs well-known metrics from the literature to analyze the performance of our model, in this subsection we perform a higher-level analysis of its results. We assess whether the behavioral characteristics of our scanpaths match those which have been reported from actual users watching 360° images [50].

Exploration time Sitzmann et al. measure the *exploration time* as the average time that users took to move their eyes to a certain longitude relative to their starting point, and measure how long it takes for users to fully explore the scene. Figure 5 (left) shows this exploration time, measured by Sitzmann et al. from captured data, for the three scenes from their dataset included in our test set (*room*, *chess*, and *robots*). To analyze whether our generated scanpaths mimic this behavior and exploration speed, we plot the exploration time of our generated scanpaths (Figure 5, center left) for the same scenes and number of scanpaths. We can see how the speed and exploration time are very similar between real and generated data. Individual results per scene can be found in the supplementary material.

Fixation bias Similar to the center bias of human eye fixations observed in regular images [24], the existence of a Laplacian-like equator bias has been measured in 360° images [50]: The majority of fixations fall around the equator, in detriment of the poles. We have evaluated whether the distribution of scanpaths generated by our model also presents this bias. This is to be expected, since the data our model is trained with exhibits it, but is yet another indicator that we have succeeded in learning the ground truth distribution. We test this by generating, for each scene, 1,000 different scanpaths with our model, and aggregating them over time to produce a *pseudo*-saliency map, which we term *aggregate map*. Figure 5 (right) shows this for two scenes in our test set: We can see how this equator bias is indeed present in our generated scanpaths.

Inter-observer congruency It is common in the literature analyzing users' gaze behavior to measure inter-observer congruency, often by means of a receiver operating characteristic (ROC) curve. We compute the congruency of our "generated observers" through this ROC curve for the three scenes in our test set from the Sitzmann et al. dataset (Figure 5, center right). The curve calculates the ability of the i^{th} scanpath to predict the *aggregate map* of the corresponding scene. Each point in the curve is computed by generating a map containing the top $n\%$ most salient regions of the aggregate map (computed without the i^{th} scanpath), and calculating the percentage of gaze points of the i^{th} scanpath that fall into that map. Our ROC curve indicates a strong agreement between our scanpaths, with around 75% of all gaze points falling within 25% of the most salient regions. These values are comparable to those measured in previous studies with captured gaze data [30, 50].

Temporal and spatial coherence Our generated scanpaths have a degree of stochasticity, to be able to model the diversity of real human observers. Works from vision science have shown a relationship between spatial and temporal mechanisms in gaze behavior at different levels. For instance, Kapoula [25] showed that fixation duration on subsequent points of interest can be determined directly by the combination of the processing done at the current fixation, and the basis of partial information available in peripheral vision when the eye was at

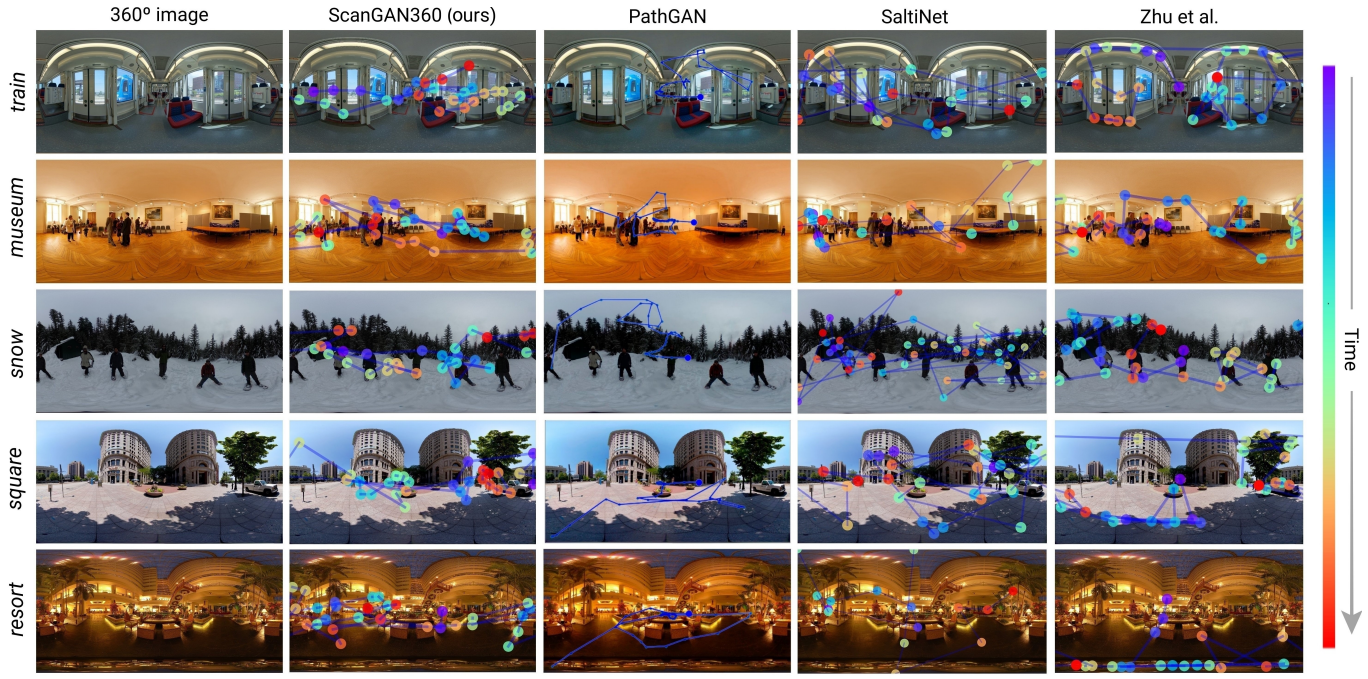

Fig. 4. Qualitative comparison to previous methods for five different scenes from Rai et al.'s dataset. In each row, from left to right: 360° image, and a sample scanpath obtained with our method, PathGAN [5], SaltiNet [6], and Zhu et al.'s [70]. Note that, in the case of PathGAN, we are including the results directly taken from their paper, thus the different visualization. Our method produces plausible scanpaths focused on meaningful regions, in comparison with other techniques. Please see text for details, and the supplementary material for a larger set of results, also including ground truth scanpaths.

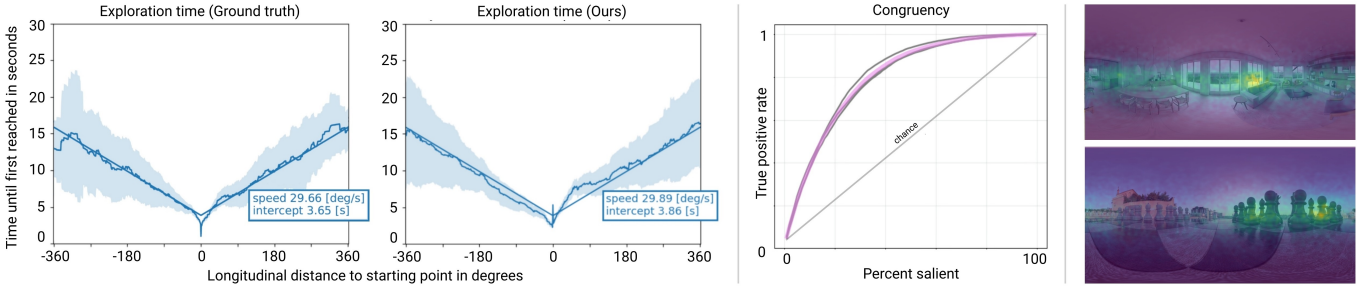

Fig. 5. *Left*: Exploration time for real captured data (*left*) and scanpaths generated by our model (*center left*). Speed and exploration time of our scanpaths are on par with that of real users. *Center right*: ROC curve of our generated scanpaths for each individual test scene (gray), and averaged across scenes (magenta). The faster it converges to the maximum rate, the higher the inter-observer congruency. *Right*: Aggregate maps for two different scenes, computed as heatmaps from 1,000 generated scanpaths. Our model is able to produce aggregate maps that focus on relevant areas of the scenes and exhibit the equator bias reported in the literature.

the preceding fixation. Thus, each gaze point is conditioned not only by the features in the scene, but also by the previous history of gaze points of the user. If two users start watching a scene in the same region, a certain degree of coherence between their scanpaths is expected, that may diverge more as more time passes, up to a point where, as the whole scene has been explored, patterns start to converge again. This has already been reported in VR, where users' attention tends to converge after approximately 17 seconds [50]. To assess whether our scanpaths actually follow a coherent pattern, we analyze the temporal coherence of generated scanpaths that start in the same region. We generate a set of random scanpaths for each of the scenes in our test dataset, and separate them according to the longitudinal region where the scanpath begins (e.g., $[0°, 40°)$, $[40°, 80°)$, etc.). Then, we estimate the probability density of the generated scanpaths from each starting region using kernel density estimation (KDE) for each timestamp (see Figure 6). During the first seconds, gaze points tend to stay in a smaller area, and closer to the starting region; as time progresses, they exhibit a more exploratory behavior with higher divergence, and eventually may reach a convergence close to regions of interest. It is also possible to see how the behavior differs depending on the starting region. More results can be found in the supplementary material.

4.3 Ablation Studies and Model Stability

We also evaluate the contribution of different elements of our model to the final result. For this purpose, we analyze a standard GAN strategy (i.e., using only the discriminative loss), as the baseline. Figure 7 shows how the model is unable to learn both the temporal nature of the scanpaths, and their relation to image features. We also analyze the results obtained by adding a term based on the MSE between the ground truth and the generated scanpath to the loss function, instead of our DTW_{sph} term (we compute a traditional MSE loss, since the only previous generative approach for scanpaths [5] relied on it for their loss term). However, MSE only measures a one-to-one correspondence between points, considering for each time instant a single point, unrelated to the rest. This hinders the learning process, leading to non-plausible results (Figure 7, second row). This behavior is corrected when our DTW_{sph}

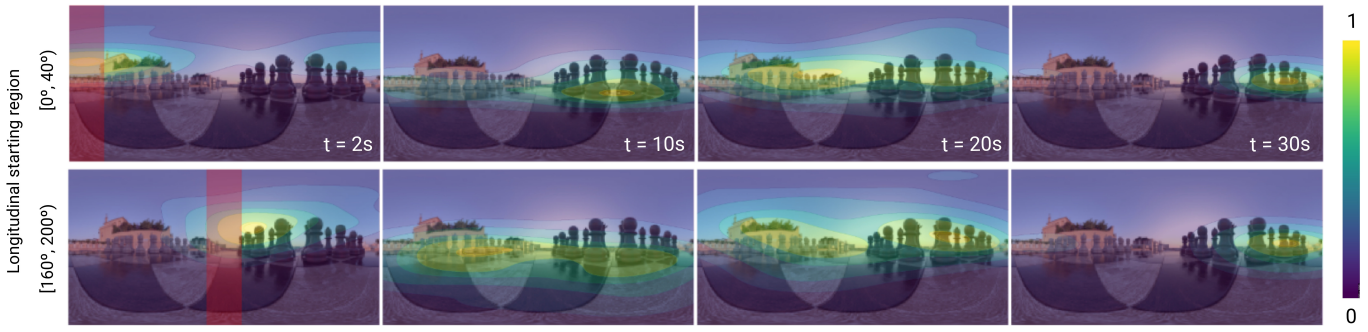

Fig. 6. In order to verify whether our scanpaths are meaningful and realistic, we analyze their exploratory behavior over time, depending on the starting point. For this figure, we take a set of 1,000 scanpaths starting at two different longitudinal ranges (depicted in red in the first image of each row, including ranges from 0° to 20° (i.e., the left-most part of the image), and from 160° to 180° (i.e., the center of the image)). For each of them, we show the kernel density estimation (KDE) of gaze points at four different timestamps $t = \{2, 10, 20, 30\}$ seconds. The scanpaths' behavior is initially more exploratory, albeit linked to the starting region, and eventually converges over time.

Table 2. Quantitative results of our ablation study. Arrows indicate whether higher or lower is better, and boldface highlights the best result for each metric (excluding the ground truth *Human BL*). Please refer to the text for details on the ablated models.

Metric	LEV ↓	DTW ↓	REC ↑	DET ↑
Basic GAN	49.42	2088.44	3.01	1.74
MSE	48.90	1953.21	2.41	1.73
DTW_{sph} (no CoordConv)	47.82	1988.38	3.67	1.99
DTW_{sph} (ours)	**46.19**	**1925.20**	**4.50**	**2.33**
Human Baseline (Human BL)	43.11	1843.72	7.81	4.07

Table 3. Quantitative results of our model for sets of generated scanpaths with different number of samples. Our results are stable regardless of the number of generated samples.

Dataset	# of samples	LEV ↓	DTW ↓	REC ↑	DET ↑
Test set from Sitzmann et al.	100	46.19	1925.20	4.50	2.33
	800	46.10	1916.26	4.75	2.34
	2500	46.15	1921.95	4.82	2.32
	Human BL	43.11	1843.72	7.81	4.07
Rai et al.'s dataset	100	40.95	1548.86	1.91	1.85
	800	40.94	1542.82	1.86	1.86
	2500	40.99	1549.59	1.72	1.87
	Human BL	39.59	1495.55	2.33	2.31

is added instead, since it is specifically targeted for time series data and takes into account the actual spatial structure of the data (Figure 7, third row). The corresponding quantitative measures over our test set from Sitzmann et al. can be found in Table 2. We also analyze the effect of removing the CoordConv layer from our model: Results in Table 2 indicate that the use of CoordConv does have a positive effect on the results, helping learn the transformation from the input to the target domain.

As previously mentioned, our model is stochastic by nature. This means that the scanpaths that it generates for a given scene are always different, simulating observer variability. We have also analyzed whether the metrics reported throughout the paper vary depending on the number of scanpaths generated, to assess the stability of our model. Results can be seen in Table 3, which shows how the metrics remain stable when the number of generated samples decreases, indicating the robustness of our method. Additionally, we have run our training procedure multiple times with different seeds to guarantee its correct reproducibility, and computed the corresponding set of metrics for each of them, obtaining stable results in our evaluations (Table 4).

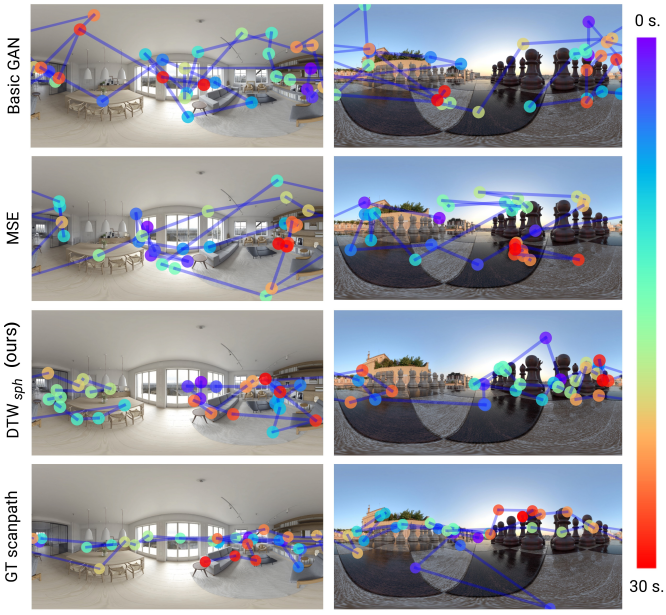

Fig. 7. Qualitative ablation results. *From top to bottom*: basic GAN strategy (baseline); adding MSE to the loss function of the former; our approach; and an example ground truth scanpath. These results illustrate the need for our DTW_{sph} loss term.

Table 4. Quantitative results of our model obtained from averaging five models trained with varying random seeds. The low standard deviations are indicative of the model's stability.

Metric	LEV ↓	DTW ↓	REC ↑	DET ↑
Mean (STD)	46.43 (0.23)	1953.56 (46.23)	4.41 (0.31)	2.39 (0.07)

5 Applications of the Model

Our model is able to generate plausible 30-second scanpaths, drawn from a distribution that mimics the behavior of human observers. As we briefly discuss through the paper, this enables a number of applications, starting with avoiding the need to recruit and measure gaze from high numbers of observers in certain scenarios. We show here two applications of our model, virtual scene design and scanpath-driven video thumbnail creation for static 360° images, and discuss other potential application scenarios.

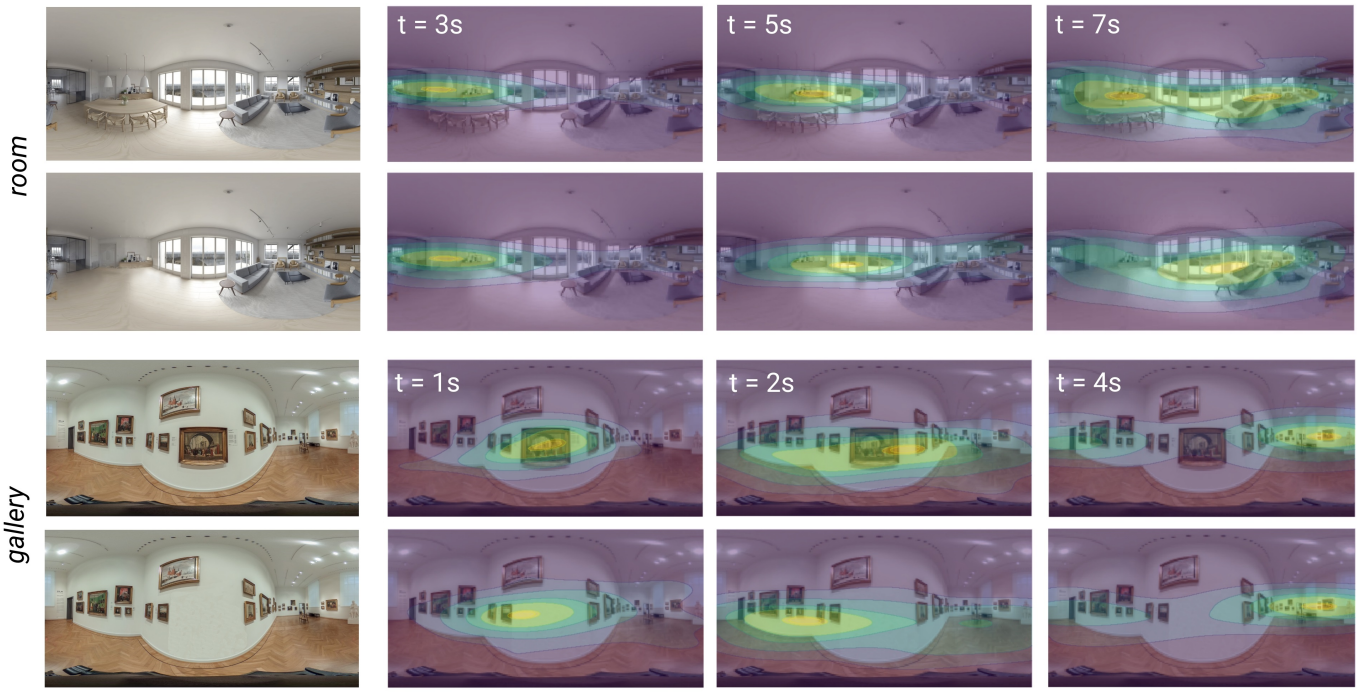

Fig. 8. Our model can be used to aid the design of virtual scenes. We show two examples, each with two possible layouts (original, and removing some significant elements). We generate a large number of scanpaths (virtual observers) starting from the same region, and compute their corresponding probability density function as a function of time, using KDE (see Section 4.2). *room* scene: The presence of the dining table and lamps (top) retains the viewers' attention longer, while in their absence they move faster towards the living room area, performing a more linear exploration. *gallery* scene: When the central picture is present (top), the viewers linger there before splitting to both sides of the scene. In its absence, observers move towards the left, then explore the scene linearly in that direction.

Virtual scene design In an immersive environment, the user has control over the camera when exploring it. This poses a challenge to content creators and designers, who have to learn from experience how to layout the scene to elicit a specific viewing or exploration behavior. Previous works have proposed predicting gaze duration to optimize the placement of visual elements in virtual environments [1]. This is not only a problem in VR, but has also received attention in, e.g., manga composition [10] or web design [45]. However, actually measuring gaze from a high enough number of users to determine optimal layouts can be challenging and time-consuming. While certain goals may require real users, others can make use of our model to generate plausible and realistic generated observers.

As a proof of concept, we have analyzed our model's ability to adapt its behavior to different layouts of a scene (Figure 8). Specifically, we have removed certain elements from a scene, and run our model to analyze whether these changes affect the behavior of our generated scanpaths. We plot the resulting probability density (using KDE, see Section 4.2) as a function of time. The presence of different elements in the scene affects the general viewing behavior, including viewing direction, or time spent on a certain region. These examples are particularly promising if we consider that our model is trained with a relatively small number of generic scenes.

Scanpath-driven video thumbnails of static 360° images 360° images capture the full sphere and are thus unintuitive when projected into a conventional 2D image. To address this problem, a number of approaches have proposed to retarget 360° images or videos to 2D [50, 52, 53]. In the case of images, extracting a representative 2D visualization of the 360° image can be helpful to provide a thumbnail of it, for example as a preview on a social media platform, but these thumbnails are static. The Ken Burns effect can be used to animate static images by panning and zooming a cropping window over a static image, producing more informative and engaging thumbnails. In the context of 360°, however, it seems unclear what the trajectory of such a moving window would be.

To address this question, we leverage our generated scanpaths to drive a Ken Burns–like video thumbnail of a static panorama. For this purpose, we use an average scanpath, computed as the probability density of several generated scanpaths using KDE (see Section 4.2), as the trajectory of the virtual camera. Specifically, KDE allows us to find the point of highest probability, along with its variance, of all generated scanpaths at any point in time. Note that this point is not necessarily the average of the scanpaths. We use the time-varying center point as the center of our 2D viewport, and its variance to drive the FOV or zoom of the moving viewport.

Figure 9 shows several representative steps of this process for two different scenes (*chess* and *street*), while full videos of several scenes are included in the supplementary video. The generated Ken Burns–style panorama previews resemble a human observer exploring these panoramas, and provide a very intuitive preview of the complex scenes they depict.

Other applications Our model has the potential to enable other applications beyond what we have shown in this section. One such example is *gaze simulation for virtual avatars*. When displaying or interacting with virtual characters, eye gaze is one of the most critical, yet most difficult, aspects to simulate [47]. Accurately simulating gaze behavior not only aids in conveying realism, but can also provide additional information such as signalling interest, aiding the conversation through non-verbal cues, facilitating turn-taking in multi-party conversations, or indicating attentiveness, among others. Given an avatar immersed within a virtual scene, generating plausible scanpaths conditioned by a 360° image of their environment could be an efficient, affordable way of driving the avatar's gaze behavior in a realistic manner.

Another potential application of our model is its use for *gaze-contingent rendering*. These approaches have been proposed to save rendering time and bandwidth in VR systems or drive the user's ac-

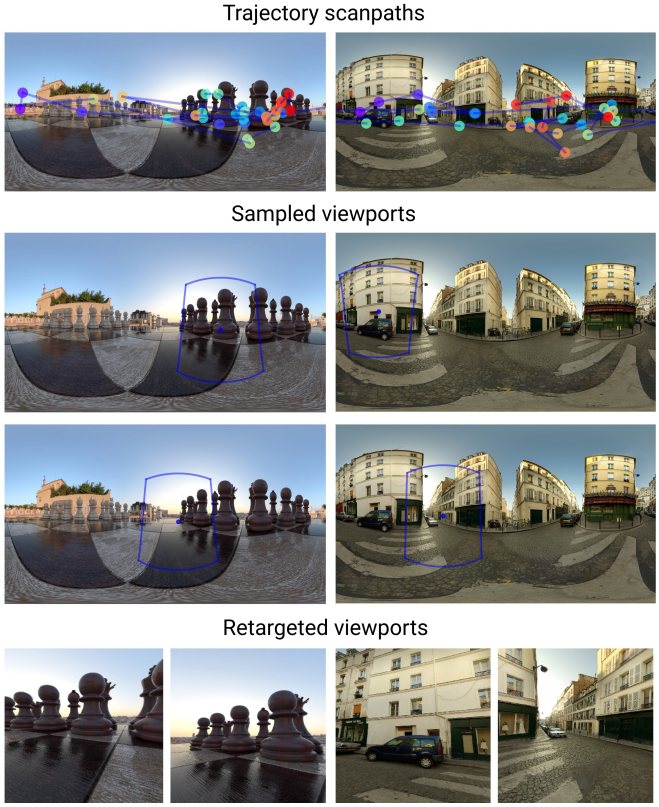

Fig. 9. Scanpath-driven video thumbnails of 360° images. We propose a technique to generate these videos that results in relevant and intuitive explorations of the 360° scenes. Top row: Points of highest probability at each time instant, displayed as scanpaths. These are used as a guiding trajectory for the virtual camera. Middle rows: Two viewports from the guiding trajectory, corresponding to the temporal window with lowest variance. Bottom row: 2D images retargeted from those viewports. Please refer to the text for details.

commodation. Eye trackers are required for these applications, but they are often too slow, making computationally efficient approaches for predicting gaze trajectories or landing positions important [3]. Our method for generating scanpaths could not only help prototype and evaluate such systems in simulation, without the need for a physical eye tracker and actual users, but also in optimizing their latency and performance during runtime.

6 Conclusion

In summary, we propose ScanGAN360, a conditional GAN approach to generating gaze scanpaths for immersive virtual environments. Our unique parameterization tailored to panoramic content, coupled with our novel usage of a DTW loss function, allow our model to generate scanpaths of significantly higher quality and duration than previous approaches. We further explore applications of our model: Please refer to the supplementary material for a description and examples of these.

Our GAN approach is well suited for the problem of scanpath generation: A *single* ground truth scanpath does not exist, yet real scanpaths follow certain patterns that are difficult to model explicitly but that are automatically learned by our approach. Note that our model is also very fast and can produce about 1,000 scanpaths per second. This may be a crucial capability for interactive applications: our model can generate *virtual observers* in real time.

Limitations and future work Our model is trained with 30-second long scanpaths, sampled at 1 Hz. Although this is significantly longer than most previous approaches [16, 30, 34, 54], exploring different or variable lengths or sampling rates remains an interesting avenue for future work.When training our model, we focus on learning higher-level aspects of visual behavior, and we do not explicitly enforce low-level ocular movements (e.g., fixations or saccades). Currently, our relatively low sampling rate prevents us from modeling very fast dynamic phenomena, such as saccades. Yet, fixation patterns naturally emerge in our results, and future work could explicitly take low-level oculomotor aspects of visual search into account.

We have focused on static images, but our parameterization and loss function are tailored to general 360° content, so future work could build on our framework and adapt it to tackle dynamic content (360° videos). In a similar spirit, a DTW-based loss function could also be applied to scanpath generation in conventional 2D images (using an Euclidean distance in 2D instead of our δ_{sph}), potentially leading to better results than current 2D approaches based on mean-squared error.

We have evaluated our model in an entirely new dataset, yielding a performance that shows how our model can generalize. There are still only a few datasets of human gaze in 360° images, and training our model with larger datasets could improve its performance and the variety of scenes it can handle. Moreover, analyzing the impact of additional modalities, such as audio, could be of interest, and could expand the range of potential applications of the model to new scenarios where multimodality plays an important role in the user's behavior. However, we believe that our work is a timely effort and a first step towards understanding and modeling dynamic aspects of attention in 360° images. We hope that our work will serve as a basis to advance this research, both in virtual reality and in conventional imagery, and extend it to other scenarios, such as dynamic or interactive content, analyzing the influence of the task, including the presence of motion parallax, or exploring multimodal experiences.

Acknowledgments

We thank Diego Gutierrez for the revision of the manuscript. This work has received funding from the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (project CHAMELEON, Grant no. 682080). This project was also supported by a 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation (the BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors). This project was in part supported by NSF award 1839974. Additionally, Daniel Martin was supported by a Gobierno de Aragon (2020-2024) predoctoral grant.

References

[1] R. Alghofaili, M. S. Solah, H. Huang, Y. Sawahata, M. Pomplun, and L.-F. Yu. Optimizing visual element placement via visual attention analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 464–473, 2019. doi: 10.1109/VR.2019.8797816

[2] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015.

[3] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[4] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE ICCV Workshops*, pp. 2331–2338, 2017.

[5] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Pathgan: visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

[6] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Scanpath and saliency prediction on 360 degree images. *Signal Processing: Image Communication*, 69:8–14, 2018.

[7] W. Bao and Z. Chen. Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 404:154 – 164, 2020.

[8] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/, 2019.

[10] Y. Cao, R. W. Lau, and A. B. Chan. Look over here: Attention-directing composition of manga elements. *ACM Trans. Graph.*, 33(4):1–11, 2014.

[11] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *2018 IEEE Int. Conf. on Multim. & Expo Workshops (ICMEW)*, pp. 01–04. IEEE, 2018.

[12] A. Colburn, M. F. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces. Technical report, Citeseer, 2000.

[13] B. Coors, A. Paul Condurache, and A. Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 518–533, 2018.

[14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[15] M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*, 2017.

[16] S. R. Ellis and J. D. Smith. Patterns of statistical dependency in visual scanning. *Eye movements and human information processing*, pp. 221–238, 1985.

[17] R. Fahimi and N. D. Bruce. On metrics for measuring scanpath similarity. *Behavior Research Methods*, pp. 1–20, 2020.

[18] K. Horley, L. M. Williams, C. Gonsalvez, and E. Gordon. Face to face: visual scanpath evidence for abnormal processing of facial expressions in social phobia. *Psychiatry research*, 127(1-2):43–53, 2004.

[19] Z. Hu. Gaze analysis and prediction in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 543–544, 2020. doi: 10.1109/VRW50115.2020.00123

[20] Z. Hu, A. Bulling, S. Li, and G. Wang. Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2681–2690, 2021. doi: 10.1109/TVCG.2021.3067779

[21] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911, 2020. doi: 10.1109/TVCG.2020.2973473

[22] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, 2019. doi: 10.1109/TVCG.2019.2899187

[23] 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.

[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE ICCV*, pp. 2106–2113. IEEE, 2009.

[25] Z. Kapoula. The influence of peripheral preprocessing on oculomotor programming in a scanning task. In *Eye movements and psychological functions*, pp. 101–114. Routledge, 2021.

[26] M. A. Kerkouri, M. Tliba, A. Chetouani, and R. Harba. Salypath: A deep-based architecture for visual attention prediction. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1464–1468. IEEE, 2021.

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. Last updated in arXiv in 2017.

[28] M. Kümmerer and M. Bethge. State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*, 2021.

[29] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[30] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, pp. 251–266, 2013.

[31] O. Le Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152 – 164, 2015.

[32] C. Li, W. Zhang, Y. Liu, and Y. Wang. Very long term field of view prediction for 360-degree video streaming. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 297–302. IEEE, 2019.

[33] S. Ling, J. Gutiérrez, K. Gu, and P. Le Callet. Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos. *IEEE Journal on Emerging and Sel. Topics in Circ. and Sys.*, 9(1):204–216, 2019.

[34] H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin. Semantically-based human scanpath estimation with hmms. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3232–3239, 2013.

[35] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Neural information processing systems*, pp. 9605–9616, 2018.

[36] Y. Lu, W. Zhang, C. Jin, and X. Xue. Learning attention map from images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[37] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Computing Surveys*, 2022. doi: 10.1145/3508361

[38] D. Martin, A. Serrano, and B. Masia. Panoramic convolutions for 360° single-image saliency prediction. In *CVPR Workshop on CV for AR/VR*, 2020.

[39] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[40] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication*, 69:26 – 34, 2018.

[41] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84, 2007.

[42] A. Nguyen, Z. Yan, and K. Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proc. ACM Intern. Conf. on Multimedia*, pp. 1190–1198, 2018.

[43] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. 2018.

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

[45] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan. Directing user attention via visual flow on web designs. *ACM Trans. on Graph.*, 35(6):1–11, 2016.

[46] 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*, pp. 205–210, 2017.

[47] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In *Computer graphics forum*, vol. 34, pp. 299–326. Wiley Online Library, 2015.

[48] M. Sela, P. Xu, J. He, V. Navalpakkam, and D. Lagun. Gazegan-unpaired adversarial image generation for gaze estimation. *arXiv preprint arXiv:1711.09767*, 2017.

[49] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia. Movie editing and cognitive event segmentation in virtual reality video. *ACM Trans. Graph. (SIGGRAPH)*, 36(4), 2017.

[50] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Trans. on Vis. and Comp. Graph.*, 24(4):1633–1642, 2018.

[51] M. Startsev and M. Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Proces.: Image Comm.*, 69:43–52, 2018.

[52] Y.-C. Su and K. Grauman. Making 360 video watchable in 2d: Learning videography for click free viewing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1368–1376. IEEE, 2017.

[53] Y.-C. Su, D. Jayaraman, and K. Grauman. Pano2vid: Automatic cinematography for watching 360° videos. In *Asian Conf. on CV*, pp. 154–171. Springer, 2016.

[54] W. Sun, Z. Chen, and F. Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2101–2118, 2019.

[55] B. W. Tatler and B. T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009.

[56] H. R. Tavakoli, E. Rahtu, and J. Heikkilä. Stochastic bottom–up fixation prediction and saccade generation. *Image and Vision Computing*, 31(9):686–693, 2013.

[57] 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.

[58] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[59] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.

[60] W. Wang and J. Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[61] W. Wang, J. Shen, X. Dong, and A. Borji. Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[62] C. Wu, R. Zhang, Z. Wang, and L. Sun. A spherical convolution approach for learning long term viewport prediction in 360 immersive video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 14003–14040, 2020.

[63] C. Xia, J. Han, F. Qi, and G. Shi. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing*, 28(7):3502–3515, 2019.

[64] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2693–2708, 2019.

[65] Y. Xu, Z. Zhang, and S. Gao. Spherical DNNs and Their Applications in 360º Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[66] C. Yang, L. Zhang, R. Lu, Huchuan, Xiang, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3166–3173. IEEE, 2013.

[67] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology*, 4:917, 2013.

[68] D. Zanca, S. Melacci, and M. Gori. Gravitational laws of focus of attention. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):2983–2995, 2019.

[69] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11:9, 2011.

[70] Y. Zhu, G. Zhai, and X. Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018.