ScanDMM: A Deep Markov Model of Scanpath Prediction for 360° Images

Xiangjie Sui¹, Yuming Fang¹*, Hanwei Zhu², Shiqi Wang², Zhou Wang³
¹Jiangxi University of Finance and Economics,
²City University of Hong Kong, ³University of Waterloo
https://github.com/xiangjieSui/ScanDMM

Abstract

Scanpath prediction for 360° images aims to produce dynamic gaze behaviors based on the human visual perception mechanism. Most existing scanpath prediction methods for 360° images do not give a complete treatment of the time-dependency when predicting human scanpath, resulting in inferior performance and poor generalizability. In this paper, we present a scanpath prediction method for 360° images by designing a novel Deep Markov Model (DMM) architecture, namely ScanDMM. We propose a semantics-guided transition function to learn the nonlinear dynamics of time-dependent attentional landscape. Moreover, a state initialization strategy is proposed by considering the starting point of viewing, enabling the model to learn the dynamics with the correct “launcher”. We further demonstrate that our model achieves state-of-the-art performance on four 360° image databases, and exhibit its generalizability by presenting two applications of applying scanpath prediction models to other visual tasks – saliency detection and image quality assessment, expecting to provide profound insights into these fields.

1. Introduction

360° images, also referred to as omnidirectional, sphere or virtual reality (VR) images, have been a popular type of visual data in many applications, providing us with immersive experiences. Nevertheless, how people explore virtual environments in 360° images has not been well understood. The scanpath prediction model that aims at generating realistic gaze trajectories has obtained increasing attention due to its significant influence in understanding users’ viewing behaviors in VR scenes, as well as in developing VR rendering, display, compression, and transmission [51].

Scanpath prediction has been explored for many years in 2D images [29]. However, 360° images are different greatly from 2D images, as a larger space is offered to interact with humans – humans are allowed to use both head and gaze movements to explore viewports of interest in the scene. In such a case, viewing conditions, e.g., the starting point of viewing, has an important impact on humans’ scanpaths [20,21,52], and leads to complex and varied scanpaths among humans. This is inherently different from what happens in 2D visuals since humans can directly guide their attention to the regions of interest. Therefore, scanpath prediction for 360° images is a more complex task.

Current 360° image scanpath prediction methods could...
be roughly divided into two categories: saliency-based [2, 70, 71] and generative methods [42, 43]. The basic idea of the former one is to sample predicted gaze points from a saliency map. The performance of such methods is highly dependent on that of the saliency maps. Furthermore, constructing a satisfactory sampling strategy to account for time-dependent visual behavior is non-trivial—the results of SaliNet [2] exhibit unstable behavior with large displacements and scarce focal regions (see Fig. 1). The latter group of methods utilizes the advance of generative models, e.g., Generative Adversarial Network (GAN), to predict realistic scanpaths. However, such methods show less attention to regions of interests (see Fig. 1). In addition, the GAN-based methods are less flexible in determining the length of scanpaths and commonly suffer from unstable training.

None of above-mentioned studies give a complete treatment of the time-dependency of viewing behavior, which is critical for modeling dynamic gaze behaviors in 360° images. For time-series data, a popular approach is to leverage sequential models, e.g., recurrent neural networks (RNNs), as exemplified in gaze prediction for 360° videos [17, 35, 45]. However, such deterministic models are prone to overfitting, particularly on small 360° databases. More importantly, they typically make simplistic assumptions, e.g., one choice is to concatenate the saliency map to the model’s hidden states [17, 45], which assumes that the network learns how the states evolve by learning from saliency maps. Nevertheless, the neuroscience research [62] reveals that in addition to top-down and bottom-up features, prior history and scene semantics are essential sources for guiding visual attention. Moreover, to be identified as interests or rejected as distractors, items must be compared to target templates held in memory [62]. Inspired by this, we argue that humans’ scanpaths in 360° scenes are complex nonlinear dynamic attentional landscapes over time as a function of interventions of scene semantics on visual working memory. We present a probabilistic approach to learning the visual attention modeling for 360° images. Specifically, we model the mechanism of visual working memory by maintaining and updating the visual state in the Markov chain. Furthermore, a semantics-guided transition function is built to learn the nonlinear dynamics of the states, in which we model the interventions of scene semantics on visual working memory.

We propose a practical strategy to initialize the visual state, facilitating our model to focus on learning the dynamics of states with correct “launcher”, as well as enabling us to assign a specific starting point for scanpath generation. Moreover, ScanDMM is capable of producing 1, 000 variable-length scanpaths within one second, which is critical for real-world applications.

- We apply the proposed ScanDMM to two other computer vision tasks—saliency detection and image quality assessment, which demonstrates our model equips with strong generalizability and is expected to provide insights into other vision tasks.

2. Related Work

2.1. Visual Attention for 2D Images

Modeling Spatial Distribution of Gazes. Visual attention depends on two distinct types of attentional mechanisms: bottom-up and top-down mechanisms [10]. The performance of classic saliency detection methods for 2D images primarily depends on bottom-up features, e.g., colour, luminance, contrast and texture, for modeling visual attention [15, 18, 19, 24]. Although the studies [40, 69] incorporate top-down features, e.g., human faces and texts, there are still obstacles in combining bottom-up and top-down visual features. Deep learning methods [36, 57, 59, 68] take advantage of large-scale databases and well-established convolutional neural networks (CNNs), and they have achieved remarkable success.

Modeling Dynamic Gaze Behaviors. The studies in neuroscience suggest that scanpath generation can be an iterative process: while the eye fixates on an image location, the brain selects the next location to look at [26, 48]. Biologically inspired methods model such mechanism in different ways, e.g., considering the inhibition-of-return mechanism [25, 49, 53], maintaining the residual perceptual information maps [58], modeling the retina transformation [1], and utilizing low-level semantic information [55, 67]. Another class of methods is statistically inspired for modeling dynamic gaze behaviors. A typical approach is to model fixation distributions by the product of the saliency map and the previous fixation location [4, 5, 13, 30, 33, 63]. Some studies regard fixation distributions as Gaussian densities, and model scanpath by leveraging component analysis [54] or hidden markov models [12]. Different from interpretable models, fitting the data using deep generative models can also achieve competitive performance [9, 42, 66].

2.2. Visual Attention for 360° Images

Modeling Spatial Distribution of Gazes. Saliency detection of 360° images is more challenging compared with 2D images, mainly due to complex viewing behaviors and insufficient available data for 360° images. According to the theory that the model is applied to, saliency detection models of 360° images can be grouped into two categories: 2D-
plane-based methods [2, 22, 38] and viewport-based methods [8, 34, 44, 64]. The former models visual attention by using hand-craft features [22, 38] or data-driven features [2] extracted from the distorted (caused by the equiradial projection) 2D plane, while the latter is designed by applying well-established 2D visual saliency detection methods on viewpoint plane with less distortions.

**Modeling Dynamic Gaze Behaviors.** Current scanpath prediction models of 360° images can be classified into two types: saliency-based models [2, 3, 70, 71] and generative models [42, 43]. The former first produces the saliency map by extracting low-level and high-level features [70, 71] or learned features from data [2, 3]. Then, scanpaths are sampled from the saliency map by maximizing information gains [70, 71] or using a stochastic approach [2, 3]. The latter group of methods models scanpaths by learning from human data. PathGAN [42] is a generative model originally developed for 2D images, and it is fine-tuned on Salient360! [47] database to be applied on 360° images. However, the generated scanpaths lose sphere properties (e.g., longitudinal continuities) due to the assumption that a 360° image is similar to a traditional 2D image. To address this problem, ScanGAN [43] propose to learn the image and coordinate representations by using spherical convolutional neural network (S-CNN) [11] and CoordConv layer [39], respectively. Besides, a loss function that measures the spherical distance is proposed for model training.

3. **The Proposed Method**

3.1. **Problem Definition**

In 360° environment, a human scanpath could be defined as a time series of gaze points \( x_{1:T} = (x_1, x_2, \ldots, x_T) \in \mathbb{R}^3 \times T \), where \( x_t \) is a three-dimensional coordinate \((x_1, y_1, z_1)\). Given a 360° image, a scanpath prediction model aims at producing realistic scanpaths \( x_{1:T} \) based on the human visual perception mechanism. This paper presents a probabilistic method – ScanDMM, for scanpath prediction for 360° images, as shown in Fig 2. We model the mechanism of visual working memory by maintaining a set of visual states \( z_{1:T} = (z_1, z_2, \ldots, z_T) \in \mathbb{R}^n \times T \) that encodes the dynamic attentional landscapes over time \((n \text{ is the dimension of the state space})\). Basically, the ScanDMM predicts a scanpath using the following generative process:

\[
\text{Transition : } z_t \sim p_{\theta_t}(z_t | z_{t-1}), \\
\text{Emission : } x_t \sim p_{\phi_{\theta_t}}(x_t | z_t), \quad t = 1, 2, \ldots, T
\]

(1) (2)

where \( \sim \) means the sample operation. \( p_{\theta_t}(z_t | z_{t-1}) \) denotes the transition probabilities between the visual states, and \( p_{\phi_{\theta_t}}(x_t | z_t) \) denotes the emission probabilities that describe how the visual state generates a gaze point. Let \( \theta \) denotes all parameters involved in the ScanDMM. We approximate the realistic scanpaths by maximizing the log-likelihood function \( \log p(x) \) (we omit subscripts here for simplicity):

\[
\max \quad \log p_\theta(x) = \log \mathbb{E}_{p_\theta(z|x)p_\theta(z)} \\
= \log \mathbb{E}_{p_\theta(z|x)} \frac{p_\theta(x, z)}{p_\theta(z|x)} \\
= \log \mathbb{E}_{p_\theta(z|x)} \frac{p_\theta(x | z)p_\theta(z)}{p_\theta(z|x)}.
\]

(3)

The posterior \( p_\theta(z|x) \) can be approximated by using Variational Inference [61], which seeks to derive a distribution \( q_\phi(z|x) \) that is parameterized by neural networks, s.t. \( q_\phi(z|x) \approx p_\theta(z|x) \). Based on Jensen’s inequality, \( f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)] \) where \( f \) is a concave function, we adopt the Evidence Lower Bound of \( \log p_\theta(x) \) as the loss function \( \mathcal{L}(\theta; \phi; x) \) to be maximized:

\[
\log p_\theta(x) \geq \mathbb{E}_{q_\phi} \log \frac{p_\theta(x | z)p_\theta(z)}{q_\phi(z|x)} \\
= \mathbb{E}_{q_\phi} [\log p_\theta(x | z) + \log p_\theta(z) - \log q_\phi(z|x)] \\
= \mathbb{E}_{q_\phi} [\log p_\theta(x | z)] - \text{KL}(q_\phi(z|x) \parallel p_\theta(z|x))
\]

(4)

where \( \text{KL}(\cdot \parallel \cdot) \) denotes the Kullback–Leibler (KL) divergence. The reconstruction term evaluates the model’s accuracy and the regularization term enforces the closeness between \( q_\phi(z|x) \) and \( p_\theta(z) \). The loss function also can be interpreted as the summation of transition \( p_\theta(z|x) \), emission \( p_\theta(x | z) \) and inference \( q_\phi(z|x) \). In the following section, we describe how we design the three functions tailored for scanpath prediction for 360° images.

3.2. **State Initialization**

Different from the common strategy that simply sets the initial state to zero vector [28] or random vector [27], we
propose a practical strategy considering the starting points of scanpaths. The motivation derives from recent studies \cite{20,21,52}, which reveal that the starting point of view has an important influence on the scanpaths. To make our model better focus on learning the dynamics of visual states with the correct “launcher” rather than random scratch, in the training stage, we directly use the starting point $x_1$ to initialize $z_0$:

$$z_0 = F(\hat{z}_0, x_1),$$

(5)

where $\hat{z}_0$ is a learnable parameter. $F$ denotes linear neural networks (see Fig. 2 for details). An advantage of such configuration is that we can assign a specific starting point for scanpath generation, which is flexible and even crucial in some visual tasks, e.g., image quality assessment (see Section 5.2). Notably, to fairly compare ScanDMM with other scanpath prediction models, in the model evaluation, we randomly sample the starting points from an Equator Bias Map that covers the whole longitude and 20% latitude.

### 3.3. Transition Function

Transition function $p_{\theta}(z_t|z_{t-1})$ controls the dynamics of the visual states, in which the history state $z_{t-1}$ functions as visual working memory that maintains the history attentional landscape. Inspired by the recent study in neuroscience \cite{62}, we propose to learn the dynamics guided by scene semantics. Due to the space-varying distortions inherent to equirectangular projections, we choose to extract scene semantics $\hat{s} \in \mathbb{R}^{n \times 1}$ using S-CNN \cite{11}. Moreover, by considering the spatial locations of semantic features are critical for scanpath prediction, we utilize the CoordConv strategy \cite{39} to give convolutions access to their own input coordinates, as suggested in \cite{43}. Given the scene semantics $\hat{s}$ and the history visual state $z_{t-1}$, transition function samples $z_t$ from the following Gaussian densities:

$$z_t \sim \mathcal{N}(\mu^t, \sigma^t),$$

(6)

where $\mu^t \in \mathbb{R}^{n \times 1}$ and $\sigma^t \in \mathbb{R}^{n \times 1}$ are the mean and scale of Gaussian densities describing an attentional landscape of the visual state $z_t$.

Here, we present how we produce the two Gaussian parameters. Firstly, we compute the new underlying gaze distributions $\hat{z}_t$ by

$$\hat{z}_t = \mathbf{W}_z^t(z_{t-1} \oplus \hat{s}) + b_z^t,$$

(7)

where $\oplus$ is the concatenate operation, $\mathbf{W}_z^t \in \mathbb{R}^{2n \times n}$ and $b_z^t \in \mathbb{R}^{n \times 1}$ are learnable parameters of the neural networks. Inspired by the Long Short-Term Memory \cite{23}, we introduce the uncertainty weighting for adaptively determining how much components of the previous visual state $z_{t-1}$ to be updated:

$$\alpha_t = \sigma(\mathbf{W}_\alpha^t z_{t-1} + b^\alpha),$$

(8)

where $\alpha_t$ is an uncertainty weighting determined by the history state $z_{t-1}$, and $\sigma$ denotes the sigmoid function. $\mathbf{W}^t_{\{\mu, \sigma\}} \in \mathbb{R}^{n \times n}$ and $b^t_{\{\mu, \sigma\}} \in \mathbb{R}^{n \times 1}$ are learnable parameters of the neural networks. Finally, the two Gaussian parameters, $\mu^t$ and $\sigma^t$, are computed by:

$$\mu^t = \alpha_t \hat{z}_t + (1 - \alpha_t) z_{t-1},$$

(9)

$$\sigma^t = \log(1 + \exp(\mathbf{W}_\sigma^t \hat{z}_t + b^\sigma)).$$

(10)

### 3.4. Emission Function

Emission function $p_{\theta}(x_t|z_t)$ describes how a gaze point $\tilde{x}_t$ is generated from the visual state $z_t$. As a gaze point is parameterized as a three-dimensional coordinate, we model the emission process by sampling $\tilde{x}_t$ from the three-dimensional Gaussian densities that are parameterized by $\mu^t$ and $\sigma^t \in \mathbb{R}^{3 \times 1}$:

$$\tilde{x}_t \sim \mathcal{N}(\mu^t, \sigma^t),$$

(11)

where $\mu^t$ and $\sigma^t$ are determined by the current visual state $z_t$:

$$\mu^t = 2\sigma(\mathbf{W}_\mu^t z_t + b^\mu) - 1,$$

(12)

$$\sigma^t = \log(1 + \exp(\mathbf{W}_\sigma^t z_t + b^\sigma)).$$

(13)

$\mathbf{W}^e_{\{\mu, \sigma\}} \in \mathbb{R}^{n \times 3}$ and $b^e_{\{\mu, \sigma\}} \in \mathbb{R}^{3 \times 1}$ are learnable parameters of the neural networks.

### 3.5. Inference

Inference for a latent state can be made using information from past and future true observations \cite{28}, i.e., GT gaze points in this study. In other words, the goal of inference is to provide an approximation to the exact posterior $p_{\theta}(z_{1:T}|x_{1:T})$. Here, we leverage the Variational Inference \cite{61} to approximate $p_{\theta}(z_{1:T}|x_{1:T})$ with a tractable family of conditional distributions $q_{\phi}(z_{1:T}|x_{1:T})$:

$$q_{\phi}(z_{1:T}|x_{1:T}) = \prod_{t=1}^T q_{\phi}(z_t|z_{t-1}, \mathcal{H}(x_{1:T}))$$

(14)

s.t. $q_{\phi}(z_t|z_{t-1}, \mathcal{H}(x_{1:T})) \sim \mathcal{N}(\mu^t, \sigma^t)$,

where $\mathcal{H}(\cdot) \in \mathbb{R}^{m \times 1}$ is the variational parameter computed by mapping a variable-length sequences $x_{t:T}$ to its $m$-dimensional space. We implement this by using RNN \cite{28}. The Gaussian parameters $\mu^t$ and $\sigma^t \in \mathbb{R}^{n \times 1}$ are produced by the following process:

$$C_t = \frac{1}{2}((\mathbf{W}_C^t z_t + b_C^t) + \mathcal{H}(x_{t:T})),$$

(15)

$$\mu^t = \mathbf{W}_\mu^t C_t + b^\mu,$$

(16)

$$\sigma^t = \log(1 + \exp(\mathbf{W}_\sigma^t C_t + b^\sigma)).$$

(17)

where $C_t$ denotes the combined feature of the previous state $z_{t-1}$ and the RNN hidden state $\mathcal{H}(x_{t:T})$. $\mathbf{W}^t_c \in \mathbb{R}^{n \times m}$, $\mathbf{W}^t_{\{\mu, \sigma\}} \in \mathbb{R}^{n \times n}$, $b^t_c \in \mathbb{R}^{n \times 1}$, and $b^t_{\{\mu, \sigma\}} \in \mathbb{R}^{n \times 1}$ are learnable parameters of the neural networks.
4. Experiments

In this section, we present the experimental results to validate the effectiveness of the proposed ScanDMM. We refer the interested readers to the supplementary file for more details on ScanDMM, datasets, and qualitative comparisons.

4.1. Dataset

We use four 360° image databases to conduct experiments, including Sitzmann [51], Salient360! [47], AOI [65], and JUFE [20]. Sitzmann [51] database contains 22 images and 1,980 scanpaths, 19 of 22 images are used for training, and the remaining are used for validation. Salient360! [47] database contains two sets: training and benchmark sets. In this paper, we consistently use its training set that contains 85 images and 3,036 scanpaths for evaluation since its benchmark set is not publicly available. AOI [65] database contains 600 high-resolution 360° images and 18,000 scanpaths. JUFE [20] database is constructed for studying 360° image quality assessment, containing 1,032 images and 30,960 scanpaths. For the latter two databases, we randomly select 20% images for model evaluation. To obtain Ground-Truth (GT) scanpaths for model training and evaluation, we sample the raw scanpaths at 1 Hz on each database. Noting that the scanpaths of the AOI database had been processed by the authors, so we did not make any post-processing.

4.2. Experimental Setup

Implementation Details. We train the proposed model using Sitzmann [51] database. To augment data, we longitudinally shift the 360° images and adjust their scanpaths accordingly [43], which yields $19 \times 6 = 114$ images for training. Before the panoramic images are fed into the network, they are resized as $(128, 256)$. The learning rate is set to $3 \times 10^{-4}$ and is decreased by a factor of 0.99998 at each epoch. The dimensions of the visual state and RNN hidden state are set to 100 and 600, respectively. The network was trained for 500 epochs.

Evaluation Metrics. We use three metrics to evaluate the performance of 360° scanpath prediction models, including the Levenshtein distance (LEV), dynamic time warping (DTW), and recurrence measure (REC), as suggested in [16, 43]. Generally, lower LEV/DTW values and higher REC values mean better prediction performance. For an image with $N$ GT scanpaths $\mathbf{x}^{1:N}$, we generate $\hat{N}=N$ fake scanpaths $\mathbf{x}^{1:\hat{N}}$ for comparison. The lengths of generated scanpaths are equal to that of the viewing time $t$. To compare a set of GT scanpaths $\mathbf{x}^{1:N}$ to another set of predicted scanpaths $\mathbf{x}^{1:\hat{N}}$, taking LEV as an example, we choose to compute each metric for all possible pairwise comparisons and average the results:

$$m\text{LEV} = \frac{1}{N\hat{N}} \sum_{i=1}^{N} \sum_{j=1}^{\hat{N}} \text{LEV}(\mathbf{x}^i, \mathbf{x}^j).$$  \hspace{1cm} (18)

To reduce the evaluation bias caused by the randomly generated scanpaths, we test each generative model with 10 times and average these 10 results to obtain the final performance. To make the results more comparable and interpretable, for each metric, we compute the human consistency [43] as a realistic upper bound for model performance (see Human in Tab. 1). Specifically, we compare each GT scanpath against all the other ones and average the results. We also compare to the results of a chance model – produce scanpaths by adding a random Brownian motion to the previous positions (see Random walk in Tab. 1).

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>LEV ↓</th>
<th>DTW ↓</th>
<th>REC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitzmann [51]</td>
<td>Random walk</td>
<td>48.942</td>
<td>2322.987</td>
<td>2.669</td>
</tr>
<tr>
<td></td>
<td>ScanGAN [43]</td>
<td>45.270</td>
<td>1951.848</td>
<td>3.241</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>44.996</td>
<td>1965.427</td>
<td>3.475</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>41.158</td>
<td>1836.986</td>
<td>6.345</td>
</tr>
<tr>
<td>Salient360! [47]</td>
<td>Random walk</td>
<td>40.802</td>
<td>2231.681</td>
<td>2.744</td>
</tr>
<tr>
<td></td>
<td>ScanGAN [43]</td>
<td>38.932</td>
<td>1721.711</td>
<td>3.099</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>37.272</td>
<td>1528.502</td>
<td>3.576</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>35.084</td>
<td>1382.590</td>
<td>5.202</td>
</tr>
<tr>
<td>AOI [65]</td>
<td>Random walk</td>
<td>13.699</td>
<td>711.516</td>
<td>2.993</td>
</tr>
<tr>
<td></td>
<td>SaltiNet [2]</td>
<td>14.695</td>
<td>596.544</td>
<td>2.244</td>
</tr>
<tr>
<td></td>
<td>ScanGAN [43]</td>
<td>12.889</td>
<td>552.446</td>
<td>3.750</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>12.127</td>
<td>537.504</td>
<td>4.024</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>9.243</td>
<td>389.477</td>
<td>6.228</td>
</tr>
<tr>
<td></td>
<td>ScanGAN [43]</td>
<td>24.209</td>
<td>1094.978</td>
<td>3.075</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>23.091</td>
<td>1086.014</td>
<td>4.329</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>18.306</td>
<td>1038.880</td>
<td>7.745</td>
</tr>
</tbody>
</table>

Table 1. Performance of scanpath prediction models on different databases. The best performance is highlighted.

4.3. Performance Comparison

Model Selection. We compare ScanDMM to four state-of-the-art scanpath prediction models: CLE [5], DeepGaze III [30], SaltiNet [2], and ScanGAN [43]. The first two models are built for traditional images, and the latter two are tailored for 360° images. For DeepGaze III and SaltiNet with larger sizes, we choose to use the pre-trained weights provided by the corresponding authors to avoid overfitting.
Figure 3. Qualitative comparison to different scanpath prediction models for four different scenes. From left to right: scanpath samples obtained by a observer, scanpath results obtained by CLE [4], DeepGaze III [30], SaltiNet [2], ScanGAN [43], and the proposed model. From top to bottom: the Room from Sitzmann [51] database, the Museum from Salient360! [47] database, the Party from AOI [65] database, and the Park from JUFE [20] database.

Table 2. Efficiency comparison in terms of the model size and running time (computed by the time cost of producing 1,000 scanpaths). Noting that CLE is a traditional model that built on saliency maps, we only count the time to produce the scanpaths.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLE [1]</td>
<td>-</td>
<td>≈ 39 seconds</td>
</tr>
<tr>
<td>DeepGaze III [30]</td>
<td>78.9MB</td>
<td>≈ 11 minutes</td>
</tr>
<tr>
<td>SaltiNet [2]</td>
<td>103.6MB</td>
<td>≈ 40 minutes</td>
</tr>
<tr>
<td>ScanGAN [43]</td>
<td>33.9MB</td>
<td>0.987 seconds</td>
</tr>
<tr>
<td>ScanDMM</td>
<td>18.7MB</td>
<td>0.737 seconds</td>
</tr>
</tbody>
</table>

Table 3. Ablation study. The effectiveness of our designs is well justified.

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>LEV ↓</th>
<th>DTW ↓</th>
<th>REC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitzmann [51]</td>
<td>ScanDMM(I)</td>
<td>45.949</td>
<td>1996.456</td>
<td>3.091</td>
</tr>
<tr>
<td></td>
<td>ScanDMM(S)</td>
<td>46.800</td>
<td>2098.412</td>
<td>2.833</td>
</tr>
<tr>
<td></td>
<td>ScanDMM(IS)</td>
<td>46.489</td>
<td>2011.157</td>
<td>2.904</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>44.966</td>
<td>1965.427</td>
<td>3.475</td>
</tr>
<tr>
<td>Salient360! [47]</td>
<td>ScanDMM(I)</td>
<td>38.340</td>
<td>1698.517</td>
<td>3.168</td>
</tr>
<tr>
<td></td>
<td>ScanDMM(S)</td>
<td>39.194</td>
<td>1787.674</td>
<td>2.744</td>
</tr>
<tr>
<td></td>
<td>ScanDMM(IS)</td>
<td>38.958</td>
<td>1720.894</td>
<td>2.853</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>37.272</td>
<td>1528.592</td>
<td>3.576</td>
</tr>
</tbody>
</table>

4.4. Ablation Study

We conduct ablation experiments to analyze the impact of the state initialization strategy and the scene semantics on the proposed model. We compare the proposed ScanDMM with three baselines: (1) ScanDMM(I), initializing state directly using learnable parameter: \( z_0 = \hat{z}_0 \). (2) ScanDMM(S), removing scene semantics \( \hat{s} \) in Eq. 7. (3) ScanDMM(IS), performing both of the two modifications. All baseline models are trained in the same settings, and the results are shown in Table 3. We can observe that the model achieves the best performance when incorporating both the state initialization strategy and the scene semantics.

5. Applications

In this section, we demonstrate the generalizability of our model by applying it to two downstream applications – saliency detection and quality assessment for 360° images.
Performance Comparison. We compare different saliency detection models on Sitzmann [51] and Salient360! [47] databases. In addition to ScanGAN [43], we also compare the proposed ScanDMM with 5 saliency detection models tailored for 360° images: BMS360 [34], GBVS360 [34], SaltiNet [8], SalNet360 [44], and ScanGAN [43]. The results are shown in Table 4 and Fig. 4. Table 4 shows that the proposed ScanDMM achieves the best performance on Salient360! database, while its performance drops apparently on Sitzmann database. One reason might be that scanpath-to-saliency models are worse at suppressing non-salient regions than the models tailored for 360° images (please refer to the supplementary file for the complete comparison), e.g., referring to Robat in Fig. 4. The results show that applying the scanpath prediction model for the saliency detection task is encouraging. Nevertheless, there is significant room for improvement, as evidenced by a large performance gap between computational models and humans.

5.2. Quality Assessment for 360° Images

The image quality assessment (IQA) model aims at predicting the perceived quality of visual images [41]. Recent studies [20,21,52] reveal that the viewing behaviors may be indispensable for quality assessment for 360° scenes, highlighting the importance of scanpath prediction in 360° IQA. In this section, we show how much gain scanpath prediction models can provide for 360° IQA. Specifically, we leverage the computational framework [52], which offers us a natural way to utilize scanpaths for 360° IQA. Basically, there have 4 steps: (1) Extracting the sequences of rectilinear projections of viewports along different generated scanpaths. (2) Computing the “frame-level” (i.e., viewport-level) quality scores by using existing 2D IQA models. (3) Computing

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>AUC</th>
<th>NSS</th>
<th>CC</th>
<th>KLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitzmann</td>
<td>BMS360 [34]</td>
<td>0.831</td>
<td>1.328</td>
<td>0.666</td>
<td>0.475</td>
</tr>
<tr>
<td></td>
<td>GBVS360 [34]</td>
<td>0.845</td>
<td>1.433</td>
<td>0.693</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>SaltiNet [2]</td>
<td>0.762</td>
<td>0.948</td>
<td>0.497</td>
<td>0.671</td>
</tr>
<tr>
<td></td>
<td>SalNet360 [44]</td>
<td>0.841</td>
<td>1.467</td>
<td>0.718</td>
<td>0.421</td>
</tr>
<tr>
<td></td>
<td>ScanGAN360 [8]</td>
<td>0.813</td>
<td>1.188</td>
<td>0.612</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>ScanDMM</td>
<td>0.767</td>
<td>0.958</td>
<td>0.521</td>
<td>0.609</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>0.877</td>
<td>2.265</td>
<td>1.000</td>
<td>0.117</td>
</tr>
</tbody>
</table>

| Salient360! | BMS360 [34] | 0.748 | 0.907 | 0.567 | 0.476 |
|             | GBVS360 [34] | 0.712 | 0.775 | 0.484 | 0.576 |
|             | SaltiNet [2] | 0.737 | 0.842 | 0.530 | 0.514 |
|             | SalNet360 [44] | 0.744 | 0.989 | 0.602 | 0.476 |
|             | ScanGAN360 [8] | 0.759 | 0.975 | 0.613 | 0.495 |
|             | ScanDMM | 0.728 | 0.774 | 0.491 | 0.556 |
|             | Human | 0.860 | 2.395 | 1.000 | 0.290 |

Table 4. Performance of saliency prediction models on Sitzmann [51] and Salient360! [47] databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>CC</th>
<th>SRCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitzmann</td>
<td>PSNR</td>
<td>0.156</td>
<td>0.016</td>
<td>0.787</td>
</tr>
<tr>
<td></td>
<td>PSNRGAN</td>
<td>0.135</td>
<td>0.014</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td>PSNRDMM</td>
<td>0.563</td>
<td>0.546</td>
<td>0.659</td>
</tr>
<tr>
<td></td>
<td>PSNRhuman</td>
<td>0.585</td>
<td>0.583</td>
<td>0.646</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.148</td>
<td>0.046</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>SSIMGAN</td>
<td>0.162</td>
<td>0.046</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td>SSIMDMM</td>
<td>0.319</td>
<td>0.309</td>
<td>0.681</td>
</tr>
<tr>
<td></td>
<td>SSIMhuman</td>
<td>0.527</td>
<td>0.532</td>
<td>0.677</td>
</tr>
<tr>
<td></td>
<td>VIF</td>
<td>0.163</td>
<td>0.096</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>VIFGAN</td>
<td>0.147</td>
<td>0.092</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>VIFDMM</td>
<td>0.507</td>
<td>0.572</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>VIFhuman</td>
<td>0.616</td>
<td>0.601</td>
<td>0.628</td>
</tr>
<tr>
<td></td>
<td>DISTS</td>
<td>0.162</td>
<td>0.081</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>DISTSGAN</td>
<td>0.176</td>
<td>0.100</td>
<td>0.784</td>
</tr>
<tr>
<td></td>
<td>DISTSDMM</td>
<td>0.662</td>
<td>0.675</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>DISTShuman</td>
<td>0.700</td>
<td>0.711</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>DeepWSD</td>
<td>0.160</td>
<td>0.044</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>DeepWSDGAN</td>
<td>0.162</td>
<td>0.072</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>DeepWSDDMM</td>
<td>0.635</td>
<td>0.628</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>DeepWSDhuman</td>
<td>0.668</td>
<td>0.667</td>
<td>0.593</td>
</tr>
</tbody>
</table>


5.1. Saliency Detection for 360° images

Visual attention prediction, commonly known as saliency detection, is the task of inferring the objects or regions that attract human attention in a scene. Intuitively, an ideal scanpath prediction model should be able to model spatially visual attention. In this section, we apply the proposed ScanDMM and ScanGAN [43] to the saliency detection task by generating 1,000 scanpaths and post-processing these gaze points to yield continuous saliency maps. Specifically, we convolve fixation maps, the two-dimensional records of the locations of all the gaze points, with a modified Gaussian [56]:

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

where $\sigma_x = \frac{\sigma_x}{\cos\theta}$, $\sigma_y = 19^\circ$ is a constant value, and $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the latitude of the gaze point.

Evaluation Metrics. We evaluate saliency detection models using 4 metrics: the Judd variant of the area under curve (AUC) [6], normalized scanpath saliency (NSS) [46], correlation coefficient (CC) [32], and Kullback–Leibler divergence (KLD) [31]. Similarly, for each metric, we compute the human consistency as a realistic upper bound for model performance [7], referring to Human in Table 4. Specifically, we first compare the fixations of two groups of $M$ observers, where $M$ varies from 1 to $N/2$ (i.e., half of the total $N$ observers). We then fit the $N/2$ performance scores to a power function (i.e., $aM^b + c$), and predict the human performance as that of two groups of infinite observers (which is equal to $c$, for $b < 0$).

Performance Comparison. We compare different saliency detection models on Sitzmann [51] and Salient360! [47] databases.
perceived quality of a “video” (i.e., a viewport sequence) by temporally pooling the frame-level quality scores. (4) Averaging the quality scores of all videos to obtain the final perceived quality of the 360° image.

**Evaluation Metrics.** We use three evaluation metrics to quantify the quality prediction performance, including CC, Spearman’s rank-order correlation coefficient (SRCC) and root mean square error (RMSE). Generally, higher CC/SRCC values and lower RMSE values mean better performance.

**Performance Comparison.** We conduct comparison experiments on JUFE [20] database, in which each image has four quality labels corresponding to four different viewing conditions (2 starting point × 2 exploration time). Considering this database is goal-directed (i.e., IQA), we retrain the ScanDMM and ScanGAN [43] on 80% training set. We generate $N_\text{fake}$ fake scanpaths for extracting viewport sequences for each image. For frame-level quality calculation, we employ five 2D IQA models, including the Peak Signal-to-Noise Ratio (PSNR), the Structural SIMilarity index (SSIM) [60], the Visual information fidelity (VIF) [50], the Deep Image Structure and Texture Similarity metrics (DISTs) [14], and the DeepWSD [37]. Then, the Gaussian weighting function [52] is chosen to temporally pool the frame-level quality scores. We add “GAN”, “DMM”, and “Human” to the five IQA methods as the subscripts to name the quality model that uses the scanpaths generated by ScanGAN [43], ScanDMM, and human, respectively. We also directly apply 2D IQA models to equirectangular projections as baselines. The results are shown in Table 5, where we can observe that all baseline models fail to predict the perceived quality, which is reasonable since they produce one quality score for a given image that has four quality labels. Conversely, our ScanDMM can simulate the viewing behaviors under the specific viewing condition thanks to the state initialization strategy, which significantly improves the performance of 2D IQA models and have competitive performance compared with Human models. ScanGAN [43], however, is not applicable due to the uncontrollable starting points. Current work only tests traditional 2D IQA models; future models that take better account for VR-specific distortions will have great potential to boost the performance.

6. Conclusion and Discussion

This paper presents a scanpath prediction model for 360° images and demonstrates its crucial capabilities, i.e., high accuracy, efficiency and generalizability, for real-world applications. Moreover, we provide a novel perspective for saliency detection and quality assessment in 360° scenes – constructing models by simulating humans’ viewing behaviors, which is an extremely natural way if we recall how we obtain the GT labels. Nevertheless, there is still much to explore, as we only simply exploit generated scanpaths without elaborately designing for a specific task. While this paper primarily focuses on static 360° images, our ScanDMM can be easily adapted to dynamic 360° videos by feeding dynamic scene semantics to the network. Moreover, as a probabilistic approach, ScanDMM is able to learn the general pattern of the viewing behaviors in 360° scenes, thus could also be applied to VR transmission, automatic cinematography, navigation, virtual scene design, etc.

**Acknowledgments**

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100601, in part by the National Natural Science Foundation of China under Grant 62132006, and in part by the Natural Science Foundation of Jiangxi Province of China under Grant 20223AE191002.
References


