Supplementary Material of Pathformer3D: A 3D Scanpath Transformer for 360° Images

Rong Quan[1], Yantao Lai[1,2], Mengyu Qiu[2], and Dong Liang[2,3] (✉)

[1] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, the Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing, 211106, China
rongquan0806@gmail.com, yantaolai@nuaa.edu.cn
[2] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
{qmengyu,liangdong}@nuaa.edu.cn
[3] Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen, China

In this document, we provide additional experiments, visualizations, details and insights regarding our Pathformer3D model. The specific sections of this document are listed below.

- We discuss a few detail about the components of Pathformer3D(Sec. 1).
- We discuss the coordinate discontinuity issue of 2D equirectangular projection(Sec. 2).
- We show more images of qualitative compare(Sec. 3).
- We provide more information about ablation experiments(Sec. 4).
- We report the ablation results on another two datasets(Sec. 5).
- We report the qualitative performance results of the ablation experiments(Sec. 6).
- We provide a study of the characteristics of center offset(Sec. 7).
- We provide a study of the characteristics of time saliency(Sec. 8).

1 Additional Architecture Details

Here, we provide additional details regarding the model architecture that were not mentioned in the main text due to space constraints.

1.1 3D Contextual Visual Feature Representation

3D feature extraction In our model, we attempted to use a network model called SphereNet [3] to extract features from 360° images. The SphereNet architecture employs a convolution method called SphereConv, which projects the 360° image onto a sphere and then performs convolution operations. After each SphereConv operation, BatchNorm and LeakyReLU operations are applied, with the LeakyReLU's negative slope set to -0.2. We used a total of 3 SphereConv operations, and the resulting feature maps of different dimensions were upsampled and concatenated to obtain the final feature map. To illustrate this point, we show the architecture of the SphereNet we used in Fig. 1.

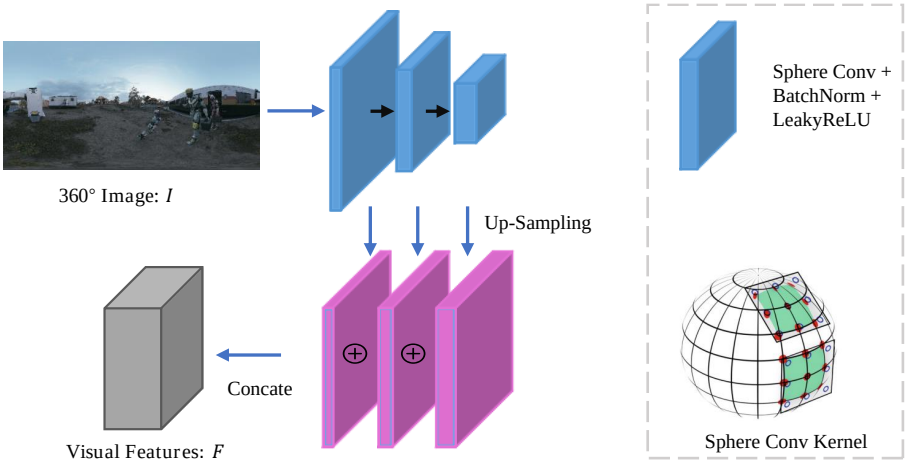


Fig. 1: The architecture of SphereNet.

3D feature encoder The 3D feature encoder is constituted by a singular linear embedding layer, an addition of a 3D position encoding, and $M = 4$ uniquely stacked layers of the standard Transformer encoder. Each of these encoder layers include elements of self-attention, layer normalization, and two consecutive linear transformations, consequently yielding an output tensor that bears the same shape as the input tensor.

1.2 Fixation Sequence Generation

Fixation decoder The Fixation decoder is constructed of a linear embedding, an addition operation introducing 1D position encoding, along with $N = 4$ stacked Transformer decoder layers. Each decoder layer includes masked self-attention, cross-attention, and two consecutive linear transformations, eventually producing an output tensor of the same size as the joint query vector Q'_0. The self-attention layer computes the attention scores between current and all historical fixations, while the cross-attention layer is responsible for computing the attention scores between the fixations and the output of the 3D Transformer Encoder.

Fixation generation We have employed a 3D Mixture Density Network to predict fixation distribution parameters, where the probability distribution is derived from $K = 5$ Gaussian distributions. When predicting parameters with the 3D Mixture Density Network, we utilized multiple independent Multi-Layer Perceptions (MLP), each containing two consecutive linear operations. We applied a ReLu function as an activation function following the first linear operation.

When predicting the weight (π), we incorporated a softmax function after the second linear layer to ensure the sum of all 3D Gaussian distributions equals

to 1. When predicting means (u), a Tanh operation is used after the second Linear operation, constraining the predicted position within a range of $[-1, 1]$. We decompose the operation of predicting the covariance matrix ($\sum$) into predicting the related variance (σ) and correlation coefficient (ρ). When predicting σ, we used the exp function after the second Linear operation to ensure that the variance is larger than 1. As for ρ, we used the Tanh function to make the correlation coefficient between $[-1, 1]$.

2 The Coordinate Discontinuity of 2D Equirectangular Projection

The coordinates of fixations in the 2D Equirectangular Projection of 360° images can be represented in two ways. They can be given by 2D coordinates $p = (Y, X)$,$Y \in [0, 1], X \in [0, 1]$, or latitude and longitude $p = (\phi, \lambda)$,$\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \lambda \in [-\pi, \pi]$. However, both representation methods for depicting scanpath in 360° images encounter the issue of discontinuity. For instance, two points with the same latitude but different longitudes of $\lambda = -180°$ and $\lambda = 180°$ actually represent the same position. However, in the above two coordinate systems, they represent two different positions which are far away from each other.

To tackle the above problems, we represent fixations in a 3D Cartesian coordinate system, where each position is given in the form of $p = (x, y, z)$. By adopting this 3D coordinate system, we can effectively solve the discontinuity issue of the coordinate systems used in the 2D equirectangular projection. In addition, the representations in the three coordinate systems can be flexibly transformed using the following formulas.

$$\phi = (Y - 0.5)\pi; \lambda = (X - 0.5)2\pi \tag{1}$$

$$x = \cos(\phi)\cos(\lambda); y = \cos(\phi)\sin(\lambda); z = \sin(\phi) \tag{2}$$

3 More Qualitative Comparison Results

We show more qualitative comparison results in Fig. 2. As we can see, our method can generate scanpaths that are closest to the ground truths.

4 More Details about Alation Studies

In the main text, due to space constraints, we did not specifically describe the structure of each model. In this section, we will further supplement some additional details about the ablation experiment models.

4.1 Pure ViT

In this structure, we executed scanpath prediction in spherical coordinate system, and adopted a method similar to ViT [4] to extract image features. Specifically, we first resize the 2D Equirectangular Projection of 360° image to 128 ×

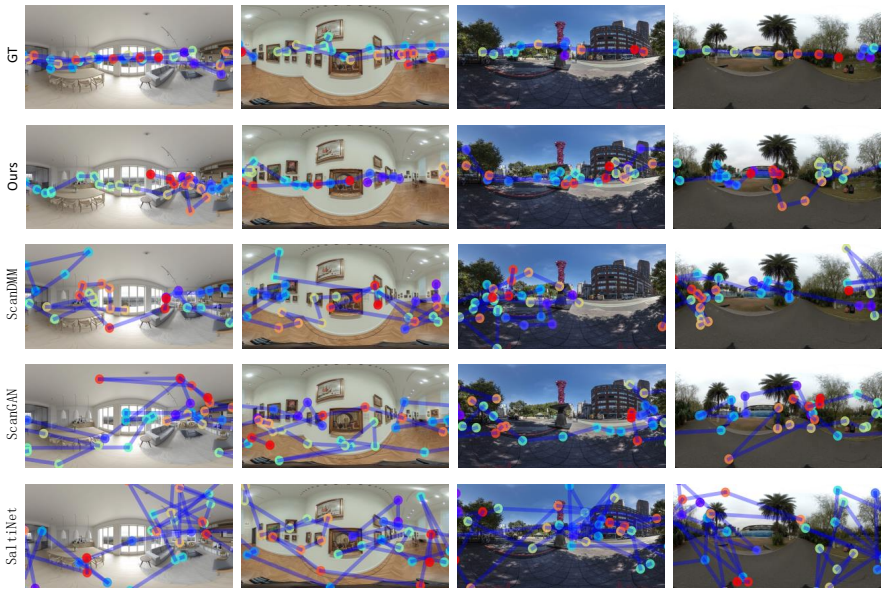


Fig. 2: More Qualitative results. The four images from left to right are sourced from Sitzmann [11], Salient360! [9], AOI [13], and JUFE [5], respectively. From top to bottom, the images represent the perspectives of Ground Truth, Our model, ScanDMM [12], ScanGAN [6], and SaltiNet [1].

256×3 (height, width, depth) with a block size of 8×8. Next, we unroll each small feature block to obtain a feature map with dimensions of $H = 128/8 = 16, W = 256/8 = 32, C = 8 \times 8 \times 3 = 192$. We then directly transform this feature sequence to the encoder's internal dimension of $D = 128$ using the Linear Embedding layer of Transformer Encoder. The remaining operations remain unchanged.

4.2 Pure 2D CNN

We executed scanpath prediction also in spherical coordinate system and replaced the SphereConv in SphereNet [3] with Conv2d. After each convolution operation, the size of the feature map is halved, and the dimension of the feature map is doubled. Other operations such as batch normalization and LeakyRelu from SphereNet are kept unchanged.

4.3 Saliency

In this structure, we follow VSPT [8] to use saliency information to help predict the scanpaths. EPSNet [14] is a recently proposed method for generating saliency maps for 360° images. The structure of its generator is the same as the SalGan [7] used in VSPT [8], which principally utilizes VGG16 [10] for feature extraction

and then employs multiple upsampling operations to generate a saliency map. To adopt the same method as VSPT [8], we extracted two feature maps with dimensions of 512 and 64 generated in the upsampling process, then stacked them together along the channel level, followed by an average pooling operation using an 8×8 convolution kernel, and then flattening the downscaled feature map. This results in a feature vector with $L = (160/8) \times (320/8) = 800, C = 512 + 64 = 576$. Then, the Embedding operation of the Transformer Encoder maps this image sequence to the D=128 dimension, with all other operations remaining unchanged.

4.4 w/o EncoderLayer

In this structure, $M = 4$ Transformer encoder layers are removed and the output of the Embedding operation in the 3D feature encoder is directly fed into the Fixation Decoder.

4.5 w/o MDN + MSE LOSS

In this structure, we used the same decoder structure as in VPT360 [2], namely removing the MDN and directly predicting the 3D coordinates of the next point through a linear operation after the Scanpath Decoder. We also adopted the same loss function as theirs.

In the training procedure, mean square error (MSE) is used as a loss function due to its simplicity to measure the distance between the predicted sequence and the ground truth sequence. This model uses two types of loss for its loss function: the speed loss of each element and the distance loss between the predicted and actual values. The motion velocity v is the root mean square value of the position in current moment and the position in the last moment, where $V = \sqrt{(p_t - p_{t-1})^2} = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2 + (z_t - z_{t-1})^2}$. Our loss function is then defined as the combination of position MSE and motion velocity MSE as:

$$L = \alpha MSE(\hat{p}, p) + \beta MSE(\hat{v}, v) \tag{3}$$

where $\hat{p}$ denotes the motion velocity in predicted position $\hat{p}$ and v denotes the motion velocity in ground truth position p. The hyper-parameters α and β are used to balance the scale of two loss components. We set $(\alpha, \beta) = (0.75, 0.25)$ in this work.

5 Ablation Results on Another Two Datasets

In the ablation experiment in the main body of this paper, due to space constraints, we only showed the ablation experiment results on Salient360! [9] and JUFE [5] datasets. Here, in Tab. 1, we have listed the ablation experiment results on the Sitzmann [11] and AOI [13] datasets. From this, we can see that our model still maintains a commendable advantage when compared to other ablation experiment models.

Table 1: Ablation study on Sitzmann [11] and AOI [13] dataset. The best result for each metric is highlighted in bold, while the second-best result is underlined.

Method	Sitzmann						AOI					
	LEV ↓	DTW ↓	TDE ↓	ScanMatch ↑	REC ↑	SS ↑	LEV ↓	DTW ↓	TDE ↓	ScanMatch ↑	REC ↑	SS ↑
Pure ViT	45.334	1997.981	<u>19.554</u>	0.476	3.331	0.270	14.113	581.322	32.131	0.373	7.571	<u>0.235</u>
Pure 2D CNN	45.173	2023.814	20.739	0.482	3.54	0.271	14.088	575.113	32.660	0.374	7.753	0.229
Saliency	48.046	2078.388	22.549	0.421	2.304	0.261	14.236	582.077	32.221	0.367	7.646	0.221
w/o Feature Encoder	45.062	2018.401	19.790	0.468	3.845	0.274	14.026	587.800	31.766	**0.378**	8.052	0.230
w/o MDN + MSE Loss	**43.439**	2135.792	26.321	0.445	3.002	0.125	14.822	595.722	<u>31.040</u>	0.333	<u>8.104</u>	0.209
K=1	45.819	1951.336	20.511	0.474	3.375	0.277	14.296	585.208	31.755	0.366	7.389	0.226
K=3	<u>44.049</u>	2117.464	23.581	<u>0.490</u>	**4.056**	0.261	14.096	593.456	32.338	0.372	7.889	0.204
K=8	46.167	1950.403	19.688	0.472	3.154	0.266	14.308	577.013	31.692	0.366	7.451	0.219
Parallel	44.766	**1884.471**	22.236	**0.501**	3.809	**0.285**	<u>13.990</u>	**531.416**	31.746	0.373	7.969	**0.278**
Ours	44.837	<u>1903.825</u>	**19.273**	0.472	<u>3.927</u>	<u>0.280</u>	**13.904**	<u>561.895</u>	**30.987**	<u>0.376</u>	**8.286**	0.225

6 Qualitative Performance Results of The Ablation Experiments

We qualitatively presented the scanpaths of the most representative models in our ablation experiments, and the results are shown in Fig. 3. From this we can see that different structures have some unique characteristics.

Although the 'parallel' model tends to have fixations predominantly located within salient regions, there is a significant issue of large displacements between fixations due to not treating this task as a sequential task. The 'w/o MDN + MSE Loss' model can't generate realistic scanpaths, failing to focus well on significant areas, and the distance between fixations is too close. The 'w/o Encoder' model can also generate more realistic scanpaths to a certain extent, but it can be seen from quantitative results that more realistic scanpaths can often be obtained by adding encoder. The fixations generated by 'saliency' models do not exhibit a strong concentration around the central region in dimension 0. These models often rely heavily on saliency information extraction models and tend to have larger parameter sizes and longer runtime compared to other methods.

7 Center Offset

In 360° images, human fixations often exhibit a stronger focus on the central region in dimension 0. To verify this characteristic, we generated a scanpath for the images used for evaluation in Sitzmann [11] and salient360! [9] datasets. We then attempted to compute saliency maps using all the generated scanpaths. We also applied the same processing to the scanpaths of ScanDMM [12], ScanGAN [6], and real human gaze data.

For the fixation point distribution, we expand latitude and longitude to a 2D coordinate system, where the horizontal coordinate is longitude and the range is $[-180, 180]$, and the vertical coordinate is latitude and the range is $[-90, 90]$. The distribution of fixations is obtained from the fixations on all pictures in the data set through Gaussian blurring. As shown in Fig. 4a and Fig. 4b, it can be

Fig. 3: Qualitative performance results of the ablation experiments. From left to right. The first two columns of images are from the Sitzmann [11] dataset, while the following three rows are from the Salient360! [9] dataset. From top to bottom, the scanpaths correspond to the Ground Truth, Ours, Parallel,w/o MDN + MSE Loss, w/o Encoder, Saliency model.

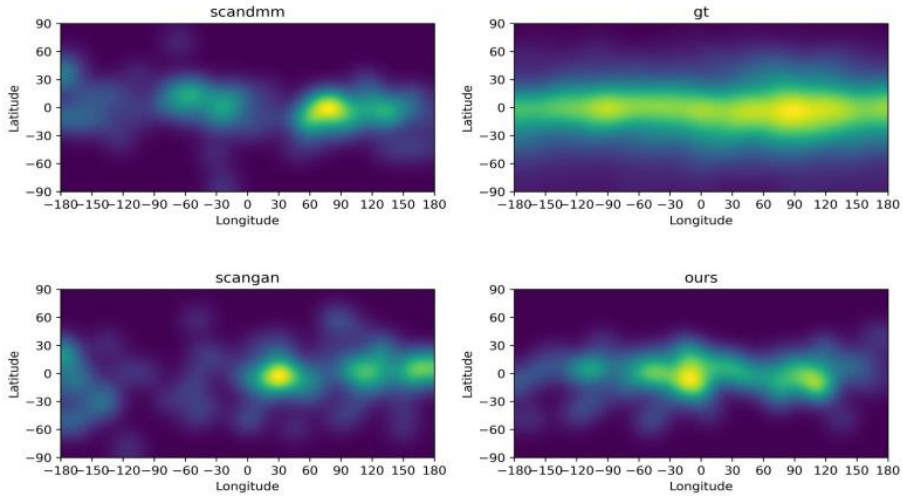
.

seen that the fixations of real human beings are mostly distributed around the region where the latitude is 0 and are evenly distributed in longitude. Our results, similar to ScanDMM [12] and ScanGAN [6], also exhibit this characteristic.

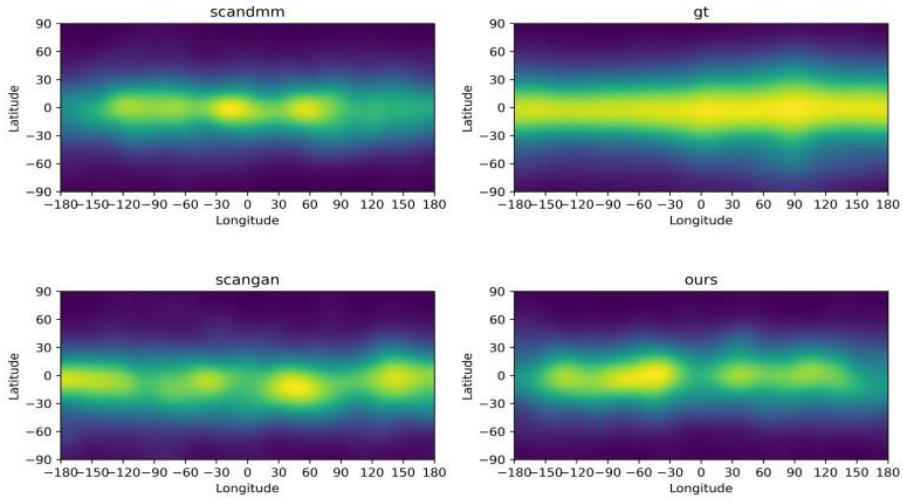
8 Time Saliency

We conducted a study on the relationship between image saliency and time. For each image, we generated 1000 scanpaths with a length of 30 fixations. We divided the time into six intervals with a 5-second interval, resulting in saliency maps for these six time intervals.

It is evident that the overall attention is distributed on informative objects. Fig. 5 illustrates the attention visualization results of our model at different time steps. Moreover, when attention is allocated to a particular object, it undergoes a process of initially staying focused for a duration, gradually converging, and then dispersing. For example, in the first 15 seconds, the attention on the chess piece on the right gradually intensifies, and after 15 seconds, the attention shifts to other regions. This attention pattern predicted by our model is valuable for guiding various visual tasks and cognitive research.



(a) Center Offset from Sitzmann



(b) Center Offset from Salient360!

Fig. 4: Qualitative results of Center Offset

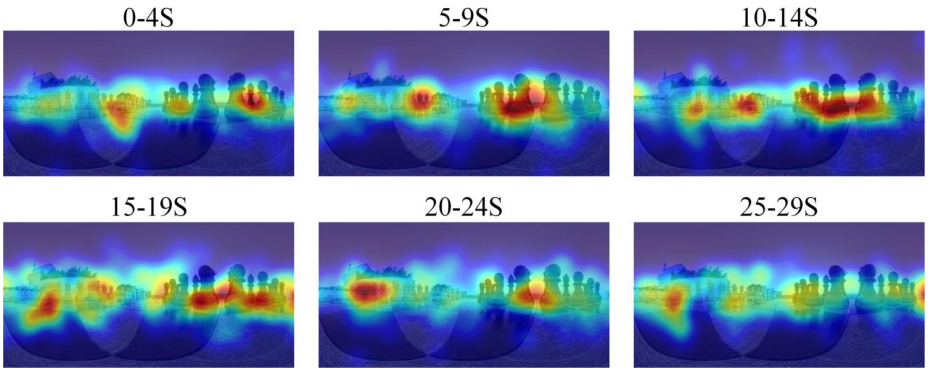


Fig. 5: Time Saliency

.

References

1. Assens Reina, M., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2331–2338 (2017)
2. Chao, F.Y., Ozcinar, C., Smolic, A.: Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In: MMSP. pp. 1–6 (2021)
3. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European conference on computer vision (ECCV). pp. 518–533 (2018)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fang, Y., Huang, L., Yan, J., Liu, X., Liu, Y.: Perceptual quality assessment of omnidirectional images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 580–588 (2022)
6. Martin, D., Serrano, A., Bergman, A.W., Wetzstein, G., Masia, B.: Scangan360: A generative model of realistic scanpaths for 360 images. IEEE Transactions on Visualization and Computer Graphics **28**(5), 2003–2013 (2022)
7. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
8. Qiu, M., Rong, Q., Liang, D., Tu, H.: Visual scanpath transformer: Guiding computers to see the world. In: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 223–232. IEEE (2023)
9. Rai, Y., Gutiérrez, J., Le Callet, P.: 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)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in vr: How do people explore virtual environments? IEEE transactions on visualization and computer graphics **24**(4), 1633–1642 (2018)
12. Sui, X., Fang, Y., Zhu, H., Wang, S., Wang, Z.: Scandmm: A deep markov model of scanpath prediction for 360∘ images. In: IEEE Conference on Computer Vision and Pattern Recognition (2023)
13. Xu, M., Yang, L., Tao, X., Duan, Y., Wang, Z.: Saliency prediction on omnidirectional image with generative adversarial imitation learning. IEEE Transactions on Image Processing **30**, 2087–2102 (2021)
14. Zou, Z., Ye, M., Li, S., Li, X., Dufaux, F.: 360° image saliency prediction by embedding self-supervised proxy task. IEEE Transactions on Broadcasting (2023)