Spherical Vision Transformer for 360° Video Saliency Prediction

Mert Cokelek[1]
mcokelek21@ku.edu.tr
Nevrez Imamoglu[2]
https://nevrez.github.io
Cagri Ozcinar[3]
https://cagriozcinar.netlify.app
Erkut Erdem[4]
https://web.cs.hacettepe.edu.tr/~erkut
Aykut Erdem[1]
https://aykuterdem.github.io

[1] KUIS AI Center, Koç University
Istanbul, TR
[2] AIST
Tokyo, JP
[3] MSK.AI
London, UK
[4] Hacettepe University
Istanbul, TR

Abstract

The growing interest in omnidirectional videos (ODVs) that capture the full field-of-view (FOV) has elevated the importance of 360° saliency prediction in computer vision. However, predicting where humans look in 360° scenes presents unique challenges, including spherical distortion, high resolution, and limited labelled data. We propose a novel vision-transformer-based model for omnidirectional videos named SalViT360 that leverages tangent image representations. We introduce a spherical geometry-aware spatio-temporal self-attention mechanism that is capable of effective omnidirectional video understanding. Furthermore, we present a consistency-based unsupervised regularisation term for projection-based 360° dense-prediction models to reduce artefacts in the predictions that occur after inverse projection. Our approach is the first to employ tangent images for omnidirectional saliency prediction, and our experimental results on three ODV saliency datasets demonstrate its effectiveness compared to the state-of-the-art. Code and models are available at: https://github.com/MertCokelek/SalViT360

1 Introduction

As an important computer vision task, visual saliency prediction aims at predicting where people look in a scene. It is widely used in various areas, such as *saliency-guided* image and video compression [1, 2, 3, 4, 5], super-resolution [6, 7], and quality assessment [8, 9, 10, 11, 12] to exploit human perceptual features for enhancement. With the growing popularity of virtual reality (VR) applications and multimedia streaming, predicting saliency in 360° videos has received more attention recently. One primary challenge in processing 360° scenes is effectively representing omnidirectional data. Equirectangular Projection (ERP), where the full-FOV scene is projected on a 2D plane, is a common representation due to its computational simplicity. However, ERP suffers from spherical distortion, particularly

towards the poles, which can significantly affect the geometric structure of the scene and degrade model performance. While previous works have proposed kernel transformations [13] and spherical convolutions [14, 15] to minimise this distortion on ERP, these methods come at the cost of computational complexity and the loss of global context in 360° scenes. Cubemap projection [16] is another common approach that addresses the distortion problem by approximately expressing the spherical scene with six faces of a cube. Although this approach eliminates the distortion to an extent, it breaks the continuity of neighbouring faces and introduces discrepancies in the predictions around the edges.

Previous works in 360° saliency prediction primarily focused on addressing this representation problem, with each method trying to balance representative power and computational complexity. Chao et al. [17] employed cubemap projection and fine-tuned the 2D image saliency model SalGAN [18] on each cube face. Cheng et al. [19] extended cubemap projection with cube-padding to address the discontinuities on the face boundaries. Chao et al. [20] extended SalGAN360 with multi-view fusion. Dahou et al. [21] proposed a two-stream architecture to compute global and local saliency in omnidirectional videos. Their approach uses global prediction as a rough attention estimate, and the local stream on cube faces predicts local saliency. Zhang et al. [22] proposed spherical convolutions for saliency prediction. Qiao et al. [23] showed that the eye fixation distribution bias depends on the viewport locations, which motivated us to introduce *spherical position information* into our model. Yun et al [24] use local undistorted patches with deformable CNNs and a ViT variant for self-attention across space and time. Djilali et al. [25] used a self-supervised pre-training based on learning the association between several different views of the same scene and trained a supervised decoder for 360° saliency prediction as a downstream task. Although their approach considers the global relationship between viewports, it ignores the temporal dimension that is crucial for video understanding.

The methods mentioned above share a common limitation in processing 360° data through projections or modified kernels and ignore the entire field-of-view, which is critical for global scene understanding and saliency prediction. Thus, there is a need for an effective omnidirectional data processing method that minimises spherical distortion while preserving the global context and avoiding computational overhead and artefacts introduced by the previous methods. Recently, Eder et al. [26] proposed tangent image representations, which use gnomonic projection to map a spherical image into multiple overlapping patches, where each patch is tangent to the faces of an icosahedron. This method tackles the problem of spherical distortion on the scene. However, to our interest, the dense-prediction models particularly suffer from discrepancies and artefacts on overlapping regions of tangent image patches after inverse projection to ERP. In this work, we propose using tangent images to process undistorted local viewports and develop a transformer-based model to learn their global association for saliency prediction in 360° videos. This is the first work that employs tangent images for omnidirectional saliency prediction, which is also capable of modelling the temporal dimension, motivated by the recent video transformer architectures, e.g., TimeSformer [27].

Proposed method and contributions: We use gnomonic projection to obtain multiple undistorted tangent images, which enables us to extract rich local spatial features using any pre-trained and fixed 2D backbone. Our extensive experiments demonstrate the effectiveness of our proposed SalViT360 model against the state-of-the-art on VR-EyeTracking [28], PVS-HMEM [29], and 360-AV-HM [30] datasets. Our contributions are three-fold:

- *Spherical geometry-aware spatial attention.* We aggregate global information on the sphere by computing self-attention among tangent viewports. We use tangent features

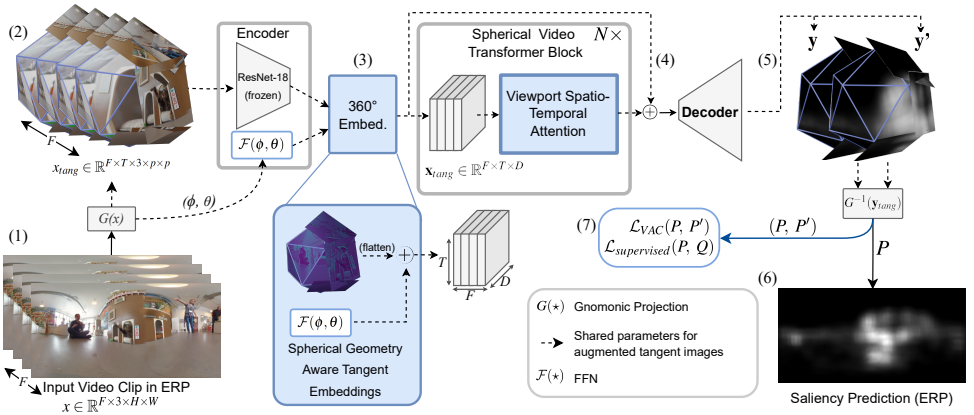

Figure 1: **Overview of the proposed SalViT360 model.** The ERP video clip of F frames (1) is projected to $F \times T$ tangent images per set (2). Each tangent image is encoded and fused with spherical-geometry-aware position embeddings (3) for the 360° video transformer to aggregate global information (4). The outputs are decoded into saliency predictions in tangent space (5), which are projected back to ERP, giving the final saliency map (6). In addition to the supervised loss, the model is trainable with $\mathcal{L}_{VAC}(P, P')$ to minimise the tangent artefacts (7). During test time, the model works with a single tangent set. For simplicity, only one set of tangent images is shown.

as viewport tokens to address the quadratic complexity associated with spatial self-attention. We introduce 360° geometry awareness to the transformer by using learnable spherical position embeddings guided by per-pixel angular coordinates (ϕ, θ), denoting *latitude* and *longitude* angles, respectively. The proposed position embedding method outperforms standard 1D embeddings and can easily be integrated into any transformer architecture designed for 360° images.

- *Viewport Spatio-Temporal Attention (VSTA).* The complexity of joint spatio-temporal self-attention increases quadratically with respect to the number of frames. With VSTA, we optimise this joint computation on tangent viewports from consecutive frames, where the spatio-temporal self-attention is performed in two stages: (1) Viewport Spatial Attention (VSA) and (2) Viewport Temporal Attention (VTA). In VSA, spherical geometry-aware self-attention is computed *intra-frame* level. The temporal information in the videos is encoded by a VTA, among tangent planes that point to the same direction in the *inter-frame* level.

- *Viewport Augmentation Consistency (VAC).* We propose an unsupervised, consistency-based loss for omnidirectional images, minimising the discrepancies in overlapping regions of tangent predictions. The loss is computed between the weight-sharing saliency predictions of two tangent image sets generated with different configurations. This regularisation method is suitable for any projection-based dense-prediction model. Importantly, VAC does not introduce any memory or time overhead during test time, as only one set of predictions is sufficient for inference.

2 Method

In Fig. 1, we present an overview of our proposed model SalViT360. We start with gnomonic projection [26] to obtain tangent images for each frame in the input video clip. These are passed to an encoder-transformer-decoder architecture. The image encoder extracts local features for each tangent viewport and reduces the input dimension for the subsequent self-attention stage. We map the pixel-wise angular coordinates to produce the proposed spherical geometry-aware position embeddings $\mathcal{F}(\phi,\theta)$ for the 360° transformer, enabling better learning of spatial representations. The transformer utilises Viewport Spatio-Temporal Attention (VSTA) to capture inter and intra-frame global information across tangent viewports in a temporal window. The transformed embeddings are then fed into a 2D CNN-based decoder, which predicts saliency on the tangent images. We then apply inverse gnomonic projection on the tangent predictions to obtain the final saliency maps in ERP. We propose an unsupervised consistency-based Viewport Augmentation Consistency Loss to mitigate the discrepancies after inverse gnomonic projection. The learnable parameters of the network are in tangent space, allowing us to use large-scale pre-trained 2D models for feature extraction, while the rest of the network is trained from scratch.

Gnomonic Projection and Encoder. We first project the input ERP clip $x \in \mathbb{R}^{F\times 3\times H\times W}$ to a set of tangent clips $x_{tang} \in \mathbb{R}^{F\times T\times 3\times p\times p}$, where F, 3, H, and W are the number of frames, channel dimension (RGB), height, and width of a given video. The resulting tangent images have a patch size of $p\times p = 224\times 224$ pixels. We pick T, the number of tangent images per frame as 18, with a *FOV* of 80°. We use a ResNet-18 encoder pre-trained on ImageNet and keep it frozen while extracting features. We downsample and flatten the encoder features to obtain tangent feature vectors with dimension $D = 512$. We map the angular coordinates (ϕ,θ) for each pixel of the tangent viewports to the same feature dimension using a 2-layer-mlp and sum these embeddings with encoder features to obtain the proposed *spherical geometry-aware embeddings* $\mathbf{x}_{tang} \in \mathbb{R}^{F\times T\times D}$ that are used in the transformer.

Viewport Spatio-Temporal Attention for 360° videos. While the pre-trained encoder extracts spatial features for each tangent image locally, the global information is essential for 360° scene understanding. We propose a self-attention mechanism for tangent viewport features to achieve this. However, since incorporating the temporal dimension of the videos increases the number of tokens and thus the computational complexity, we approximate spatio-temporal attention with two stages: temporal attention (1) among the same tangent viewports from consecutive F frames and spatial attention (2) among T tangent viewports in the same frame. This reduces the self-attention complexity from $F^2\times T^2$ to F^2+T^2. In this way, we effectively model the global context required for 360° video understanding. Fig. 2 illustrates the proposed *Viewport Spatio-Temporal Attention*,

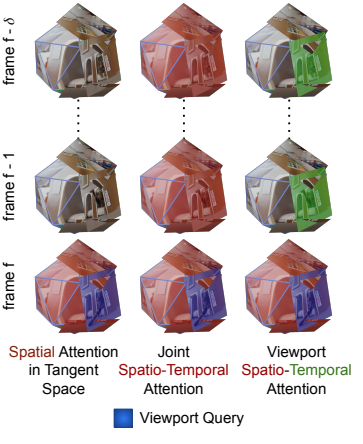

Figure 2: **The proposed VSTA** (*right*), as compared to Viewport Spatial Attention (VSA) (*left*) and Joint Spatio-Temporal Attention (JSTA) (*middle*). Red and Green viewports denote the self-attention neighborhood for each scheme.

which is formally defined as:

$$\begin{aligned}
\mathrm{VSTA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{VSA}(\mathrm{VTA}(\mathbf{z}^{(l)}_{(t,f)})) \\
\mathrm{VTA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{SM}\left(\mathbf{q}^{(l)}_{(t,f)} \cdot \left\{\mathbf{k}^{(l)^\top}_{(t,f')}\right\}\right) \cdot \left\{\mathbf{v}^{(l)}_{(t,f')}\right\}_{f'=1\ldots F} \\
\mathrm{VSA}(\mathbf{z}^{(l)}_{(t,f)}) &= \mathrm{SM}\left(\mathbf{q}^{(l)}_{(t,f)} \cdot \left\{\mathbf{k}^{(l)^\top}_{(t',f)}\right\}\right) \cdot \left\{\mathbf{v}^{(l)}_{(t',f)}\right\}_{t'=1\ldots T}
\end{aligned} \tag{1}$$

where $\mathbf{z}^{(l)}_{(t,f)} \in \mathbb{R}^{1\times D}$ corresponds to the tangent features ($\mathbf{x}_{tang}$) of viewport t in frame f at l-th transformer block, $\mathbf{q},\mathbf{k},\mathbf{v}$ are the query, key, and value projections of $\mathbf{z}$, and SM is the softmax operator. Attention heads and the scale multiplication of dot-product attention are not given for presentation clarity. Our baseline 360° video transformer consists of 6 VSTA blocks with an embedding dimension of 512 and 8 attention heads. We use 1D learnable position embeddings with a window size of $F = 8$ for the temporal position embeddings. Following the approach in [27], we apply temporal self-attention followed by spatial self-attention in alternation.

Decoder. The decoder comprises four upsample layers followed by 3×3 convolutions and normalisation layers. For a set of tangent clips, it takes the skip connection of encoder and transformer features $\mathbf{x}^{f}_{tang} \in \mathbb{R}^{T\times D\times 7\times 7}$ of the last frame as input and outputs saliency prediction $\mathbf{y} \in \mathbb{R}^{T\times 56\times 56}$ on tangent planes. The final ERP saliency maps are obtained by passing the tangent predictions to inverse gnomonic projection. We aggregate global information among tangent images through the transformer; however, the decoder predicts each tangent plane separately.

Viewport Augmentation Consistency (VAC). Our model effectively learns the overall saliency distribution with the supervised loss. However, since each tangent plane is predicted separately, the final ERP saliency map contains discrepancies in overlapping regions of tangent patches. To tackle this issue, we propose an unsupervised loss strategy, called *Viewport Augmentation Consistency* (VAC), for improving the consistency between the saliency predictions P and P' from two tangent projection sets. Specifically, we generate the second tangent set by applying different configurations, such as horizontally shifting the tangent planes on the sphere, using a larger FOV for the same viewports, and varying the number of tangent images at different viewports. We provide the details and comparison of these approaches in the supplementary. VAC does not require any additional memory or time overhead since it uses the shared parameters of the whole model, and the forward pass is done in parallel. Furthermore, since the ERP predictions from the original P and augmented P' tangent sets are expected to be consistent, only one tangent set is sufficient during inference. The VAC loss is defined as:

$$\begin{aligned}
\mathcal{L}_{VAC}(P,\,P') &= \mathcal{L}^{weighted}_{KLD}(P,\,P') \;+\; \mathcal{L}^{weighted}_{CC}(P,\,P'), \\
\mathcal{L}^{weighted}_{KLD}(P,\,P') &= \sum_{i,j} P_{i,j} \log\left(\varepsilon + \frac{P_{i,j}}{P'_{i,j}+\varepsilon}\right) \cdot w_{i,j}, \\
\mathcal{L}^{weighted}_{CC}(P,\,P') &= 1 - \frac{\sum(P\cdot P')\cdot w_{i,j}}{\sum(P\cdot P)\cdot\ \sum(P'\cdot P')}
\end{aligned} \tag{2}$$

where $P,\ P'$ are the saliency predictions from original and augmented viewports, and w is an optional weight matrix obtained from gnomonic projection to weigh the overlapping pixels of gnomonic projection on ERP predictions.

Table 1: **Performance analysis of *SalViT360* against the state-of-the-art 360° saliency models** on VR-EyeTracking, PVS-HMEM, 360AV-HM datasets. While the scores in **bold** highlight the best performance, the underlined ones are the second best.

	VR-EyeTracking [28]				PVS-HMEM [29]				360AV-HM [30]			
Method	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**	**NSS↑**	**KLD↓**	**CC↑**	**SIM↑**
SalGAN360 [17]	1.753	10.845	0.370	0.355	1.513	4.394	0.314	0.291	0.719	25.301	0.065	0.036
CP-360 [19]	0.624	15.338	0.165	0.240	0.576	4.738	0.162	0.198	0.689	24.426	0.061	0.041
MV-SalGAN360 [20]	1.818	8.713	0.382	0.357	1.546	4.112	0.316	0.295	0.716	25.322	0.066	0.036
ATSal [21]	1.317	12.259	0.336	0.318	0.732	4.303	0.183	0.219	0.727	24.141	0.058	0.041
PAVER [24]	1.511	13.267	0.307	0.294	0.750	3.736	0.224	0.269	0.732	23.944	0.065	0.035
Djilali et al. [25]	**3.183**	6.570	0.565	0.475	1.688	2.430	0.447	0.404	1.727	22.889	0.148	0.085
SalViT360 (Ours)	2.630	**5.744**	**0.586**	**0.492**	**2.191**	**1.841**	**0.626**	**0.495**	**1.946**	**22.711**	**0.168**	**0.093**

3 Experiments

3.1 Setup

Datasets and pre-processing. We use the publicly available *VR-EyeTracking* [28] dataset for training, which includes 134 train, 22 validation, and 52 test videos viewed by at least 31 subjects, lasting between $20-60$ seconds. We sampled the videos at 16 fps with a resolution of 960×1920. For cross-dataset evaluation, we use the *PVS-HMEM* [29] and *360AV-HM* [30] datasets, which respectively contain 76 and 21 videos viewed by 58 and 15 subjects. The videos in the *PVS-HMEM* dataset have varying durations between $10-80$ secs, while those in the *360AV-HM* dataset have a duration of 25 secs. The videos in both datasets have a frame rate between 24-60 fps.

Evaluation metrics and loss functions. We evaluate the performance of the models using the four most commonly used saliency evaluation metrics [31]: Normalised Scanpath Saliency (NSS), KL-Divergence (KLD), Correlation Coefficient (CC), and Similarity Metric (SIM). We use a weighted differentiable combination of KLD, CC, and Selective-MSE (MSE on normalised saliency maps at only eye-fixation points [32]) for the supervised loss:

$$\mathcal{L}_{supervised}(P,\, Q_s,\, Q_f) = \mathcal{L}_{KLD}(P,\, Q_s) \,+\, \mathcal{L}_{CC}(P,\, Q_s) \,+\, \alpha \mathcal{L}_{SMSE}(P,\, Q_s,\, Q_f) \qquad (3)$$

where $P,\, Q_s,\, Q_f$ are the predicted saliency, gt density and fixation maps, and $\alpha = 0.005$.

Architecture and optimization details. We train our SalViT360 model using the AdamW optimizer [33] with an initial learning rate of $1e-5$, default weight decay, and momentum parameters of $1e-2$ and $(\beta_1,\, \beta_2) = (0.9,\, 0.999)$, with a batch size of 16 for five epochs with early stopping. We conducted our experiments on a single 32G Tesla V100 GPU.

3.2 Comparison with the State-of-the-art

In Table 1, we present the results of our SalViT360 model and the existing models. We evaluate the performance of SalViT360 with six state-of-the-art models for 360° image and video saliency prediction, namely CP-360 [19], SalGAN360 [17], MV-SalGAN360 [20], Djilali et al. [25], ATSal [21], PAVER [24]. ATSal, PAVER, and CP-360 are video saliency models; the rest are image-based models developed for the omnidirectional domain. On the VR-EyeTracking test set, SalViT360 outperforms the state-of-the-art on three metrics and gives competitive results on NSS. NSS metric penalises false positives and negatives equally, which puts our model at a potential disadvantage in this scenario. Given our model's

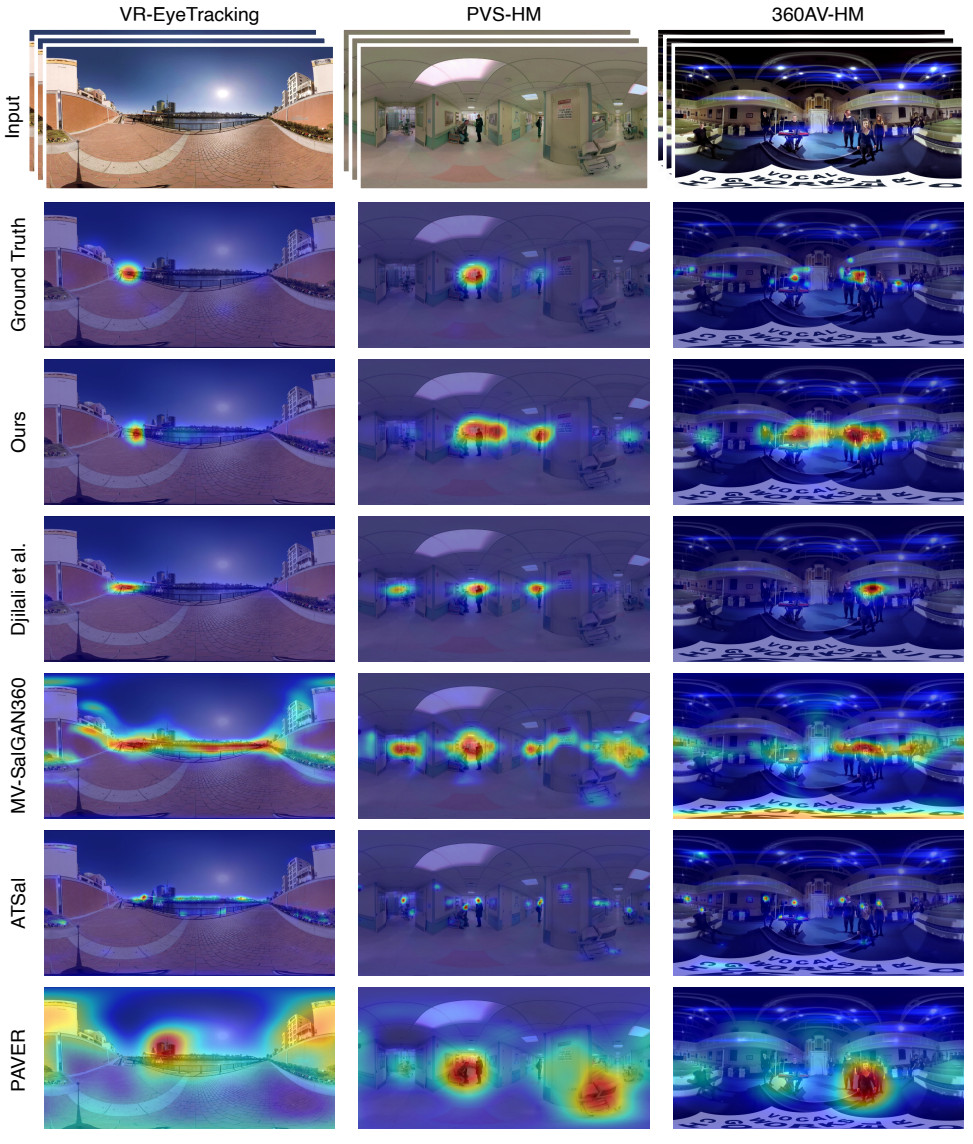

Figure 3: **Qualitative comparison on VR-EyeTracking [28], PVS-HM [29], and 360AV-HM [30] datasets.** Our proposed approach gives better results compared to existing models. The saliency predictions of our model better resemble the ground truth fixation density maps, producing sparse estimates while covering the most dominant modes better.

tendency to provide broader saliency coverage (see Fig. 3), it may introduce false positives at a higher rate than ground-truth density maps. Consequently, this could lead to a slight reduction in NSS scores, despite the model's proficiency in capturing the salient regions. In Table 1, SalViT360 outperforms the state-of-the-art on all metrics on PVS-HMEM and 360-AVHM datasets. These results demonstrate the cross-dataset generalisation capability of our

Table 2: **Ablation study** for each component in our approach on VR-EyeTracking dataset.

Method	# trainable params	NSS↑	KLD↓	CC↑	SIM↑
VSA (*w/ 1D Pos. Emb.*)	18.76M	2.518	6.445	0.560	0.472
+ *Spherical Pos. Emb.*	19.26M	2.575	6.221	0.563	0.475
VSTA (*w/ Sph. Pos. Emb.*)	25.56M	**2.664**	6.174	0.570	0.479
+ *VAC (w/o mask)*	25.56M	2.624	6.011	0.576	0.490
+ *VAC (w/ mask)*	25.56M	<u>2.630</u>	<u>5.744</u>	<u>0.586</u>	<u>0.492</u>
+ *Late-Fusion*	25.56M	2.578	**4.654**	**0.592**	**0.495**

model compared to the existing methods. The qualitative comparison in Fig. 3 also shows the effectiveness of our approach in highlighting the salient regions more accurately.

3.3 Experimental Analysis and Ablation Studies

To assess the contribution of each component of our approach and to provide an in-depth analysis of spatio-temporal modelling, we perform report additional experiments in Table 2 and Table 3. In these experiments, we consider a baseline model comprised of a 2D ResNet-18 backbone, a vision transformer with Viewport Spatial Attention, and a CNN decoder.

Spherical Position Embeddings. We compare the performance of our proposed *spherical geometry-aware spatial position embeddings* with regular 1D learnable position embeddings. The results on all four metrics show that our proposed embedding method outperforms it, demonstrating that it is more suitable for processing spherical data with Vision Transformers.

Viewport Augmentation Consistency and Late-Fusion. We train our VSTA baseline using only the supervised loss. We then compare this single-scale baseline to the one trained with a weighted combination of the supervised loss ($\mathcal{L}_{supervised}$) and the consistency loss ($\mathcal{L}_{VAC}$). Table 2 shows that VAC outperforms the VSTA baseline on three distribution-based saliency evaluation metrics with a performance gain of 3.2% on KLD, 6.4% on CC, and 5.8% on SIM, respectively. Lastly, we investigated the effect of using the predictions of two tangent sets in the final prediction. We perform this with an optional late-fusion as an element-wise multiplication of two ERP predictions to better highlight the consistently predicted salient regions. This simple optional fusion improves the performance on three metrics significantly, with zero memory- and 0.5× time-overhead. We provide sample results for qualitatively comparing these components in Fig. 4.

Spatio-Temporal modelling. We conducted several experiments to evaluate the effectiveness of our proposed Viewport Spatio-Temporal Attention (VSTA) mechanism. In Table 2, we compare Viewport Spatial Attention (VSA) and VSTA blocks to assess the contribution

Table 3: **Spatio-temporal modelling performance** of *SalViT360* compared against two alternative approaches on VR-EyeTracking dataset.

Method	# trainable params	NSS↑	KLD↓	CC↑	SIM↑
VSA + *2+1D-CNN Enc.*	17.31M	2.568	**5.915**	0.568	0.477
VSA + *Offline EMA*	19.27M	2.591	6.018	0.566	0.477
VSTA	25.56M	**2.664**	6.174	**0.570**	**0.479**

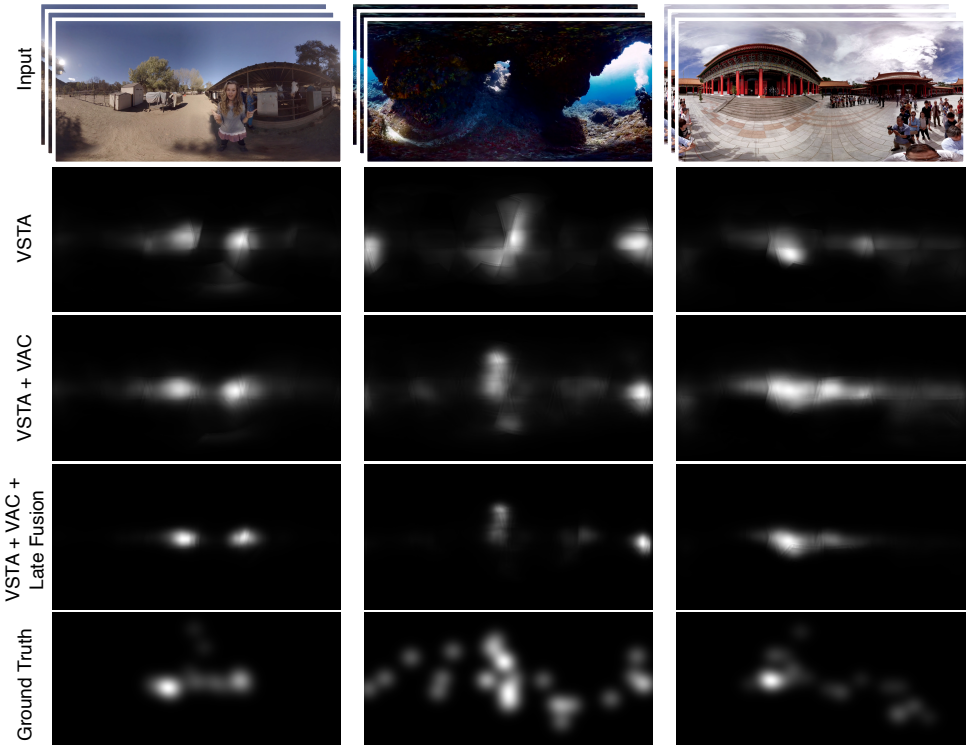

Figure 4: **Qualitative comparison** for our *VSTA baseline (third row)*, with the proposed *VAC loss (fourth row)*, and the optional *late-fusion (last row)*, compared to the ground truth.

of temporal information processing in omnidirectional videos. In Table 3, we compare our VSTA with two distinct approaches, namely *2+1D-CNN* [34] backbone and *Offline EMA*. *2+1D-CNN* backbone is an R2+1D model [35] which performs convolution over consecutive frames, pre-trained on undistorted *normal-FOV* crops in 360° videos. We replace ResNet-18 + VSTA with *2+1D-CNN* + VSA to introduce temporal features for spatial self-attention.

In the other setting, we keep ResNet-18 and VSA and apply a weighted exponential moving average on F consecutive predictions for temporal aggregation. Additionally, we ablate on transformer depth, which shows that our VSTA blocks gradually learn better spatio-temporal representations in deeper layers. Our experiments demonstrate that the proposed VSTA mechanism outperforms the spatial-only setting and the other two spatio-temporal approaches. We refer the reader to the supplementary for comprehensive experiments on *the transformer depth,* and *temporal window size F*, along with *joint spatio-temporal attention.*

3.4 Failure Cases

We analyse some failure cases of our method in this section. A common reason for failure cases is due to texts on the video added later, uncommon textures, and multiple people in the centre of videos that cause centre bias. Fig. 5.

Figure 5: **Failure cases.** *First column:* Although SalViT360 detects the saliency of two men in the scene, it predicts a higher saliency for the drawing on the wall. *Second column:* Similarly, the model predicts false positives for the text added to the video later. *Third column:* When multiple people with similar visual attributes are located along the equator, the model tends to assign higher saliency near the centre due to the centre bias encountered in the ground truth fixations.

4 Conclusion

In this study, we proposed SalViT360, a transformer-based framework using tangent image representations for 360° video saliency prediction. We also introduced a spatio-temporal attention mechanism on tangent viewports to capture the global and temporal context in omnidirectional videos effectively. Our framework employs a new (model-agnostic) spherical geometry-aware position embedding structure based on angular coordinates. Lastly, we suggested an unsupervised, consistency-based loss function as a regulariser to the artefacts commonly observed in projection-based dense-prediction models. Our experiment results obtained from three omnidirectional video saliency datasets demonstrate that our proposed SalViT360 model outperforms the state-of-the-art qualitatively and quantitatively. As a future research direction, we plan to investigate the impact of spatial audio in 360° videos and extend our proposal to the audio-visual domain.

Acknowledgements: This work was supported in part by KUIS AI Center Research Award, Unvest R&D Center, TUBITAK-1001 Program Award No. 120E501, GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem, and BAGEP 2021 Award of the Science Academy to A. Erdem.

References

[1] Shengxi Li, Mai Xu, Yun Ren, and Zulin Wang. Closed-form optimization on saliency-guided image compression for hevc-msp. *IEEE Transactions on Multimedia*, 20(1):155–170, 2018.

[2] Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018.

[3] Dipti Mishra, Satish Kumar Singh, Rajat Kumar Singh, and Divanshu Kedia. Multi-scale network (mssg-cnn) for joint image and saliency map learning-based compression. *Neurocomputing*, 460:95–105, 2021.

[4] Shiping Zhu, Chang Liu, and Ziyao Xu. High-definition video compression system based on perception guidance of salient information of a convolutional neural network and hevc compression domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):1946–1959, 2020.

[5] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph. (TOG)*, 35(6):1–12, 2016.

[6] Haoxiang Wang, Zhihui Li, Yang Li, Brij B Gupta, and Chang Choi. Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, 130:64–72, 2020.

[7] Zejiang Hou and Sun-Yuan Kung. Multi-dimensional dynamic model compression for efficient image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–643, 2022.

[8] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1383–1391, 2019.

[9] Mengmeng Zhu, Guanqun Hou, Xinjia Chen, Jiaxing Xie, Haixian Lu, and Jun Che. Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1953–1962, 2021.

[10] Jingwei Guan, Shuai Yi, Xingyu Zeng, Wai-Kuen Cham, and Xiaogang Wang. Visual importance and distortion guided deep image quality assessment framework. *IEEE Transactions on Multimedia*, 19(11):2505–2520, 2017.

[11] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. Assessing visual quality of omnidirectional videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3516–3530, 2019.

[12] Miaomiao Qiu and Feng Shao. Blind 360-degree image quality assessment via saliency-guided convolution neural network. *Optik*, 240:166858, 2021.

[13] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018.

[15] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

[16] Ned Greene. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications*, 6(11):21–29, 1986.

[17] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *Proc. IEEE ICMEW*, 2018.

[18] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.

[19] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. IEEE/CVF CVPR*, pages 1420–1429, 2018.

[20] F. Chao, L. Zhang, W. Hamidouche, and O. Deforges. A multi-fov viewport-based visual saliency model using adaptive weighting losses for 360-degree images. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[21] Yasser Dahou, Marouane Tliba, Kevin McGuinness, and Noel O'Connor. ATSal: An attention based architecture for saliency prediction in 360 videos. In *International Conference on Pattern Recognition*, pages 305–320, 2021.

[22] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360-degree videos. In *Proc. ECCV*, September 2018.

[23] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji. Viewport-dependent saliency prediction in 360 video. *IEEE Trans. Multimed.*, 23:748–760, 2020.

[24] Heeseung Yun, Sehun Lee, and Gunhee Kim. Panoramic vision transformer for saliency detection in 360-degree videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 422–439. Springer, 2022.

[25] Yasser Abdelaziz Dahou Djilali, Tarun Krishna, Kevin McGuinness, and Noel E. O'Connor. Rethinking 360deg image visual attention modelling with unsupervised learning. In *Proc. IEEE/CVF ICCV*, pages 15414–15424, October 2021.

[26] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proc. IEEE/CVF CVPR*, pages 12426–12434, 2020.

[27] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[28] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *Proc. IEEE/CVF CVPR*, pages 5333–5342, 2018.

[29] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[30] F. Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 355–358, 2020.

[31] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2018.

[32] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing*, 120:104395, 2022.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33, 2020.

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.