CIT. Journal of Computing and Information Technology, Vol. 34, No. 1, March 2026, 65–79
doi: 10.20532/cit.2026.1006291

A Lightweight Spatiotemporal Saliency Detection Framework for VR Panoramic Dynamic Scenes

Dezhi Kong, Huijuan Hao and Bo Gao

Hebei University of Water Resources and Electric Engineering, Cang Zhou, Hebei, China

Saliency detection in virtual reality (VR) panoramic dynamic scenes faces two major challenges: geometric distortion caused by equirectangular projection (ERP) and the high computational cost of modeling long-term temporal dependencies. To address these issues, we propose TAD-Net, a lightweight spatiotemporal saliency detection framework that integrates cubemap projection (CMP), temporal attention, knowledge distillation, and adversarial training. CMP efficiently reduces panoramic distortion while enabling standard 2D convolutional processing. A dual-stream network extracts spatial appearance and temporal motion features, and a temporal attention module enhances dynamic saliency discrimination. To reconcile the accuracy–latency trade-off, a heavy teacher model transfers long-range temporal knowledge to a lightweight student model via distillation, while adversarial training improves boundary sharpness. Extensive experiments on Salient360-Dynamic and VR-EyeDynamic demonstrate that TAD-Net achieves state-of-the-art performance, improving AUC-Judd by up to 5.2% while maintaining real-time inference at 35.1 FPS on an RTX 3080 GPU. Cross-dataset evaluation confirms robust generalization under domain shifts. The results indicate that the proposed projection–perception–distillation pipeline effectively balances geometric correction, temporal reasoning, and real-time constraints in VR applications.

ACM CCS (2012) Classification: Computing methodologies → Computer graphics → Image manipulation → Image processing

Keywords: VR panoramic images, dynamic scenes, saliency detection, spatiotemporal features, attention mechanism, lightweight optimization

1. Introduction

Virtual reality (VR) panoramic content has become an essential component of immersive computing applications, including interactive entertainment, cultural heritage visualization, remote inspection, and simulation-based training [1]. In such systems, visual saliency detection (SD) plays a critical role in enabling adaptive rendering, gaze-guided compression, bandwidth optimization, and user interaction enhancement. Accurate identification of salient regions allows VR platforms to allocate computational and display resources efficiently, thereby improving perceptual quality while maintaining low latency.

However, saliency detection in panoramic dynamic scenes presents two fundamental technical challenges. First, most VR content is encoded using equirectangular projection (ERP), which maps spherical visual data onto a two-dimensional rectangular plane. This representation introduces severe geometric distortion, particularly near polar regions, disrupting the spatial invariance assumptions underlying conventional 2D convolutional neural networks (CNNs). Although spherical convolution and tangent-based projections have been proposed to address distortion, these approaches incur high computational costs or complex boundary management, limiting their applicability in real-time VR systems [2].

Second, dynamic scene saliency depends not only on spatial appearance but also on temporal

motion patterns [3]. Conventional static saliency models fail to capture motion-driven attention shifts, while video saliency models designed for planar data do not explicitly account for spherical deformation. Moreover, advanced temporal modeling techniques such as long-range attention or transformer-based architectures often introduce significant computational overhead, making them unsuitable for latency-sensitive VR applications where real-time inference (typically ≥ 30 FPS) is required.

Existing panoramic saliency approaches tend to optimize either geometric correction or temporal modeling, but rarely under strict computational constraints. Projection-based methods mitigate distortion yet neglect dynamic dependencies, whereas heavy spatiotemporal networks improve accuracy at the cost of inference speed [4]. Therefore, a unified framework capable of jointly addressing spherical distortion, motion modeling, and real-time efficiency remains an open research problem.

To address these limitations, this study proposes a lightweight spatiotemporal saliency detection framework termed TAD-Net (Temporal Adversarial Distillation Network). The design follows a projection–perception–distillation paradigm. First, cubemap projection (CMP) is employed to convert ERP images into six low-distortion planar faces, preserving local spatial structures while enabling efficient use of standard 2D CNNs. Second, a dual-stream architecture decouples spatial appearance features from temporal motion cues extracted via optical flow [5]. A temporal attention module models inter-frame dependencies to enhance dynamic saliency discrimination. To reconcile the accuracy–latency trade-off, knowledge distillation transfers long-term temporal reasoning from a computationally heavy teacher network to a lightweight student network used during inference. Additionally, adversarial training is introduced to refine saliency boundary sharpness without increasing inference complexity.

The main contributions of this work are summarized as follows:

1. Projection-aware spatiotemporal modeling: A cubemap-based preprocessing strategy is integrated into the network architecture to mitigate ERP distortion while maintaining computational efficiency,
2. Lightweight dynamic saliency architecture: A dual-stream network with temporal attention effectively captures motion-dependent saliency in panoramic scenes,
3. Distillation-driven efficiency optimization: Knowledge distillation compresses long-range temporal reasoning into a lightweight inference model, enabling real-time deployment,
4. Comprehensive evaluation and generalization validation: Experiments on Salient360-Dynamic and VR-Eye Dynamic demonstrate state-of-the-art accuracy with real-time performance, and cross-dataset testing confirms robustness under domain shifts.

By systematically integrating geometric correction, temporal modeling, and model compression within a unified architecture, TAD-Net provides a computationally efficient solution for VR panoramic dynamic saliency detection.

2. Literature Review

In related research fields, SD and related technologies have been applied and developed in different scenarios. Li Q. *et al.* proposed a micro-oxidation salient object detection model (MO-SOD) suitable for planar images to address the problems regarding the fact that micro-oxidation of oxygen-free copper materials is difficult to identify by the naked eye and that manual detection is costly and subjective. This model integrates small target feature extraction, key target attention pyramid fusion and anchor box-free decoupling detection modules while combining complete intersection over union (CIOU) loss and focus loss to optimize the loss function, thus achieving efficient and accurate detection [6]. Lin J. *et al.* constructed the red-green-blue-depth (RGB-D) video salient object detection (SOD) dataset ViDSOD-100 and proposed the attentive triple-fusion network (ATF-Net) model, providing an effective solution for RGB-D video SOD [7]. Zhang Y. *et al.* constructed the Panoramic Audio-Visual Saliency 10K dataset PAVS10K to advance the development of SOD. PAVS10K outperformed all the

comparison models, providing a foundation for the study of panoramic image SOD [8].

In terms of processing new sensor data, Li D. *et al.* focused on dynamic and active-pixel vision sensor (DAVS) cameras. In response to the problem that it is difficult to effectively fuse spatiotemporal cues between asynchronous events and frame modes, they proposed the Streaming Object Detection with Transformer (SODFormer) model. This model mined spatiotemporal cues through the spatiotemporal Transformer module and integrated heterogeneous modalities using the asynchronous attention fusion module, which performed well in high-speed motion, low light and other scenarios [9]. In the specific application of VR and panoramic images, Vasic I. *et al.* developed a user behavior (UB) tracking algorithm that combines the salient regions of panoramic images to explore the interaction mode between users and virtual cultural heritage in order to optimize the experience. They used the Italian virtual museum as input, extracted the region of interest (ROI) through the VR engine and counted the visitor behavior. They found that the ROI was mostly concentrated on the artworks, which clarified the core interests of users [10]. Liu Y. *et al.* combined Bézier curves with 3D models to construct a new model. Experiments showed that the average recognition accuracy of the model under different lighting conditions was better than that of traditional methods [11]. Wang J. *et al.* used multidimensional information analysis technology to construct an integrated VR system. This system improved the naturalness and fidelity of human-computer interaction [12]. Liu R. *et al.* explored the relationship between urban park landscape quality and visitor experience through eye-tracking experiments. Their research results could provide a reference for the design of landscape VR scenes [13]. Cammarasana S. *et al.* analyzed the temporal changes of geometric and kinematic characteristics in continuous frames to classify actions. This method provided a technical reference for related detection optimization [14].

Existing panoramic approaches largely focus on static distortion correction while neglecting temporal continuity, whereas standard video saliency models capture motion but fail to address spherical geometric deformations. TADNet distinguishes itself by bridging this gap through a unified architecture that simultaneously resolves panoramic distortion via lightweight projection and incorporates long-term temporal dependencies through knowledge distillation, thereby overcoming the limitations of single-domain baselines.

3. Methodology

The proposed architecture implements a 'projection-perception-distillation' pipeline adapted for resource-constrained VR environments. The lightweight projection is structurally aligned with the dual-stream network to minimize computational redundancy associated with distortion correction. Furthermore, a knowledge distillation strategy is employed to transfer temporal reasoning capabilities from a teacher network to a lightweight student model. This optimization aims to balance the wide spatiotemporal receptive fields required for analysis with the low-latency constraints of VR interaction.

3.1. Panoramic Image Preprocessing and Feature Extraction

When processing dynamic panoramic images for VR, the primary challenge comes from their unique projection method. Standard equirectangular projection (ERP) formats induce severe polar distortions that compromise the translation invariance of 2D CNN kernels. To mitigate this without incurring the computational cost of spherical convolutions, the proposed framework utilizes cubemap projection (CMP) to remap the spherical field into six low-distortion planar faces, strictly preserving local spatial structures for efficient feature extraction. [15–16]. To solve this core problem and meet the requirements of VR applications for high real-time performance, this study first introduced a lightweight panoramic projection module. The core function of this module is to efficiently remap the input ERP format image into a two-dimensional plane representation with less distortion. The study adopts cubemap projection (CMP) as the specific implementation of this lightweight module, as shown in Figure 1.

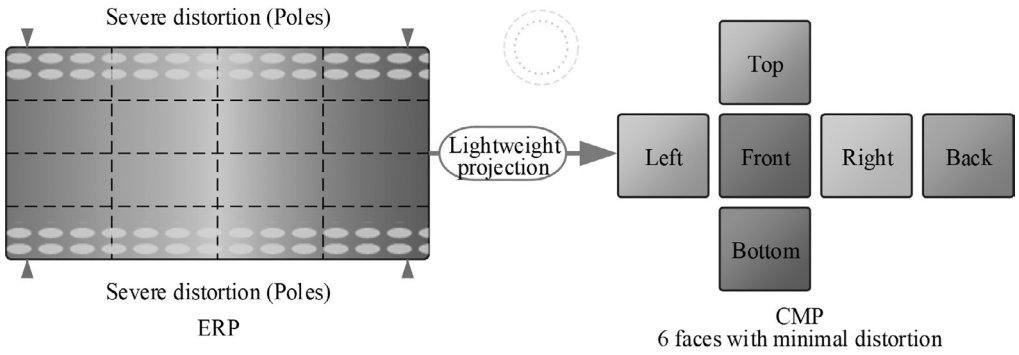


Figure 1. Lightweight panoramic projection module.

As shown in Figure 1, CMP projects a spherical image onto six cubic planes surrounding the sphere, with each plane (front, back, left, right, top, and bottom) corresponding to a 90° × 90° field of view. Compared to spherical convolutions that incur high computational overhead and tangent representations that complicate continuity handling, CMP provides the optimal trade-off between geometric fidelity and inference efficiency. It minimizes local distortion within faces while enabling the direct utilization of highly optimized standard 2D CNNs, thereby avoiding the latency penalties associated with specialized spherical geometric operations. After solving the spatial distortion problem of static images, the second key step is to build a feature extraction network that can effectively handle "dynamic scenes". The salience of dynamic scenes depends not only on the appearance of static objects, but more importantly on capturing the motion information and temporal context of objects. Therefore, a dual-stream spatiotemporal network was constructed to extract and process these two complementary pieces of information separately, as shown in Figure 2.

The network in Figure 2 contains a spatial stream and a temporal stream. The spatial stream is responsible for extracting appearance features, and its input is the six cubemap faces of the current frame I_t after processing by the lightweight projection module described above. The backbone network of the spatial stream adopts a lightweight CNN architecture (such as MobileNetV3 or a lightweight version of ResNet) to ensure high inference speed of the model. This stream learns the static semantics of the scene from high-resolution RGB information, such as object category, texture, color, and contour. Meanwhile, the temporal stream focuses on capturing motion features. The input of the temporal stream is the dense optical flow map O_t calculated between two frames I_t and I_{t-1}, as shown in equation (1).

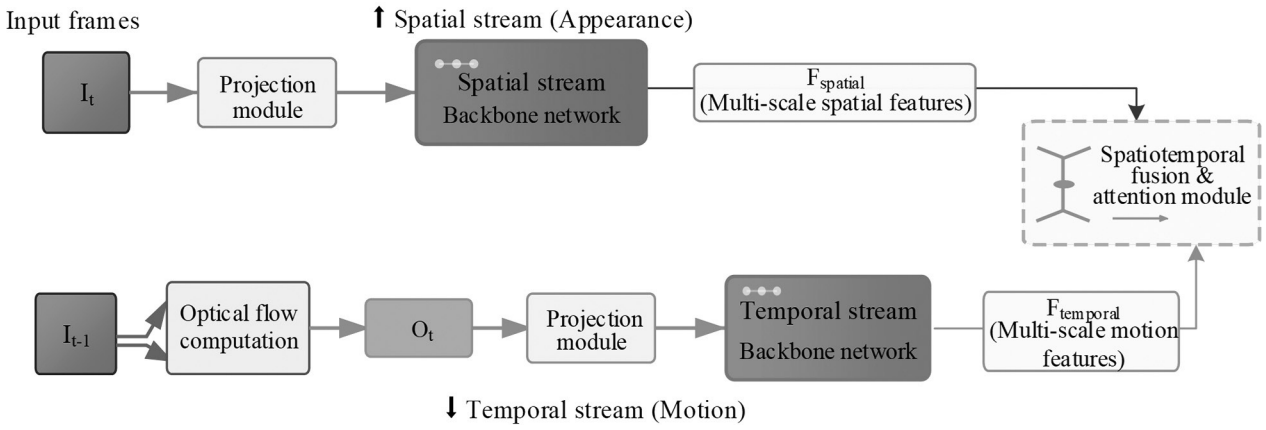


Figure 2. Feature extraction framework of dual-stream spatiotemporal network.

$$O_t = \mathcal{F}_{flow}(I_t, I_{t-1}) \quad (1)$$

In equation (1), I_t and I_{t-1} represent the current frame and the previous frame, respectively. $\mathcal{F}_{flow}$ represents an efficient optical flow estimation algorithm that estimates a two-dimensional motion vector for each pixel. The generated optical flow map O_t is also projected through a cubemap (to match the input format of the spatial flow) and then fed into a temporal backbone network with a similar structure to the spatial flow but with independent weights. To enable the subsequent attention module to fuse information of different granularities, both data streams employ a multi-scale feature extraction strategy, as shown in Figure 3.

As shown in Figure 3, feature maps are extracted from different stages (*i.e.*, shallow, middle, and deep) of the two backbone networks to obtain a set of hierarchical spatiotemporal features. This process is shown in equation (2).

$$\begin{cases} F^k_{spatial} = \mathcal{N}^k_{spatial}(\{\mathcal{T}_{proj}(I_t)\}_{i=1}^6) \\ F^k_{temporal} = \mathcal{N}^k_{temporal}(\{\mathcal{T}_{proj}(O_t)\}_{i=1}^6) \end{cases} \quad (2)$$

In equation (2), $\mathcal{T}_{proj}$ represents the cube mapping projection operation, and i = 1..6 represents the six faces of the cube. $\mathcal{N}^k_{spatial}$ and $\mathcal{N}^k_{temporal}$ represent the feature extraction subnetworks (*i.e.*, part of the backbone network) at the k-th scale of the spatial flow and temporal flow, respectively. $F^k_{spatial}$ and $F^k_{temporal}$ are the extracted appearance and motion feature maps at the k-th scale. Through multi-scale spatiotemporal feature extraction and fusion, the model efficiently decouples the dynamic scene into a rich set of multi-scale spatiotemporal feature representations while preserving the spatial structure integrity of the panoramic image.

3.2. Optimization of the SD Model in Dynamic Scenarios

After the constructed dual-stream spatiotemporal network provides rich but separate multi-scale appearance features, the current core task is to design an efficient optimization model to achieve accurate and real-time saliency discrimination of dynamic scenes [17-19]. Simply fusing spatiotemporal features is insufficient to cope with the complex dynamics in VR scenes, such as the instantaneous appearance of objects, occlusion, and global optical flow interference caused by camera movement [20]. To justify the architectural complexity, the proposed pipeline employs a complementary "Capture-Compress-Refine" strategy. While standard temporal attention effectively models dynamic dependencies on the low-distortion CMP plane without requiring heavy spherical geometric adaptations, its computational cost inhibits real-time performance. Consequently, knowledge distillation is strictly employed to transfer this high-level temporal reasoning into a lightweight student network, while adversarial training is introduced specifically to counteract the prediction blurring often caused by distillation-based regression, ensuring that the final saliency maps retain sharp boundaries. The study first introduced a temporal attention module to enhance the model's understanding of dynamic evolution and its ability to discriminate key temporal nodes, as shown in Figure 4.

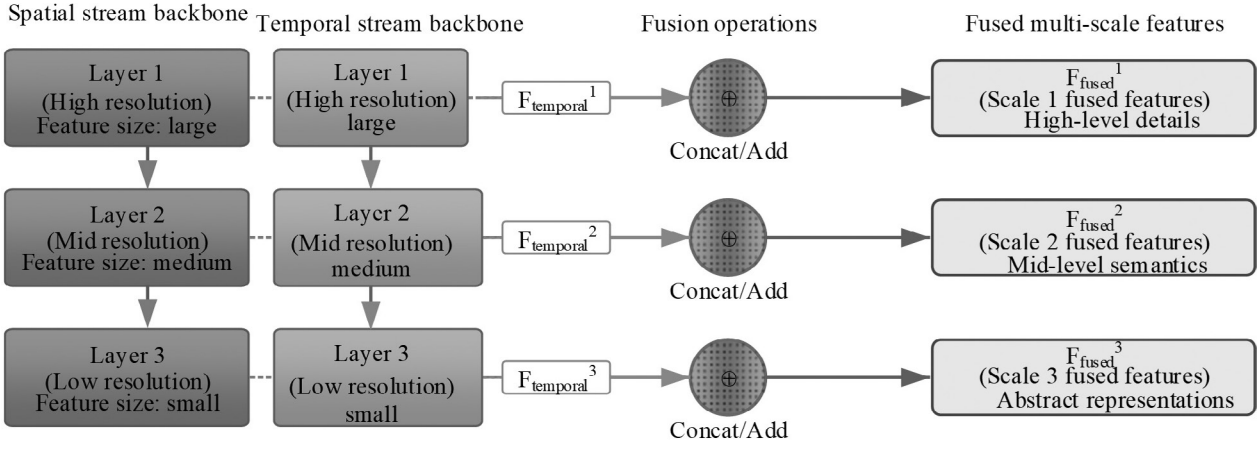


Figure 3. Multi-scale spatiotemporal feature extraction and fusion.

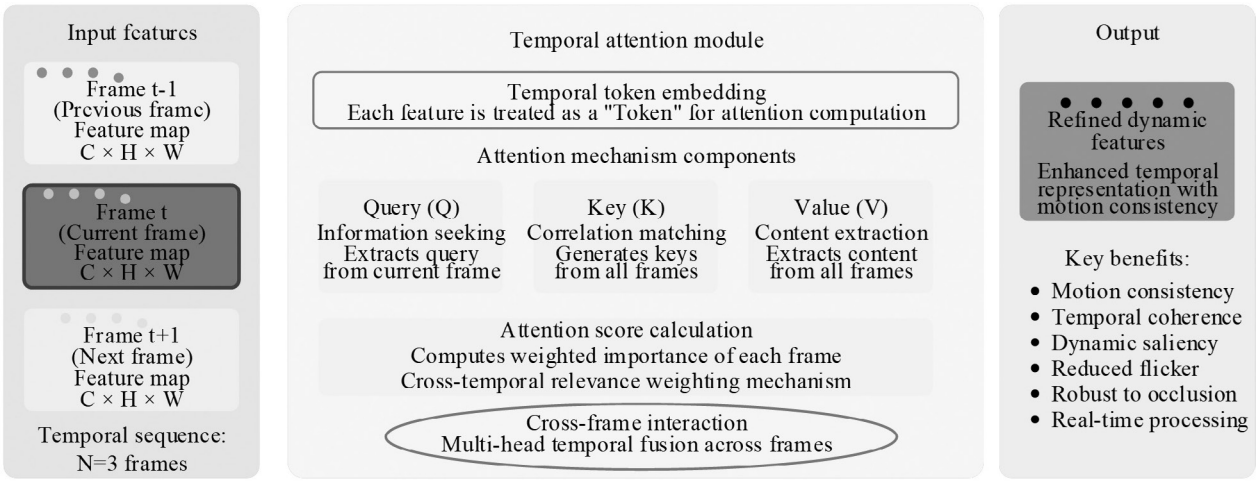


Figure 4. Structure of the temporal attention module.

In Figure 4, operating on a sequence of *N* consecutive spatiotemporal feature maps, the module utilizes self-attention to dynamically recalibrate feature importance along the temporal dimension. By treating feature maps as tokens, the mechanism explicitly models the correlation between frames, enabling the network to distinguish consistent motion patterns from transient noise without relying on manual feature engineering. This allows the module to learn the fact whether an object becomes salient because of its continuous motion or a suddenly appearing object is more salient than an object moving at a constant speed in the background, thereby effectively strengthening the discrimination ability of dynamic salient regions and suppressing non-salient motion interference. The output of the module is a highly refined salient feature map after temporal context awareness and weighting. However, although the complex attention mechanism can improve accuracy, it usually brings huge computational burden, which is contrary to the 35 FPS real-time inference speed required by VR applications [21-22]. In order to solve this inherent contradiction between accuracy and speed, the study adopted two advanced training strategies to optimize a lightweight student model. The first strategy is KD, as shown in Figure 5.

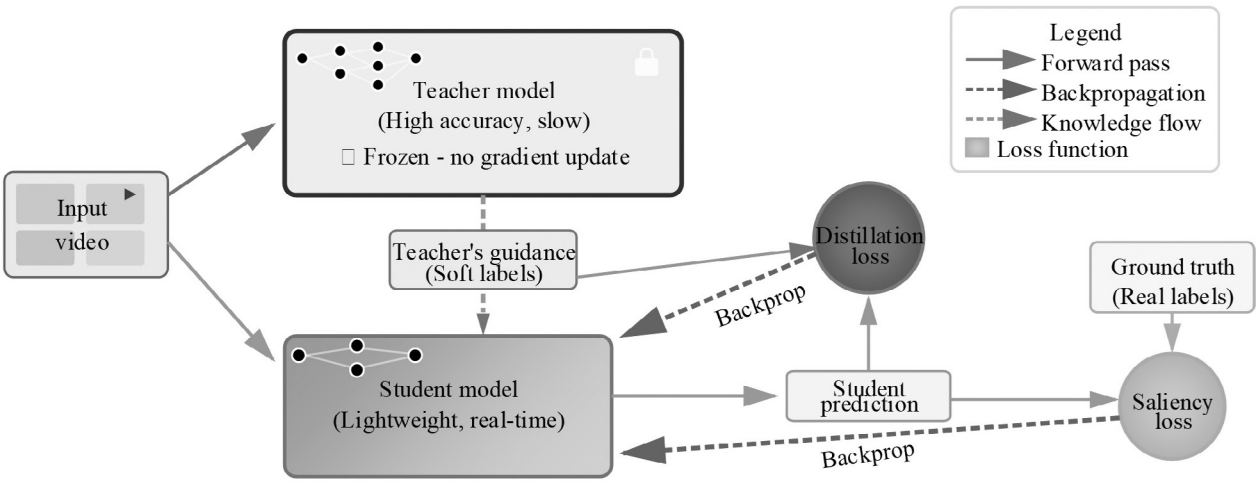


Figure 5. KD training strategy.

As shown in Figure 5, this model can have deeper networks (such as ResNet-101), more complex attention structures, and longer temporal dependencies (such as N=10), with the goal of achieving the highest accuracy without considering speed. During training, a teacher model T generates $S_{Teacher}$, and the lightweight student model G not only learns to minimize the significance loss $L_{saliency}$ of the true label S_{GT}, but is also forced to ensure that the logits distribution $S_{Student}$ of its output approximates the distribution of $S_{Teacher}$. This learning by imitation is achieved through a special distillation loss L_{KD}, which forces the student model G to learn the hidden knowledge and stronger generalization ability of the teacher model T, as shown in equation (3) [23].

$$L_{KD} = KL(S_{Student} / \tau \,||\, S_{Teacher} / \tau) \tag{3}$$

In equation (3), KL is the Kullback-Leibler divergence, τ is the "temperature" hyperparameter of distillation used to smooth the output distribution of the teacher model. Secondly, to further improve model performance, especially the edge sharpness and visual realism of the saliency map, an adversarial training strategy was introduced, as shown in Figure 6.

In Figure 6, the student model G is considered as a generator whose task is to generate a salient map that is indistinguishable from the real one. At the same time, a discriminator network (Discriminator D) is introduced, which is a binary classification CNN, whose task is to distinguish between the G-generated salient map $S_{Student}$ and the real labeled map S_{GT}. For the generator G, the goal is to minimize the following loss, that is, to maximize the probability that the discriminator D classifies its generated result as "real", as shown in equation (4).

$$L_{ADV}(G) = -\log(D(S_{Student})) \tag{4}$$

In equation (4), $D(S_{Student})$ is the probability that the discriminator D classifies $S_{Student}$ as a real image. Finally, the total loss function of the model is a weighted sum that combines all optimization objectives and is used to train the student network G end-to-end. First, the basic saliency loss $L_{saliency}$ is defined, which consists of pixel-level binary cross-entropy (BCE) loss and region-level intersection over union (IoU) loss, as shown in equation (5).

$$L_{saliency} = L_{BCE} + L_{IoU} \tag{5}$$

In equation (5), the BCE loss is the core of the SD task, and its calculation method is shown in equation (6).

$$L_{BCE} = -\frac{1}{N}\sum_{i}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{6}$$

In equation (6), N is the total number of pixels, y_i is the true label (0 or 1) of the i-th pixel, and $\hat{y}_i$ is the probability that the model predicts the pixel to be significant. The loss L_{IoU} is used to optimize the overlap between the predicted map and the true map on the region contour. Combining all loss terms, the total loss L_{Total} is shown in equation (7).

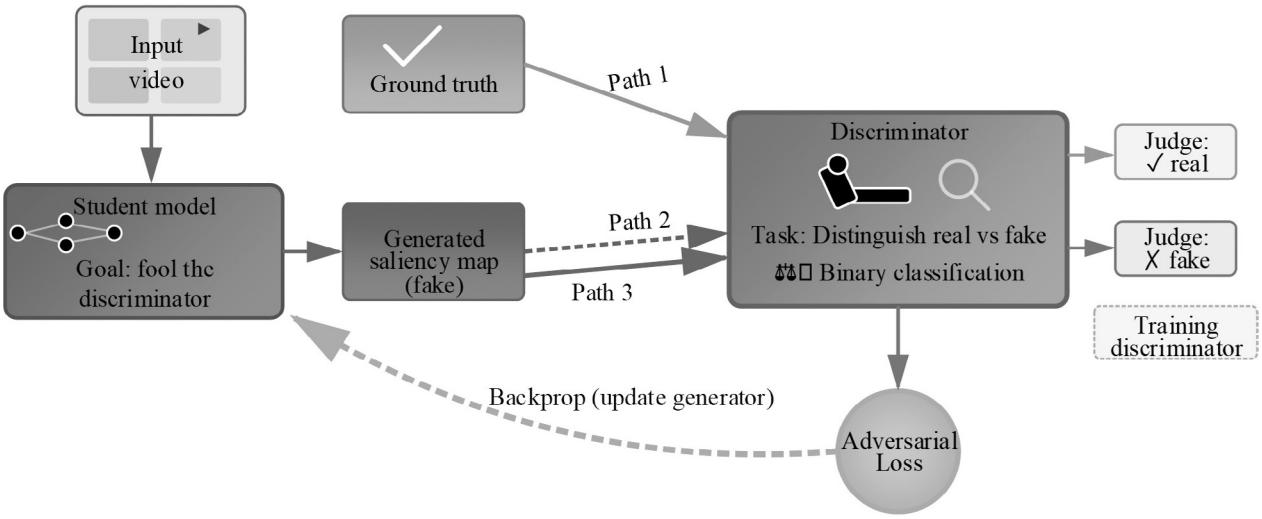


Figure 6. Optimization strategy for adversarial training.

$$L_{Total} = \lambda_{sal} \cdot L_{saliency} + \lambda_{kd} \cdot L_{KD} + \lambda_{adv} \cdot L_{ADV}(G) \quad (7)$$

In equation (7), λ_{sal}, λ_{kd}, and λ_{adv} are hyperparameters used to balance the importance of various tasks. Through this complete optimization process that combines temporal attention, KD, and adversarial training, the study can comprehensively improve the accuracy, speed, and generalization of SD in dynamic VR scenes. The proposed method is named temporal adversarial distillation saliency network (TAD-Net).

4. Results and Analysis

4.1. Performance Testing of the VR Panoramic Image SD Method

To conduct model training and performance comparisons, all experiments were performed in a unified hardware and software environment. The hardware configuration consisted of an NVIDIA GeForce RTX 3080 (10 GB) GPU, an AMD Ryzen 7 5800X CPU, and 32 GB of RAM. The software environment used Ubuntu 20.04, with PyTorch 1.10 as the deep learning framework and CUDA 11.3 for acceleration. To verify the effectiveness of each model component, ablation experiments were conducted on the Salient360-Dynamic dataset. M1 served as the baseline model, containing only spatial flow. M2 added temporal flow to M1. M3 added a temporal attention module. M4 added adversarial training (GAN). M5 added KD to M4. In the metrics, ↑ indicates higher performance and ↓ indicates lower performance. The results of the model ablation experiments are shown in Table 1.

As shown in Table 1, the comparison between the spatial-only baseline (M1) and the dual-stream model (M2) reveals that removing optical flow leads to a sharp performance drop (approximately 3% in AUC-Judd), strictly validating the necessity of motion cues despite the computational overhead. To mitigate potential flow unreliability in fast-motion or low-texture scenarios, the temporal attention module (M3)

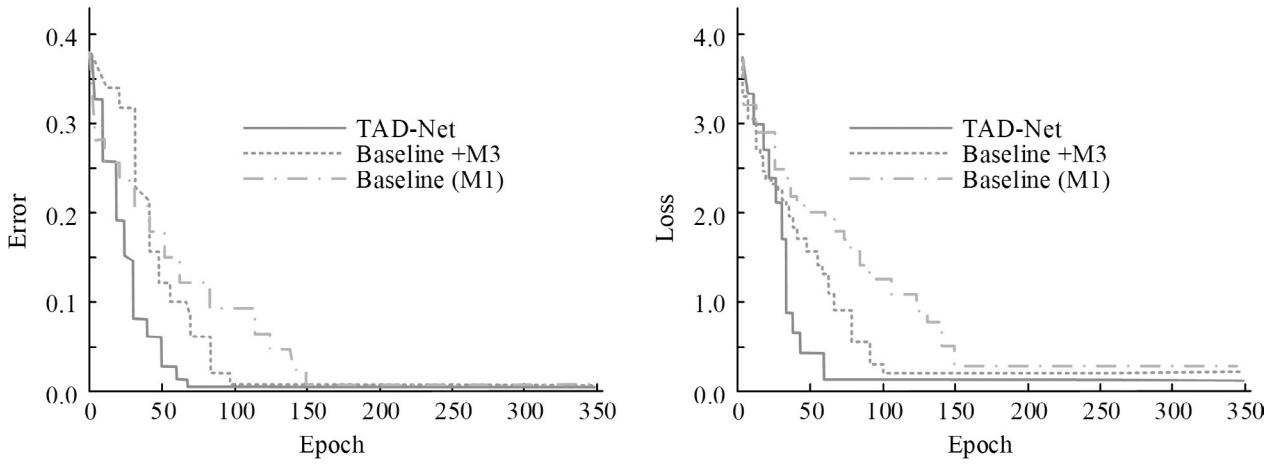
Table 1. Results of model ablation experiment.

Model	Components	AUC-Judd (↑)	F-measure (Fbw) (↑)	MAE (↓)	NSS (↑)
M1	Baseline (spatial stream only)	0.8217 ± 0.006	0.6133 ± 0.007	0.1204 ± 0.003	2.1039 ± 0.016
M2	M1 + temporal stream	0.8509 ± 0.005	0.6721 ± 0.006	0.1033 ± 0.002	2.4518 ± 0.013
M3	M2 + temporal attention	0.8711 ± 0.003	0.7018 ± 0.004	0.0917 ± 0.002	2.6844 ± 0.010
M4	M3 + adversarial training (GAN)	0.8724 ± 0.003	0.7153 ± 0.004	0.0921 ± 0.002	2.7103 ± 0.010
M5	Ours (M4 + KD)	0.8803 ± 0.002	0.7296 ± 0.003	0.0864 ± 0.001	2.8022 ± 0.008

effectively functions as a reliability gate, dynamically down-weighting corrupted motion features and shifting focus to spatial representations when optical flow estimation degrades. Interestingly, after adding adversarial training, M4 showed a significant improvement in F-measure (Fbw), consistent with the expectation that GANs could optimize edge contours. However, the MAE metric showed a slight deterioration of 0.0004, possibly due to the slight sacrifice of pixel-level average accuracy in adversarial game theory. Finally, M5 (TAD-Net), after incorporating KD, comprehensively improved all metrics, especially correcting the deterioration of M4 in MAE, reaching an optimal value of 0.0864. This demonstrates the comprehensive improvement of model generalization and robustness brought about by the KD strategy. For training parameter settings, all models used the AdamW optimizer with an initial learning rate of $1e^{-4}$, using cosine annealing for learning rate decay. The batch size was set to 8. The input image resolution was uniformly adjusted to 1024×512. To ensure statistical reliability and reproducibility, all quantitative experiments were repeated five independent times using fixed random seeds. Consequently, the reported metrics represent the mean ± standard deviation (SD). Statistical significance was validated using paired t-tests based on the results of the five independent runs ($N = 5$). These tests were performed separately for each evaluation metric, where p-values < 0.05 confirmed that performance improvements over comparison methods were statistically significant. Figure 7 shows the model's training convergence over 350 epochs.

Figure 7(a) shows the change in error with the number of training epochs, and Figure 7(b) shows the change in loss with the number of training epochs. Both figures clearly show that the proposed TAD-Net method significantly outperformed the comparative methods in both convergence speed and performance. TAD-Net's curve showed a rapid descent, with both error and loss converging to their minimum points at approximately 60 epochs, demonstrating marked training efficiency. Table 2 shows a comparison of the accuracy of TAD-Net and state-of-the-art (SOTA) methods on the dynamic panopticon dataset.

As shown in Table 2, TAD-Net surpassed comparison methods in all accuracy metrics across the two mainstream datasets. Method-A and Method-B suffered from performance limitations due to their lack of ability to handle panoramic distortion and dynamic temporal sequences. Method-C achieved an AUC-Judd value of 0.8368 on the Salient360-Dynamic dataset, while TAD-Net reached 0.8803. Table 3 compares the complexity and inference speed of TAD-Net with SOTA methods.



(a) Error. *(b)* Loss.

Figure 7. Comparison of model decoding performance.

Table 2. Accuracy comparison of TAD-Net and SOTA methods on dynamic panoramic datasets.

Dataset	Method	AUC-Judd (↑)	F-measure (Fbw) (↑)	MAE (↓)	CC (↑)	NSS (↑)
Salient360-Dynamic	Method-A (CVPR'22)	0.8316 ± 0.005	0.6508 ± 0.006	0.1119 ± 0.002	0.7102 ± 0.007	2.2138 ± 0.015
	Method-B (ECCV'22)	0.8104 ± 0.004	0.6277 ± 0.005	0.1253 ± 0.003	0.6921 ± 0.006	2.0594 ± 0.012
	Method-C (ICCV'23)	0.8368 ± 0.003	0.6802 ± 0.004	0.0974 ± 0.001	0.7419 ± 0.005	2.5011 ± 0.010
	Ours (TAD-Net)	0.8803 ± 0.002	0.7296 ± 0.003	0.0864 ± 0.001	0.7833 ± 0.004	2.8022 ± 0.008
VR-EyeDynamic	Method-A (CVPR'22)	0.8011 ± 0.006	0.5912 ± 0.007	0.1403 ± 0.003	0.6517 ± 0.008	1.9813 ± 0.014
	Method-B (ECCV'22)	0.7933 ± 0.005	0.5839 ± 0.006	0.1477 ± 0.004	0.6439 ± 0.007	1.9127 ± 0.013
	Method-C (ICCV'23)	0.8126 ± 0.004	0.6137 ± 0.005	0.1296 ± 0.002	0.6808 ± 0.006	2.1051 ± 0.011
	Ours (TAD-Net)	0.8549 ± 0.003	0.6582 ± 0.004	0.1171 ± 0.001	0.7246 ± 0.005	2.3416 ± 0.009

Table 3. Comparison of complexity and inference speed between TAD-Net and SOTA methods.

Method	Backbone	Params (M) (↓)	GFLOPs (↓)	Memory (MB) (↓)	FPS (↑)
Method-A (CVPR'22)	ResNet-50	45.1	102.3	1152	18.2
Method-B (ECCV'22)	Spherical-Net	38.2	151.8	1789	12.7
Method-C (ICCV'23)	Swin-T	30.1	75.4	1024	26.8
Method-D (TPA-MI'22)	ViT-B/16	86.5	170.6	2156	9.9
Method-E (CVPR'23)	MobileNetV3-L	7.5	35.9	641	30.5
Ours (TAD-Net)	Custom-light	12.7	41.21	718	35.1

Results presented in Table 3 illustrate significant efficiency advantage of TAD-Net. Methods B and D had the lowest FPS. Methods A and C failed to meet the standards for real-time VR applications. Method E's FPS was still slightly lower than the proposed method. In contrast, TAD-Net, through its lightweight backbone network and efficient optimization strategies, achieved only 12.7M parameters and maintained a low GFLOPs of 41.2. Most importantly, TAD-Net achieved an inference speed of 35.1 FPS on an RTX 3080, making it the only method among the compared methods to meet the real-time processing requirements. TAD-Net achieved SOTA accuracy while maintaining optimal inference efficiency, thus achieving the best balance between accuracy and speed. While the reported 35.1 FPS on an RTX 3080 validates real-time capabilities for high-performance workstations, deployment on resource-constrained standalone VR headsets may exhibit lower frame rates. However, given the model's low computational complexity (41.2 GFLOPs), preliminary theoretical estimates suggest that mid-range mobile GPUs (*e.g.*, NVIDIA Jetson series) can sustain >15 FPS, which remains viable for non-competitive VR interaction. Future work will focus on INT8 quantization and TensorRT optimization to fully bridge the gap for embedded deployment. Crucially, the apparent architectural complexity is confined strictly to the offline training phase, where heavy components like the teacher network and GAN discriminator are utilized to transfer knowledge. During inference, these modules are discarded, leaving only the lightweight student model. Therefore, the 5.2% improvement in AUC-Judd represents a substantial breakthrough within the strict constraints of real-time processing, justifying the training complexity by overcoming the performance bottleneck that limits standard lightweight baselines.

4.2. Simulation Validation of the VR Panoramic Image SD Method

The study selected scenes from the Salient360-Dynamic test dataset to evaluate the robustness of the model in complex dynamic environments. In Figure 8, Baseline 1 and Baseline 2 correspond to two SOTA comparison methods with higher and lower overall performance, respectively. All input images were uniformly set to 1024×512 resolution.

Figure 8(a) shows the input frame, and Figure 8(b) shows the ground truth label. In the first row, the "Fast Motion" scene, the input frame contained a subject with significant motion blur. The prediction result of Baseline 1 (Figure 8(d)) was significantly larger than the real object and Baseline 2 (Figure 8(e)) failed to capture the complete shape. In contrast, TAD-Net (Figure 8(c)) effectively processed motion information and generated a saliency map that was almost identical to GT. In the second row, the "Occlusion" scene, the subject was divided into two parts by a foreground pillar (Occluder). Baseline 1 (Figure 8(d)) incorrectly identified the occluder (pillar) as a saliency region and Baseline 2 (Figure 8(e)) almost completely lost the smaller occluded part. TAD-Net (Figure 8(c)) accurately identified the two visible salient parts and completely ignored insignificant occlusions, demonstrating the model's powerful scene understanding and discrimination capabilities. Although CMP introduces boundary discontinuities, the proposed framework mitigates seam artifacts through the expansive receptive fields of deep convolutional layers and the consistency constraints of knowledge distillation. By forcing the student network to mimic the teacher's continuous probability distribution, the model learns to implicitly interpolate features across cube faces. Qualitative evidence in Figure 8 confirms this geometric consistency, where large dynamic objects spanning multiple faces are detected as coherent, non-fragmented regions. While CMP inherently creates geometric discontinuities at cube edges, the architecture ensures seam consistency through the aggregation of deep features with wide receptive fields and the regularization effect of knowledge distillation. The student model is constrained to replicate the teacher's seamless global attention, effectively bridging projection gaps. As evidenced qualitatively in Figure 8, salient regions spanning across adjacent cube faces exhibit smooth continuity without fragmentation, demonstrating the model's robustness against projection artifacts. A visualization comparison of different methods in complex backgrounds and small object scenes is shown in Figure 9.

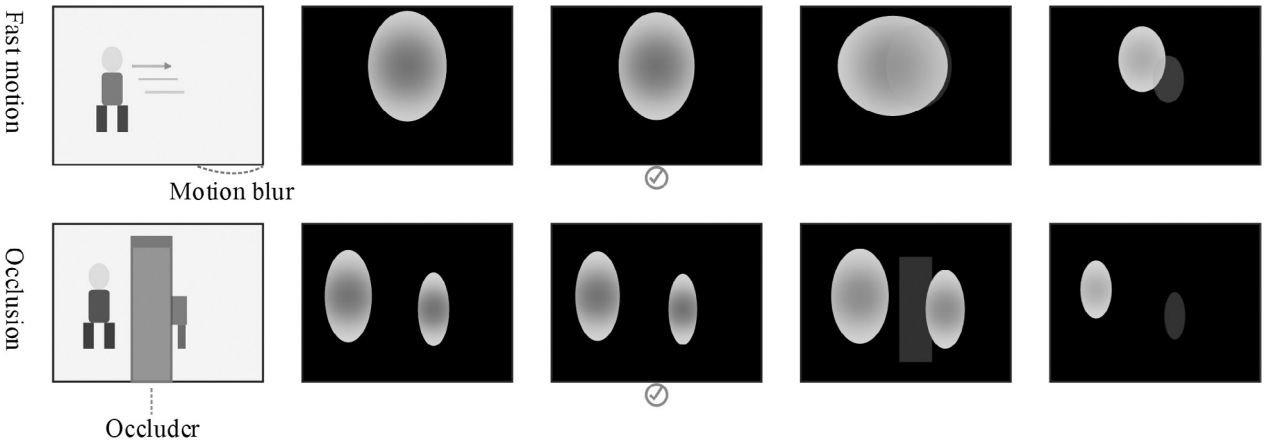


Figure 8. Visual comparison of TAD-Net and two baseline methods in dynamic and occluded scenes.

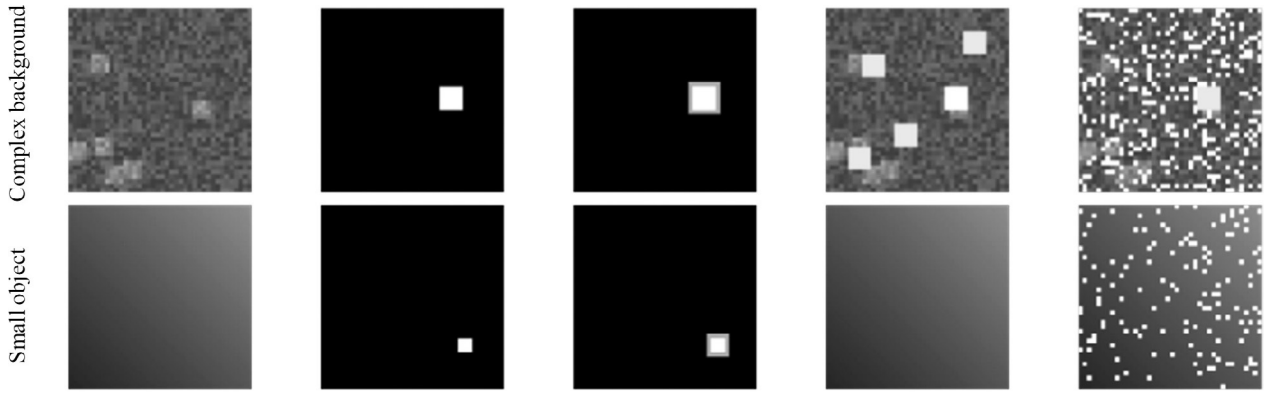


Figure 9. Visual comparison of different methods in scenes with complex backgrounds and small objects.

In Figure 9, in the first row, "Complex Background," both Baseline 1 and Baseline 2 were severely affected by high-frequency noise, resulting in numerous false positives (Figure 9(d) and Figure 9(e)). TAD-Net (Figure 9(c)) successfully suppressed all background noise and accurately located the unique target. In the second row, "Small Object" scene, both Baseline methods (Figure 9(d) and (e)) failed to detect the small target, leading to missed detections. TAD-Net (Figure 9(c)) still successfully captured the target, demonstrating that TAD-Net significantly outperformed the baseline models in both noise robustness and detection sensitivity.

4.3. Cross-Dataset Generalization Analysis

To evaluate the model's robustness against domain shifts and diverse VR content acquisition conditions (*e.g.*, varying lighting and capture devices), cross-dataset validation was conducted. Specifically, the model trained on the Salient360-Dynamic dataset was directly evaluated on the VR-EyeDynamic dataset without fine-tuning, and vice versa. The results are presented in Table 4.

As indicated in Table 4, although a minor performance decline is observed in the cross-dataset settings due to inherent distribution differences between datasets, TAD-Net maintains a

Table 4. Cross-dataset generalization performance of TAD-Net.

Training Set	Testing Set	Condition	AUC-Judd	CC	NSS
Salient360	Salient360	Within-Dataset	0.8803	0.7833	2.8022
Salient360	VR-Eye	Cross-Dataset	0.8412	0.7015	2.2841
VR-Eye	VR-Eye	Within-Dataset	0.8549	0.7246	2.3416
VR-Eye	Salient360	Cross-Dataset	0.8256	0.6988	2.4103

robust performance level (*i.e.*, AUC > 0.82). This confirms that the proposed architecture learns generalized spatiotemporal saliency representations rather than overfitting specific dataset biases, demonstrating strong adaptability to unseen real-world VR scenarios.

5. Conclusion

This study presented TAD-Net, a lightweight spatiotemporal saliency detection framework designed for VR panoramic dynamic scenes. The framework addresses two primary challenges in panoramic saliency detection: geometric distortion introduced by equirectangular projection and the high computational cost associated with long-range temporal modeling. By integrating cubemap projection, a dual-stream spatiotemporal architecture, temporal attention, knowledge distillation, and adversarial refinement within a unified pipeline, the proposed approach achieves a balance between accuracy and real-time efficiency.

Experimental results on Salient360-Dynamic and VR-EyeDynamic datasets demonstrate that TAD-Net consistently improves saliency detection performance while maintaining real-time inference capability. The ablation analysis confirms the complementary contributions of motion modeling, attention refinement, and distillation-based compression. Cross-dataset evaluation further indicates that the learned representations generalize across different acquisition conditions and scene distributions, suggesting robustness to domain shifts.

From a computational perspective, the proposed projection–perception–distillation strategy provides an effective solution to the inherent trade-off between spherical distortion correction and temporal dependency modeling. By confining heavy architectural components to the offline training stage and deploying only the lightweight student network during inference, the framework maintains low computational overhead without sacrificing discriminative capacity.

Despite these advances, several limitations remain. The current framework relies on pre-computed optical flow for motion modeling, which introduces additional preprocessing cost and potential sensitivity to flow estimation errors. Future work will investigate end-to-end motion encoding strategies that reduce dependency on explicit optical flow computation. In addition, further optimization through model quantization, pruning, or hardware-aware acceleration (*e.g.*, TensorRT or edge deployment strategies) may improve adaptability to resource-constrained VR devices. Finally, expanding evaluation to larger and more diverse panoramic datasets will help assess scalability and robustness in more complex real-world VR environments.

Overall, this work demonstrates that distortion-aware projection combined with distil-

lation-driven model compression is a viable direction for efficient panoramic dynamic saliency detection in immersive computing systems.

Declaration of Competing Interests

The authors declare no conflict of interest.

Funding

This research was jointly supported by the Self-funded Project of Cangzhou Science and Technology Plan for the 2023-2024 Fiscal Year: Study on Visual Saliency Detection of Panoramic Images Based on VR Technology (Grant No.:23244101032) and Hebei Provincial Water Science and Technology Plan Project (2025): Design of an Intelligent Water Quality Monitoring and Pollution Early Warning System for the Grand Canal Based on Deep Learning (Grant No.: 2025-36).

Data Availability

Data used in this study is proprietary.

References

[1] Y T. Khanh *et al.*, "Digital Pathways to Sustainability: Empirical Evidence of Tourism Industry Transformation in the Industry 5.0 era", *Journal of Management Changes in the Digital Era*, vol. 2, no. 1, pp. 110–119, 2025. http://dx.doi.org/10.33168/JMCDE.2025.0108

[2] J. Nalivaikė, "Evolution of Online Marketing Communication Tools: Classification, Technological Integration, and Functional Analysis Across Web Generations", *Journal of Management Changes in the Digital Era*, vol. 2, no. 1, pp. 10–24, 2025. http://dx.doi.org/10.33168/JMCDE.2025.0102

[3] Y.-T. Bau *et al.*, "Android Malware Multiclass Classification Using Machine Learning: Evaluating the Performance of Random Forest, Artificial Neural Network, and Convolutional Neural Network", *Journal of Logistics, Informatics and Service Science*, vol. 11, no. 10, pp. 1–19, 2024. http://dx.doi.org/10.33168/JLISS.2024.1001

[4] S. Liu, "Research on Computational Methods and Algorithms for Dimensionality Reduction and Feature Selection in High Dimensional Data", *Journal of Logistics, Informatics and Service Science*, vol. 10, no. 3, pp. 1–12, 2023. http://dx.doi.org/10.33168/JLISS.2023.0301

[5] W. Zhang, "Investigation on the Use of Virtual Reality in the Teaching of Engineering Education Based on Functional Linked Neural Network", *Studies in Informatics and Control*, vol. 33, no. 3, pp. 103–113, 2024. http://dx.doi.org/10.24846/v33i3y202410

[6] Q. Li *et al.*, "MO-SOD: Micro-oxidation Small Object Detection Model for Oxygen-free Copper Surfaces Based on Microscopic Imaging System", *ACS Omega*, vol. 8, no. 7, pp. 6608–6620, 2023. http://dx.doi.org/10.1021/acsomega.2c07043

[7] J. Lin *et al.*, "Vidsod-100: A New Dataset and a Baseline Model for rgb-d Video Salient Object Detection", *International Journal of Computer Vision*, vol. 132, no. 11, pp. 5173–5191, 2024. http://dx.doi.org/10.1007/s11263-024-02051-5

[8] Y. Zhang *et al.*, "PAV-SOD: A New Task Towards Panoramic Audiovisual Saliency Detection", *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–26, 2023. http://dx.doi.org/10.1145/3565267

[9] D. Li *et al.*, "Sodformer: Streaming Object Detection with Transformer Using Events and Frames", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 14020–14037, 2023. http://dx.doi.org/10.1109/TPAMI.2023.3298925

[10] I. Vasic *et al.*, "3VR: Vice Versa Virtual Reality Algorithm to Track and Map User Experience", *ACM Journal on Computing and Cultural Heritage*, vol. 17, no. 3, pp. 1–19, 2024. http://dx.doi.org/10.1145/365634

[11] Y. Liu, "Application of Fusing Bézier Curves and 3D Models in VR Stereo Vision", *Multimedia Systems*, vol. 31, no. 3, pp. 1–14, 2025. http://dx.doi.org/10.1007/s00530-025-01781-x

[12] J. Wang *et al.*, "Integrated Design System of Voice-visual VR Based on Multi-dimensional Information Analysis", *International Journal of Speech Technology*, vol. 24, no. 1, pp. 1–8, 2021. http://dx.doi.org/10.1007/s10772-020-09696-w

[13] R. Liu and X. Wang, "Visual Quality Analysis of Urban Park Landscapes Based on Eye Tracking: A Case Study of Nanjing Xiaohong Stone Carving Park", *Journal of Landscape Research*, vol. 16, no. 4, pp. 5–12, 2024. http://dx.doi.org/10.16785/j.issn1943-989x.2024.4.002

[14] D. Martin *et al.*, "Spatio-temporal Analysis and Comparison of 3D Videos", *The Visual Computer*, vol. 39, no. 4, pp. 1335–1350, 2023. http://dx.doi.org/10.1007/s00371-022-02409-1

[15] D. Martin *et al.*, "Scangan360: A Generative Model of Realistic Scanpaths for 360 Images", *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2003–2013, 2022. http://dx.doi.org/10.1109/TVCG.2021.3139238

[16] X. Y. Zhang *et al.*, "Simulation of Guided Crowd Evacuation Scheme of High-Speed Train Carriage", *International Journal of Simulation Modelling*, vol. 22, no. 1, pp. 110–120, 2023. http://dx.doi.org/10.2507/IJSIMM22-1-638

[17] G. F. Chai and Y. Z. Xia, "Multi-Robot Path Optimization and Simulation for Multi-Route Inspection in Manufacturing", *International Journal of Simulation Modelling*, vol. 22, no. 1, pp. 121–132, 2023. http://dx.doi.org/10.2507/IJSIMM22-1-CO1

[18] A. Petrovas *et al.*, "Gestalt Principles Governed Fitness Function for Genetic Pythagorean Neutrosophic WASPAS Game Scene Generation", *International Journal of Computers Communications & Control*, vol. 18, no. 4, p. 5475, 2023. http://dx.doi.org/10.15837/ijccc.2023.4.5475

[19] M. Zheng, "Enhancing Automation with Label Defect Detection and Content Parsing Algorithms", *Journal of Computing and Information Technology*, vol. 31, no. 1, pp. 1–19, 2023. http://dx.doi.org/10.20532/cit.2023.1005734

[20] L. Zheng, "A Visual Cortex-Attentive Deep Convolutional Neural Network for Digital Image Design", *Journal of Computing and Information Technology*, vol. 31, no. 1, pp. 21–37, 2023. http://dx.doi.org/10.20532/cit.2023.1005695

[21] Y. H. Wu *et al.*, "EDN: Salient Object Detection via Extremely-Downsampled Network", *IEEE Transactions on Image Processing*, vol. 31, pp. 3125–3136, 2022. http://dx.doi.org/10.1109/TIP.2022.3164550

[22] Z. Wang *et al.*, "Video Saliency Prediction via Joint Discrimination and Local Consistency", *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1490–1501, 2022. http://dx.doi.org/10.1109/TCYB.2020.2989158

[23] P. W. Chen *et al.*, "Viewing Bias Matters in 360° Videos Visual Saliency Prediction", *IEEE Access*, vol. 11, pp. 41534–41546, 2023. http://dx.doi.org/10.1109/ACCESS.2023.3269564

Received: December 2025
Revised: February 2026
Accepted: February 2026

Contact addresses:
Dezhi Kong*
Hebei University of Water Resources and Electric Engineering
Cang Zhou
Hebei
China
e-mail: dezhikong1011@163.com
*Corresponding author

Huijuan Hao
Hebei University of Water Resources and Electric Engineering
Cang Zhou
Hebei
China
e-mail: huijuanhao_1@163.com

Bo Gao
Hebei University of Water Resources and Electric Engineering
Cang Zhou
Hebei
China
e-mail: gaobo@hbwe.edu.cn

Dezhi Kong graduated from Beijing University of Technology with a master's degree in mathematics. Currently she holds the professional title of lecturer at the School of Hebei University of Water Resources and Electric Engineering. Her research interest include mathematics and applied mathematics.

Huijuan Hao received her Master's degree from the YanShan University. Currently, she works at the School of Hebei University of Water Resources and Electric Engineering. Her research interest include digital media technology and computer application technology.

Bo Gao holds a Master's degree from Harbin Institute of Technology and is currently a PhD candidate in electrical engineering at the Shenyang University of Technology whileserving as an Associate Professor at the Department of Power Engineering at Hebei University of Water Resources and Electric Engineering. Research interests center on artificial intelligence and electrical automation control, particularly the integration of intelligent algorithms into modern power systems and industrial automation processes.