Circuits, Systems, and Signal Processing (2025) 44:3983–4012
https://doi.org/10.1007/s00034-024-02983-w

HSNet: A Novel Edge-Preserving Hierarchical Separable Network for Video Shadow Detection

Hemraj Singh[1] · Mridula Verma[2] · Ramalingaswamy Cheruku[1]

Received: 18 April 2024 / Revised: 20 November 2024 / Accepted: 20 November 2024 /
Published online: 25 January 2025

Abstract

Video Shadow Detection (VSD) is an emerging research direction that holds significant importance in surveillance video analysis for multiple industrial applications. Recently, several research efforts in the field of VSD have focused on improving the performance and accuracy of shadow detection by utilizing state-of-the-art deep-learning models, overlooking the challenges associated with practical deployment on resource-constrained devices. To maintain this trade-off between accuracy and computational complexity, we propose a novel edge-preserving lightweight Hierarchical Separable Network (HSNet) for VSD tasks, which hierarchically extracts the attention-based multi-scale geometric spatiotemporal shadow features from videos to improve shadow detection performance while keeping the number of network parameters and floating point operations low. As far as we know, this is the first work that extracts the attention-based multi-scale geometric spatial and temporal features hierarchically. Additionally, a Geometric Attention Information Module (GAIM) is designed, which extracts geometric spatial and temporal resolution information from video frames and preserves the edge information. Next, a novel Edge-enhanced Detection Network (EDNet) is proposed to extract geometric spatial and temporal features and enhance edge information. To enhance the diversity of the existing datasets with visually complex shadow scene variations, we collected new annotated examples. Lastly, Shadow Region Intensity (SRI) loss is proposed to minimize the training loss and differentiate the geometric variation of the background and foreground of the objects. Extensive experimental results demonstrate that HSNet outperformed existing state-of-the-art

✉ Ramalingaswamy Cheruku
rmlswamy@nitw.ac.in

Hemraj Singh
hs720079@student.nitw.ac.in

Mridula Verma
vmridula@idrbt.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology Warangal, Hanamkonda, Telangana, India

[2] Institute for Development and Research in Banking Technology, Hyderabad, Telangana, India

models with 2.82% of increased accuracy while achieving 87.9 and 81.4% reduction in the number of parameters and FLOPS, respectively, on VSD. Our code and data samples and the corresponding annotations are available at https://github.com/shemraj/HSNet.

Keywords Geometric multi-scale encoder · Multi-level feature extraction · Hierarchical spatiotemporal features · Multi-scale feature · Video shadow detection · Deep learning

1 Introduction

The development of a sustainable edge computing (SEC) framework holds significant importance in the Industrial Internet of Things (IoT) context. This framework can potentially enhance the efficiency of data processing and transmission while enabling seamless adaptation to data-intensive applications. Surveillance video analytics is a prominent application in which extracting video semantics is crucial. Video semantics offer valuable insights and information that can be leveraged to establish trust in the data or trigger alerts for potential risks. Consequently, extracting video semantics is pivotal in deploying IoT solutions for real-time surveillance applications, such as video object tracking [1], video object segmentation [2], video object recognition [3, 4], and video scene understanding [5] by identifying the visual regions within a video frame that stands out the most in comparison to their surroundings across various spatial scales. In this direction, Video Shadow Detection (VSD) has also emerged as a crucial task that localizes shadow regions in videos and provides valuable insights, such as the direction of illumination [6], inferring the shape of occluding objects [7], localization [8], and scene geometry [9]. Looking at the applicability in multiple application areas such as the coal industry [10], robotics industry [11], road transportation industry [12], shopping mall [9], automobile industry [13], manufacturing [14], etc., VSD has become an urgent task.

Shadow detection in images and videos provides additional visual semantics that may impact the main task positively or negatively, as shown in Fig. 1. For instance, shadows may provide additional information about an object (the second row of the figure) [15–17], which may improve the detection performance. For example, the objects are not directly visible (fully or partially), but with the help of shadows of the object, they can be detected. In contrast, shadows may also create difficulties in detecting or segmenting the object, downgrading the performance of computer vision tasks [12, 18, 19] (the first row of the Fig. 1). In both cases, the effective detection of shadows is crucial.

With the advanced development of multiple deep learning models and architectures in the domain of computer vision over the past few years, the performance of these models has also been successfully evaluated on shadow detection tasks to learn the discriminative shadow features from images [20–22]. In videos, shadow information extraction presents more intricate challenges due to interference caused by a moving background, motion blur, and geometric changes in the foreground and background. Recent deep-learning approaches [2, 7, 23–25] model the relationships and

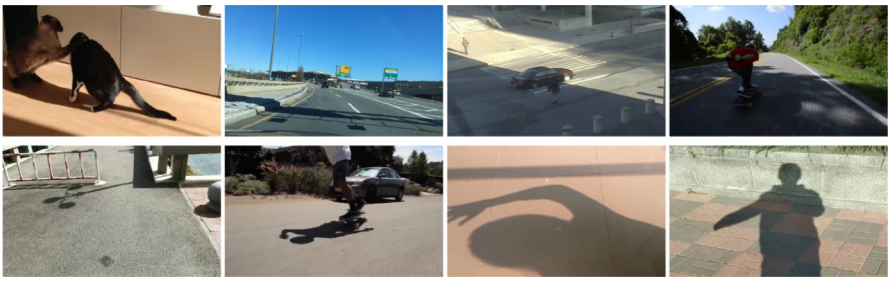


Fig. 1 Illustration of negative (first row) and positive (second row) impacts of shadow on the object detection task

correlations between extracted features from different video frames by leveraging the temporal dependencies between frames by applying convolution transformations to improve the performance. However, the bigger and computationally complex Convolution Neural Network (CNN)-based models are challenging to deploy in the practical setting for applications like smart homes [9], smart cities [26], smart traffic systems [27], etc., where the devices are usually resource-constrained (IoT devices, mobile, edge devices). In line with the current research trend, which focuses on developing lightweight models to balance computational complexity, energy consumption, and detection performance, our objective is to apply this approach to the unexplored domain of video shadow detection.

Existing methods of VSD [2–4] often suffer from losing low-level local information during the encoding process, due to which they face difficulties in capturing the global distribution and feature representation needed for complex and diverse shadows. This loss also hampers the feature recovery during decoding. To overcome the above problems, the transformer-based models [12, 15, 28–30] have been proposed, which rely on pixel-wise supervision to overlook the structural relationships between shadow boundaries and adjacent shadow or non-shadow regions. These models demonstrate strong performance in capturing global shadow semantics, such as partial shadow and large-scale variations. However, these models struggle to accurately represent local, global, and mixed shadow semantic regions, which causes difficulty in detecting tiny shadow objects, multiple shadow objects, complex shadow scenes, and geometric variation shadow scenes, particularly around fine boundaries. Additionally, these methods struggle to effectively merge low-level local features with high-level global features during decoding. The inefficient fusion of contextual information can result in poor restoration of images, making it challenging to distinguish between foreground and background, such as black objects and shadow areas. To overcome these problems, we propose a novel, efficient, lightweight Hierarchical Separable Network (HSNet), including Geometric Multi-scale Encoder Module (GMEM), Multi-level Feature Extraction Module (MFEM), Geometric Attention Fusion Module (GAFM), and Saliency Generation Module (SGM), which extracts the attention-based multi-scale geometric spatiotemporal features from video sequences with the help of several newly designed modules. These modules help to extract better geometric high spatial and temporal resolution information with fewer network parameters and floating

point operation and increase the inference speed of the network. We further modified the EfficientNet [31] with the help of the proposed Geometric Attention Information Module (GAIM), including the Self-Attention Geometric Feature Module (SGFM), which extracts and fuses the geometric attention information and makes the backbone network light. As per our knowledge, this is the first work, where attention-based multi-scale geometric spatial and temporal features are extracted for effective VSD.

Another limitation is associated with the evaluation methods of VSD models. Previous methods [5–7, 32] based on cross-entropy loss functions mainly focus on shadow pixels, neglecting non-shadow pixels, which can also provide more discriminative features. Moreover, due to poor detail perception, the cross-entropy loss function tends to have weak detection capabilities for tiny shadows. To exploit each pixel region locally and globally and reduce the divergence inconsistency while discriminating the foreground and background region intensity boundary, a Shadow Region Intensity Loss (SRI) is proposed, mainly focusing on significant regions near adjacent pixels or fine edges. Moreover, to enhance the existing datasets' diversity with visually intricate shadow scene variations, we introduce additional examples. To demonstrate the enhanced performance of the proposed model, it is evaluated on both the existing benchmark and the new samples. The proposed model demonstrates the capability to achieve comparable performance to that of the state-of-the-art (SOTA) models with 7.2 million parameters, 4.24 gigabytes (G) of floating point operation, and 90 FPS (frame per second) inference speed.

The primary contributions of this paper are summarized as follows:

1. A novel lightweight Hierarchical Separable Network (HSNet) is proposed, which extracts the multi-scale geometric spatiotemporal features from video sequences with the help of newly designed modules such as GMEM, MFEM, GAFM, and SGM.
2. A novel lightweight Edge Detection Network (EDNet) is proposed to extract the geometric variations of edges from the video frames. It utilizes a newly designed Contrast Edge Distillation (CED) with the help of Fast Fourier Convolution (FFC) [33] and Spatial Spectral Transformation (SST) [34], which adaptively learns the spatial and spectral edge information from different video frames and generates enhanced features.
3. New examples have been added to the benchmark datasets with visually intricate shadow scene variations.
4. A new plugin named GAIM is designed to make the backbone EfficientNet [14] lighter and extract the efficient geometric high spatial and temporal resolution information.
5. The proposed model is tested on the benchmark datasets ViSha [7], VISAD [23], and our new set of samples and compared with SOTA models in terms of model computational complexity and performance. Further, a detailed ablation study is provided to support the efficiency and effectiveness of the technical contributions.

2 Related Works

2.1 Shadow Detection in Image/Video

The current shadow detection models [35, 36] detect shadow pixels based on a single input image and frame, which have at least one shadow. For detecting shadows, the recent methods used the CNN-based shadow detection models [3, 6, 7, 16], which extracts deep features from images and videos based on superpixels. In [7], a triple parallel network has been designed to learn the discriminative features at intra/inter-video levels using a dual-gated co-attention module and derive features from its adjacent frames in the video. In [36], contextual-based spatial features have been extracted in a direction-aware manner. In contrast, Zhu et al. [22] suggested a Recurrent Attention Residual (RAR) model for combining the contextual information at a stack of CNN layers to detect the shadow. In [6], a distraction-aware information extraction technique has been designed to predict false positives and false negatives for detecting shadows. Ding et al. [8] introduced a shadow consistent correspondence (SC-Cor) method to enhance pixel-wise similarity across shadow regions for VSD tasks, which follows weakly-supervised learning of pixel-wise correspondence across frames, eliminating the need for dense pixel-to-pixel labels. Chen et al. [12] designed PSTNet, a pioneering data-driven model for video shadow removal, leveraging physical properties, spatio-temporal relationships, and temporal coherence. It utilized a dedicated physical branch with mask-guided attention and a progressive aggregation module to address dataset limitations. Barua et al. [4] proposed a novel privacy-preserving deep feature engineering model and validated using a custom shadow video dataset. It extracted the deep feature using AlexNet, producing a 9192-dimensional feature vector, the Chi^2 selector is refined into 1000 features and classified by support vector machine (SVM). Wu et al. [37] introduced a moving target shadow detection method for VideoSAR images, addressing high false alarm and missed detection rates while extracting shadow and local contrast information for robust background reconstruction for super-pixel segmentation. However, these models have generated false detection accuracy and suppressed false alarms across multiple frames. Zhou et al. [38] designed the timeline and boundary-guided diffusion (TBGDiff) network for video shadow detection, integrating dual scale aggregation (DSA) for enhancing temporal context, shadow boundary aware attention (SBAA) for edge-based shadow features, and space-time encoded embedding (STEE) modules enable advanced temporal guidance, significantly improving shadow detection. Wei et al. [28] introduced a spatial-temporal feature interaction strategy that refines global shadow semantics using local priors for inter-frame shadow relations. Additionally, a structure-aware shadow prediction module was designed, which modeled distance relations between local shadow edges and regions. Jie et al. [39] introduced RMLANet, a random multi-level attention network that utilizes shuffled feature aggregation and sparse attention to efficiently process high-resolution inputs. By reducing unnecessary dense attention, it significantly lowered computational complexity while maintaining high accuracy in feature fusion. Zhang et al. [40] designed a GNN-JFL, a novel graph neural network-based joint motion-appearance feature extraction method that enhanced shadow tracking accu-

racy by leveraging graph relationships. It integrated multi-object tracking with GNNs for robust feature representation and improved association of shadows in complex scenarios, which uses more upsampling and downsampling operations, causing information loss and increasing the network complexity as its limitation. However, these models have drawbacks in that they require more computational power and storage space, decrease the inference speed, and are unable to detect the geometric variation of the shadow at multiple scales.

2.2 Lightweight Video Shadow Detection

In real-world scenarios, visual recognition tasks must be performed efficiently, considering factors such as speed, power consumption, and memory usage while operating within computational resource limitations. So, designing a lightweight video shadow detection method is important in various computer vision applications, including object recognition, surveillance, and autonomous vehicles, etc. Liu et al. [15] presented a shadow deformation attention (SODA) module to detect the large shadow deformation in videos. Further, the shadow contrastive learning module (SCOTCH) is presented to extract the unified shadow information. Chen et al. [25] designed a semi-supervised video shadow detection method leveraging existing labeled image datasets to generate temporally consistent pseudo-labels using a spatio-temporally aligned network (STANet). Further, it integrates an uncertainty-guided learning strategy and a lightweight memory-propagated long-term network (MPLNet) with memory propagation for enhanced long-term consistency in shadow detection. Lin et al. [41] presented a method for real-time video shadow detection using a fast and lightweight algorithm. It focuses on exploiting the temporal consistency of shadows to reduce computation. Lu et al. [23] proposed a Spatio-Temporal Interpolation Consistency Training (STICT) framework that integrates unlabeled video frames with labeled images for enhanced shadow detection. It introduces spatial and temporal interpolation schemes and scale-aware constraints to improve pixel-wise classification and temporal prediction consistency. Wang et al. [30] proposed an effective and simple method to fine-tune the Segment anything model (SAM) for detecting the shadows. Further, it uses long short-term attention methods to detect the shadow in videos. Wu et al. [42] introduced a new shadow annotation method using graph convolution networks, which extracts the complete shadow mask information. In [43], a multi-input and multi-output (MIMO) strategy was proposed to extract the spatiotemporal information and reduce the computational complexity of a 3D separable convolution layer-based model. In this, Lin et al. [29] proposed a novel CNN leveraging motion-guided multi-scale memory features for enhanced video shadow detection, which integrates global, local, and motion memories with a multi-scale motion-guided long-short transformer (MMLT) module, dense-scale transformers, and memory-read pooling attention to accurately detect shadows of varying sizes.

2.3 Edge Detection Methods

The recent methods [44–47] mostly focus on pixel-level edge information to increase the performance of deep-learning methods. Qin et al. [48] designed a hybrid loss, which provides training supervision for correctly detecting salient objects at patch, map, and pixel levels. Zhao et al. [49] designed an edge-based deep learning model (EGNet) to preserve the salient object boundaries from across the salient object and edge information. In [45], the Swin Transformer is designed to extract the multi-modality features and optimize intra-level cross-modality features. Zhou et al. [46] designed an edge-guided recurrent positioning network (ERPNet), which performs two operations: (1) edge extraction and (2) feature fusion. These edge methods mostly face short connection problems and less correlation geometric variation between the previous and next frames. Luo et al. [21] designed an edge-aware spatial pyramid fusion network, which extracts multi-task features in airborne remote sensing images. Jiao et al. [50] designed a spectral feature-scalable framework Permutohedral Refined UNet with a conditional random field (CRF) for precise cloud and shadow segmentation. This pipeline efficiently refines edges using multi-spectral bilateral kernels, significantly improving shadow retrieval. Dong et al. [51] proposed a novel Additive Contour Model (ACM) based on Shadow Image and Reflection Edge (SIRE) for precise image segmentation by leveraging mean-filtered shadow images and reflection edges derived from energy function optimization. It combined with level set minimization, accurately captures target boundaries and enhances robustness through optimized length and distance regularization terms. However, these models require more computational complexity and fail to detect the geometrical change of its position dynamically at multiple scales. To overcome the above problems, an Edge Detection Network (EDNet) is proposed, which extracts the geometric multi-scale spatial and temporal features and produces enhanced edge maps without increasing the network complexity.

3 Proposed Hierarchical Separable Network

3.1 Model Architecture

The proposed Hierarchical Separable Network (HSNet) architecture is shown in Fig. 2. At first, the appearance frames are sent to the Geometric Attention Information Module (GAIM) (discussed in Sect. 3.2), which extracts the compressed geometrical spatiotemporal representations. This information aids in extracting the geometrically rich spatiotemporal backbone features X_k and edge information Y_k via the baseline EfficientNet architecture [31]. The three Receptive Field Blocks (RFBs) [52] control the eccentricities of the receptive field at different scales and generate more discriminative geometric spatiotemporal features. These RFBs are connected to the stack of Geometric Multi-scale Encoder Module (GMEM), which are designed to extract multi-scale spatial attention. Next, the Multi-level Feature Extraction Module (MFEM) (details are in Sect. 3.4.1) extracts multi-level features at different dilation rates with the help of skip-connections to preserve the quality of low- and high-level features through multiple layers. The respective features are extracted by the stack of RFBs,

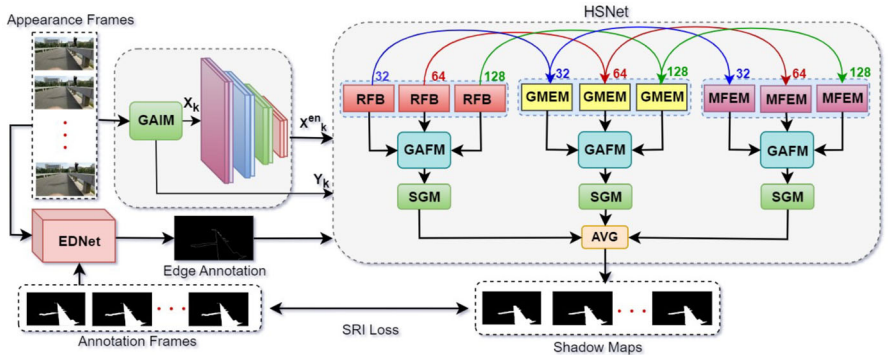


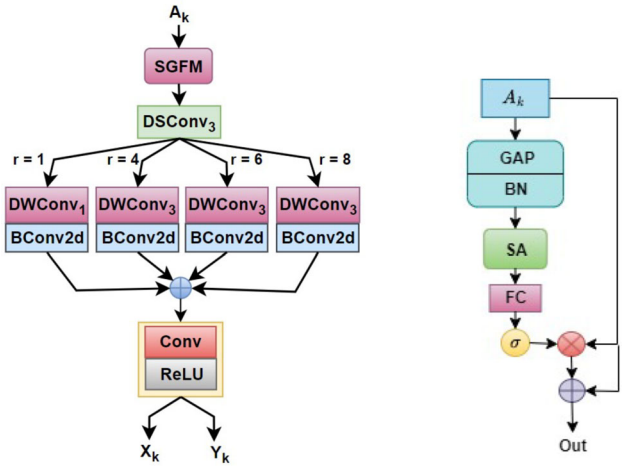
Fig. 2 The pipeline architecture of HSNet

GMEMs, and MEFMs at different scales and are then fused via the Geometric Attention Fusion Module (GAFM) (details are in Sect. 3.5). The fused outputs are passed to the three saliency generation modules (SGM), which generate the three saliency shadow maps. Further, we take the average (AVG) of the generated saliency maps and enhance the representation. The edge maps are extracted by our proposed EDNet (details are in Sect. 3.7), which refine the final saliency map based on edge-preserving contextual information. The details of each module in the proposed model are provided in the subsequent subsections.

3.2 Geometric Attention Information Module (GAIM)

The Geometric Attention Information Module (GAIM) is proposed to extract the geometric variation of low-level spatial and temporal resolution information, as shown in Fig. 3a. It configures with a deformable separable convolution (DSConv) layer with 3×3 filter and four depth-wise convolution (DWConv) layers with different filter sizes (1×1, 3×3, 3×3, 3×3) and different dilation rates (1, 4, 6, 8) in parallel fashion followed by basic 2d-convolution (BConv2d) layers with 1×1 filters. First, Fast Fourier Transformation (FFT) followed by FFTShift operation is performed to extract the geometric variations of the input appearance A^t_k frames. Further, the self-attention geometric feature module (SGFM) extracts high-level spatial and temporal information, and the output is passed to the DSConv layer, which extracts the depth-wise high-level geometric spatial and temporal information. Then four DWConv layers followed by BConv2d layers extract the multi-scale geometric high-level spatial and temporal information in a parallel fashion. Next, these features are fused together using element-wise addition operations to enhance the feature quality. At last, the point-wise 1×1 convolution layer (Conv) followed by the non-linear activation function (ReLU) is used to normalize and generate the multi-scale geometric high-level spatial and temporal information X_k and edge maps Y_k.

The geometric high-level spatial and temporal information X_k and low-level edge maps Y_k provided by GAIM are first passed to EfficientNet to extract the backbone high-level spatial and temporal information X_k. The extraction of the geometric high-



(a) GAIM (b) SGFM

Fig. 3 Architecture of GAIM and SGFM Modules. GAP is the global average pooling, BN is batch normalization, SA is a self-attention layer, FC is a fully connected layer, Conv is the convolution layer, DSConv is deformable separable convolution, DWConv is the depth-wise convolution, BConv2d is basic convolution, $\oplus$ is element-wise addition operation, $\otimes$ is element-wise multiplication operation, and r is the dilation rates

level spatial and temporal information procedure is given in Eq. 1.

$$\begin{aligned} X_k, Y_k &= \text{GAIM}(A_k^k) \\ X_k^{en} &= \text{EfficientNet}(X_k) \end{aligned} \tag{1}$$

3.2.1 Self-Attention Geometric Feature Module (SGFM)

The existing hierarchical attention-based feature extraction methods [2, 12, 13] have limitations towards the performance due to multiple down-sampling operations and are unable to detect geometric structural variation of shadow in multi-view. To overcome this problem, the self-attention geometric feature module (SGFM) is designed as shown in Fig. 3b, which captures the crucial local region context and global context information dynamically. The input appearance frames are passed to the SGFM, which extracts attention-based features using the Global Average Pooling (GAP) operation. Then, Batch Normalization (BN) is used to normalize the feature quality. Further, Self-Attention (SA) [53] extracts the attention-based global spatial and temporal features and enhances the information of spatial and temporal content. SA is used to handle long-range dependency between the present frame and the previous frame during information extraction. Next, the fully connected layer is applied densely to generate the feature maps. To normalize the feature map, the Sigmoid operation is used, and element-wise multiplication and skip connection are applied to enhance and preserve the feature quality. At last, element-wise addition operation balances the input and output feature maps to preserve the original feature quality and generate the high spatial and temporal resolution information output.

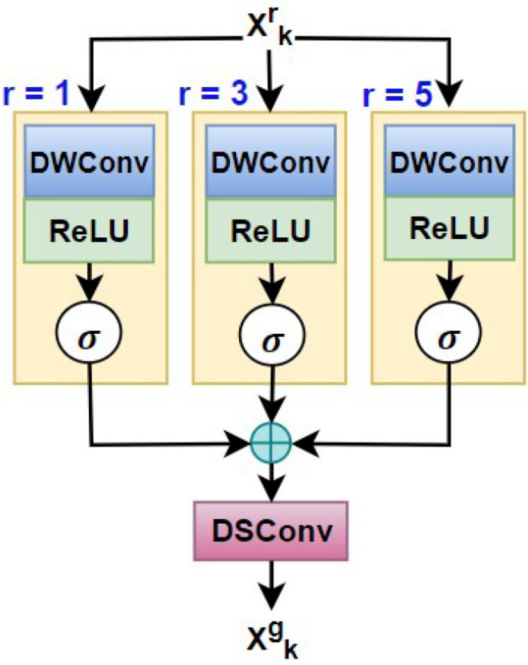


Fig. 4 Architecture of GMEM, which extracts the geometric multi-scale spatial and temporal information. The DWConv is the depth-wise convolution, $\oplus$ is the element-wise addition operation, σ is the Sigmoid operation, and r is the dilation rates

3.3 Multi-Scale High-level Spatiotemporal Feature Extraction

The output of the modified EfficientNet X_k^{en} is passed to the three Receptive Field Blocks (RFB) [52], each with three branches for the three dilation rates (r= 3, 5, 7). Each branch is designed with one basic 2-D convolution layer with 1×1 filter and the other two blocks with 3×3 filters, followed by DWConv layers with 3×3 filters. RFBs extract the geometric high-level spatiotemporal information ($X^r_{k_1}, X^r_{k_2}, X^r_{k_3}$) with different contexts and generate the outputs at three scales k_1 =32, k_2 =64, and k_3 =128, respectively. The procedure is given as follows,

$$\begin{aligned} X^r_{k_1} &= \mathrm{RFB}(X^{en}_k) \\ X^r_{k_2} &= \mathrm{RFB}(X^{en}_k) \\ X^r_{k_3} &= \mathrm{RFB}(X^{en}_k) \end{aligned} \tag{2}$$

3.3.1 Geometric Multi-scale Encoder Module (GMEM)

The existing approaches [15, 16, 25, 54] used pretrained network, or convolutional neural network, or UNet-like architecture to extract spatiotemporal information, which needs more computational complexity and fail to extract the geometric variation of spatiotemporal information at multiple scales due to their use of fixed kernel structure. To overcome these problems, a Geometric Multi-scale Encoder Module (GMEM) is proposed, which dynamically extracts the multi-scale geometric high-level spatial and

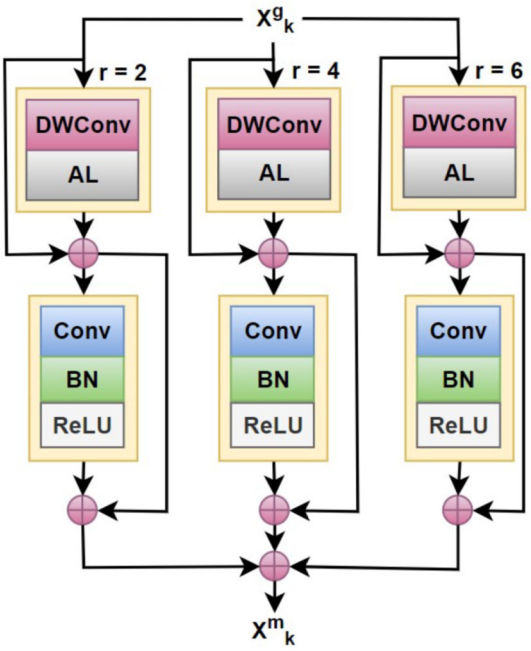


Fig. 5 Illustration of MFEM, which extracts the multi-level spatial and temporal information

temporal information without increasing the computational complexity. Additionally, it exploits the high-level spatiotemporal correlation information between center and side regions to efficiently locate and segment the salient shadow at three scales: k_1, k_2, and k_3. It is the combination of three DWConv layers with 3×3 filters at three different dilation rates ($r = 1, 3, 5$), three ReLU, and three Sigmoid operations (σ), fusion using element-wise addition $\oplus$, and DSConv as shown in Fig. 4. The outputs of the RFBs are passed to the three GMEM modules at the scale of 32, 64, and 128, respectively, to improve the representation of the information and generate the multi-scale high-level spatial and temporal information $X^g_{k_1}$, $X^g_{k_2}$, $X^g_{k_3}$. The procedure is given in Eq. 3.

$$\begin{aligned} X^g_{k_1} &= \text{GMEM}(X^r_{k_1}) \\ X^g_{k_2} &= \text{GMEM}(X^r_{k_2}) \\ X^g_{k_3} &= \text{GMEM}(X^r_{k_3}) \end{aligned} \tag{3}$$

3.4 Multi-Scale Multi-level Spatiotemporal Feature Extraction

3.4.1 Multi-level Features Extraction Module (MFEM)

The MFEM is shown in Fig. 5, which extracts the information from three branches at different scales (32, 64, 128) in a hierarchical fashion. It extracts the geometric variation of low-level spatiotemporal features using DWConv and AL layers with skip connections, while high-level spatiotemporal features are extracted using Conv,

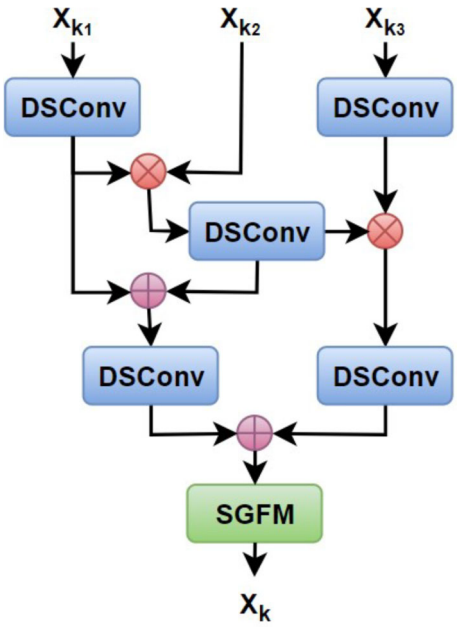


Fig. 6 Illustration of GAFM, which extracts and fuses the attention-based spatial and temporal information

followed by BN and ReLU activation layers with skip connections to preserve the quality of the features. Generally, the multi-level (low-level to high-level) spatiotemporal information contains effective semantic global information and characterizes the most important geometric variation features. The skip connection is used locally between the input and output and passes the useful information directly to the extractor blocks. The procedure is given as follows,

$$\begin{aligned} X_{k_1}^m &= \mathrm{MFEM}(X_{k_1}^g) \\ X_{k_2}^m &= \mathrm{MFEM}(X_{k_2}^g) \\ X_{k_3}^m &= \mathrm{MFEM}(X_{k_3}^g). \end{aligned} \tag{4}$$

3.5 Geometric Attention Fusion Module (GAFM)

In contrast to fusing the multi-modality features using concatenation, graph learning, or attention-based methods [2, 27, 55], we use a hierarchical strategy for fusing the multi-scale geometric and multi-level spatiotemporal features. The architecture of GAFM is shown in Fig. 6, which dynamically adapts the characteristics of DSConv and connects in a hierarchical way to strengthen the geometric transformation modeling and fusion capability of the proposed model. Next, an SGFM is used to extract geometric spatial and temporal information hierarchically and fuse it. The outputs of RFBs, GMEMs,

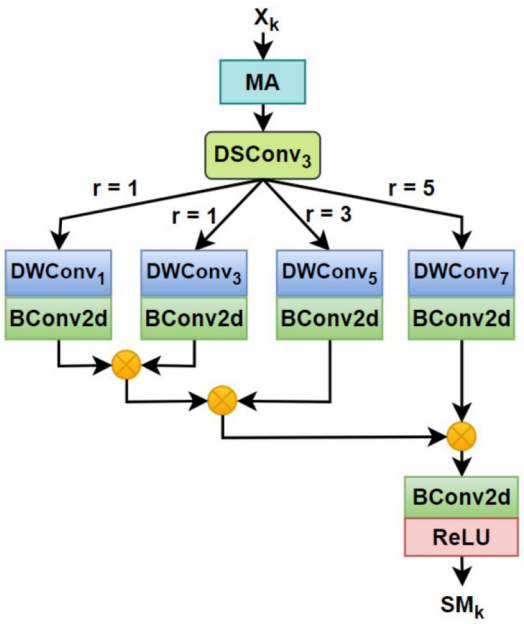


Fig. 7 Illustration of SGM, which converts the high-level semantic of spatial and temporal features into low-level detailed information to generate the saliency maps

and MEFMs are fused as follows,

$$\begin{aligned} X_k^r &= GAFM(X_{k_1}^r, X_{k_2}^r, X_{k_3}^r), \\ X_k^g &= GAFM(X_{k_1}^g, X_{k_2}^g, X_{k_3}^g), \\ X_k^m &= GAFM(X_{k_1}^m, X_{k_2}^m, X_{k_3}^m). \end{aligned} \tag{5}$$

3.6 Saliency Generation Module (SGM)

The boundary contour of an object or region within an image can be considered the demarcation or separation line between the salient (important or foreground) regions and the non-salient (unimportant or background) regions. Boundary information helps to detect and segment the salient objects efficiently. Existing VSD methods [15, 16, 30, 38, 42] use the multiple decoders, stacked of the convolution neural network, graph convolution network, and diffusion module to generate saliency maps, which require more computational complexity and perform poorly on coarse boundaries, deform objects, and tiny objects at multiple scales. To overcome these problems, the Saliency Generation Module (SGM) is proposed as shown in Fig. 7. In SGM, first, the shadow masking is performed using Mask Attention (MA), which uses Spatial Attention (SA) to differentiate the background and foreground shadow boundaries with the help of Sigmoid operation, multiplied with (-1) and addition of $(+1)$ to preserve the foreground regions, and Channel Attention (CA), which performs the Sigmoid operation to get the channel information of the shadow foreground regions and expand the channel of the feature to get the original feature attention. The output of MA is passed to a stack of DSConv, four branches of DWConv with different dilation rates

and filters followed by BasicConv2d, and at last, BasicConv2d, followed by the ReLU activation operation, to improve the region boundary and semantic spatial regions to generate the shadow maps SM_k. The outputs of three different GAFM modules are passed to three SGM modules to generate the three saliency maps. Finally, averaging the outputs of the three SGM modules and generating the final shadow maps SM_k as follows,

$$SM_k = AVG(SGM(X_k^r), SGM(X_k^g), SGM(X_k^m)). \tag{6}$$

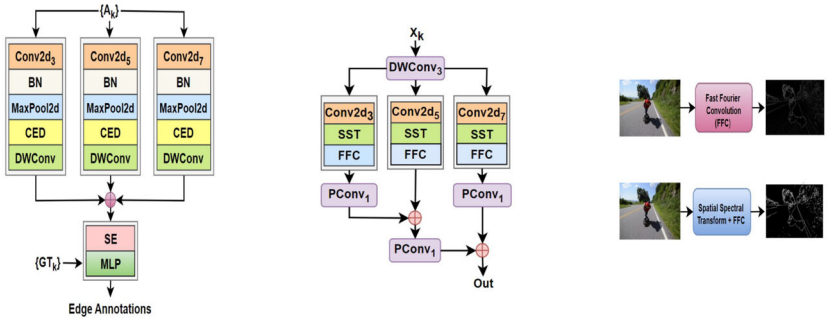
3.7 Edge Detection Network (EDNet)

A newly designed EDNet (shown in Fig. 8a) is used to extract the essential high-level edge information with multiple filter sizes, which was mostly ignored in the existing edge detection models [21, 44, 45, 49, 56]. Moreover, in contrast to the existing models, we are able to use re-parameterizable branches without compromising the cost of increased training time. In EDNet, the appearance frames A_k are passed to three branches with different filter sizes, 3×3, 5×5, and 7×7. Each branch is designed with a stack of convolution, BN, and max-pooling layers. The main idea behind extracting enhanced edges is to adaptively learn the geometric and spectral correction of spatiotemporal edge features at different dilation rates. This is achieved by combining a Spatial-Spectral Transform (SST) with the Fast Fourier Convolution (FFC) [33] operation in a newly designed Contrast Edge Distillation (CED) component. SST enlarges the receptive field of convolution to the full resolution of the input feature map in an efficient way. FFC performs three operations: (1) a local feature is extracted using small-kernel convolution, (2) semi-global features are extracted using convolution operation, and (3) global features are extracted using convolution kernel and converted image-level spectrum. The impact of applying the SST with FFC is shown in Fig. 8c. The detailed architecture of Contrast Edge Distillation (CED) is shown in Fig. 8b.

The edge features at multiple scales are then fused and passed to the saliency edge (SE) module, which combines the Convolution layer followed by Sigmoid operation and sequential layer with 3×3 filter to generate the edge region-based spatio-temporal information. At last, a Multi-Layer Perceptron (MLP) is used, which performs a sequential, fully connected convolution operation to classify and generate the edge maps. For calculating the edge loss, binary cross-entropy is used as follows,

$$l_k = -\frac{1}{N}\sum_{i=1}^{N} X_i \, log(p(X_i)) + (1 - X_i) \, log(1 - p(X_i)), \tag{7}$$

where X_i = input frame, $p(X_i)$ = predicted frame, l_k = loss at k^{th} sample and, N = the total number of frames at k^{th} sample. Finally, EDNet generates the refined representation of the edge map.



(a) Edge Detection Network (EDNet) (b) Contrast Edge Distillation (CED) (c) Comparison of SST and FFC result

Fig. 8 Illustration of (**a**) EDNet, (**b**) CED. Here, Conv2d is the convolution 2D layer, MaxPool2d is the max pooling 2D layer, BN is the batch normalization layer, CED is the contrast edge distillation, DWConv is the depth-wise convolution layer, SE is the saliency edge, MLP is the Multi-Layer Perceptron and $\oplus$ is the element-wise addition operation. In (**c**), the impact of Spatial Spectral Transformation (SST) with the Fast Fourier Convolution layer (FFC) is shown

3.8 Shadow Region Intensity Loss (L_{SRI})

To exploit each pixel region locally and globally and reduce the divergence inconsistency while discriminating the foreground and background region intensity boundary, motivated by [57], a Shadow Region Intensity Loss (SRI) is proposed, which is a hybrid of weighted versions of the Binary Cross Entropy (BCE) [5], Intersection Over Union (IoU) [3], and L1 Loss [44]. A pixel intensity weight, $w_{i,j}$ is calculated, which helps to overcome the difficulty of differentiating the coarse background and foreground regions. The larger the weight, the more complex it is to differentiate the pixels. In other words, the $w_{i,j}$ shows the important region pixel, calculated between the center pixel region intensity and its neighbors. Shadow region intensity is calculated in two steps:

1. The non-linear log transformation is performed on the predicted shadow map (SM_k) to effectively indicate the important local pixel-level region intensity information and differentiate the foreground and background region intensity. The procedure is as follows,

$$T_{SM_k} = \left(\sum_{k \in K} \frac{log(SM_k + 1)}{log(1 + max(SM_k))} \times 255\right) \tag{8}$$

2. The transformed output T_{SM_k} is divided after applying the Sigmoid operation on annotation maps GT_k and multiplied with override weight $(1 - \eta)$ to generate the shadow intensity weight $w_{i,j}$, as follows,

$$w_{i,j} = (1 - \eta) \sum_{k \in K} \frac{T_{SM_k}}{Sigmoid(GT_k)} \tag{9}$$

Here $(1-\eta)$ is calculated with the help of average pooling operation on annotation maps (GT_k using K number of the filter sizes = 1, 10, 20), and $w_{i,j}$ lies between [0, 1]. The shadow region intensity weight $w_{i,j}$ is assigned to each region pixel, and for difficult (noisy, coarse, and cluttered, etc.) region pixels, the $w_{i,j}$ is a high and simple region low.

For local region intensity structure information, weighted BCE (L_{wBCE}) loss is calculated as follows,

$$L_{wBCE}\left(SM_k, GT_k^a\right) = -\sum_{i=1}^{H}\sum_{j=1}^{W}(1-w_{i,j})(SM_k log(GT_k^a) + (1-SM_k) log(1-GT_k^a)) \tag{10}$$

where SM_k is the saliency map, GT_k^a denotes the annotation maps of k^{th} sample, and $w_{i,j}$ denotes pixel region intensity which is calculated using Eq. 9. For global region intensity structure information, weighed IoU (L_{wIoU}) loss is used, which gives the proper guidance to the network about the clear salient shadow details. The weighted IoU loss is calculated as follows,

$$L_{wIoU}\left(SM_k, GT_k^a\right) = 1 - w_{i,j}\frac{\sum_i^H \sum_j^W SM_k GT_k^a}{\sum_i^H \sum_i^W \left(SM_k + GT_k^a - SM_k GT_k^a\right)} \tag{11}$$

The weighed L1 (L_{wL1}) is used to remove the inconsistency between foreground and background regions and calculated as follows,

$$L_{wL1}\left(SM_k, GT_k^a\right) = \sum_{i}^{H}\sum_{j}^{W}|SM_k - GT_k|\left(1-w_{i,j}\right) \tag{12}$$

In the L_{wBCE}, L_{wIoU}, and L_{wL_1}, each pixel of shadow regions is assigned a weight $w_{i,j}$ to calculate the pixel intensity, as follows

$$L_{SRI}(SM_k, GT_k^a) = L_{wBCE}(SM_k, GT_k^a) + L_{wIoU}(SM_k, GT_k^a) + L_{wL1}(SM_k, GT_k^a) \tag{13}$$

Finally, the total loss is calculated with the SRI loss of multi-level multi-scale geometric saliency map $L_{SRI_k^a}\left(SM_k, GT_k^a\right)$ and edge enhance maps $L_{SRI_k^e}\left(Y_k, E_k^e\right)$. Here, T is the total number of iterations, Y_k is the edge enhance feature maps, and E_k^e is the corresponding annotations.

$$L_{SRI} = \sum_{k=1}^{T} L_{SRI_k^a}\left(SM_k, GT_k^a\right) + L_{SRI_k^e}\left(Y_k, E_k^e\right) \tag{14}$$

Table 1 Comparative analysis of the Proposed HSNet model, twelve SOTA VSD Models, and twelve Lightweight VSD Models across three datasets

	Model Yr. Ref.	Backbone Network	# Param (M)	FLOPs (G)	MPE (m)	Speed (FPS)	Our Samples S_α	Our Samples F_β	Our Samples MAE	ViSha S_α	ViSha F_β	ViSha MAE	**VISAD** S_α	**VISAD** F_β	**VISAD** MAE
L-VSD Models	$BDRAR_{18}$ [58]	ResNeXt-101	65.8	22.36	10.34	27.8	0.892	0.875	0.032	0.881	0.865	0.053	0.543	0.504	0.088
	$Triple\text{-}Unet_{20}$ [59]	Unet	75.2	24.53	5.85	45.2	0.895	0.865	0.030	*0.914*	0.880	0.035	0.601	0.590	0.098
	$GMTI_{21}$ [60]	VGG-16	92.1	32.43	10.83	32.1	0.857	0.811	0.052	0.829	0.819	0.043	0.509	0.488	0.120
	$GCN\text{-}based_{22}$ [42]	MobileNet-V2	*2.9*	*6.12*	3.8	**174.6**	0.822	0.813	0.060	0.816	0.807	0.049	0.523	0.503	0.099
	$LAB\text{-}Net_{22}$ [43]	MIMO+3D-CNN	**0.36**	28.43	***2.50***	*154.0*	0.846	0.817	0.049	0.828	0.809	0.038	0.629	0.601	0.083
	$STICT_{22}$ [23]	ResNet-50	62.2	24.28	6.28	22.2	0.835	0.803	0.040	0.891	0.802	0.034	0.673	0.646	0.065
	$VS\text{-}Net_{22}$ [61]	ResNet-50	10.9	8.7	*2.41*	66.0	0.900	0.880	0.031	0.860	0.841	0.040	0.700	0.670	0.064
	$STANet_{22}$ [25]	RexNeXt-101	78.4	33.93	8.65	23.4	0.885	0.862	***0.029***	0.865	0.853	0.032	0.708	*0.683*	*0.059*
	$DSNet_{23}$ [62]	ResNet-50	25.5	18.5	2.80	80.0	0.904	0.879	0.032	0.889	0.870	0.030	0.705	0.677	***0.062***
	$LDRA_{23}$ [24]	MobileNetv2	***3.9***	**0.68**	***2.62***	37.6	0.874	0.858	0.039	0.860	0.848	0.035	0.581	0.470	0.136
	$ShadowSAM_{23}$ [30]	ViT	7.6	9.41	5.5	44.6	0.833	0.819	0.056	0.813	0.789	**0.024**	0.621	0.576	0.072
	$DSFNet_{24}$ [63]	DRNet	23.4	20.0	2.60	84.0	***0.914***	***0.890***	***0.029***	0.901	***0.896***	**0.024**	***0.709***	0.805	0.054
VSD Models	$FEELVOS_{19}$ [64]	DeepLabv3+	129.7	23.87	6.84	29.7	0.854	0.821	0.037	0.863	0.843	0.033	0.498	0.473	0.132
	$DSDNet_{19}$ [6]	ResNext-101	133.3	27.86	7.65	33.7	0.832	0.796	0.044	0.812	0.788	0.052	0.438	0.408	0.068
	$SAR\text{-}GMTI_{19}$ [65]	ResNet-50	155.6	29.81	8.49	25.6	0.830	0.808	0.043	0.836	0.811	0.045	0.418	0.399	0.139
	GFM_{20} [66]	WASM	139.8	30.27	8.45	29.8	0.795	0.782	0.058	0.796	0.754	0.059	0.581	0.549	0.113
	$TVSD\text{-}Net_{21}$ [7]	ResNeXt-101	148.7	25.27	6.25	38.7	0.757	0.750	0.064	0.781	0.770	0.030	0.662	0.634	***0.062***
	$STF\text{-}Net_{22}$ [41]	Res2Net	162.7	144.29	5.26	30.0	0.828	0.801	0.046	0.794	0.761	0.047	0.632	0.612	0.076
	$S\text{-}DNet_{22}$ [3]	ResNet-50	165.8	153.38	8.06	11.1	0.886	0.852	0.035	0.887	0.877	0.045	0.649	0.623	0.067
	$TCRN_{22}$ [54]	T2T-ViT	103.1	114.48	7.65	28.4	0.840	0.815	0.038	0.845	0.833	0.041	0.588	0.539	0.079
	$SC\text{-}Cor_{22}$ [8]	ResNext-101	109.9	24.34	6.50	29.9	0.817	0.791	0.050	0.832	0.812	0.059	0.623	0.587	0.070
	$SE\text{-}SA\text{-}AAN_{23}$ [2]	ResNet-50	187.4	157.51	5.2	42.3	0.861	0.829	0.047	0.852	0.810	0.051	0.532	0.458	0.143
	$SCOTCH\ and\ SODA_{23}$ [15]	MiT-B3	211.8	126.46	9.15	36.5	0.851	0.824	0.033	0.830	0.793	***0.029***	0.613	0.578	0.079
	$DSCNet+_{23}$ [67]	ResNet-50	100.0	88.3	2.9	45.0	*0.918*	*0.902*	*0.028*	***0.910***	*0.900*	0.030	*0.712*	***0.680***	*0.059*
	HSNet	Modified Efficient	7.2	***6.7***	**2.28**	***90.0***	**0.921**	**0.900**	**0.026**	**0.915**	**0.902**	*0.028*	**0.723**	**0.689**	**0.056**

The top three results are visually highlighted in bold font, italic font, and bolditalic font

Table 2 Comparison of our collected Samples with other datasets

Statistics	ViSha	VISAD	Our Additional Examples
Training Videos	50 (total 4786 frames)	81 (total 12566 frames)	25 videos (total 1870 frames)
Testing Videos	70 (total 6897 frames)	26 (total 3063 frames)	35 videos (total 2335 frames)

Challenging Scenarios	ViSha	VISAD	Our Additional Examples
Small-Object	24	03	40
Geometric Variation	28	10	28
Motion Blur	19	12	24
Defocus	17	10	22

4 Datasets, Experiments and Result Analysis

4.1 Datasets

For the task of Shadow Detection in videos, there are two existing benchmark datasets available in the literature, namely ViSha Dataset [7] and VISAD [23]. The statistics of the datasets are given in Table 2. Note that in VISAD, only 33 videos with 4188 frames are annotated with a pixel-level shadow mask, and the remaining videos are unlabeled, whereas, in ViSha, each frame is annotated with a pixel-level shadow mask. The main challenge is that both datasets provide limited examples for complex scenarios, such as partial occlusion with motion blur, deformation of scenes and objects, and geometric variations.

To make the benchmark more comprehensive, we added new videos reflecting challenging scenarios. This additional dataset is prepared from VOS dataset [68], DAVIS [64], DAVSOD [16], MOT [25], and ISTD [3] and annotated according to complex scenarios of the shadow visualization. It has a total of 60 videos with 4205 frames; the minimum duration of the video is 1 s, and the maximum duration is 5 s. The training and testing data statistics are provided in Table 2. The dataset is manually annotated using the Labelbox tool.[1]

4.2 Experimental Setup and Performance Measure

All the experiments are performed on a 64-bit Ubuntu 18.04 operating system. The GPU system has 32GB of RAM, a 16GB NVIDIA P5000/PCIe/SSE2 GPU, and a 1 TB hard disk with 200 GB SSD. The GPU machine is equipped with Anaconda 3.7 and PyTorch version 1.10.0, which utilizes CUDA 10.2 and NVIDIA Driver 470. All input frames are uniformly resized to dimensions of 352×352. The model has been trained using the Adam optimizer with a weight decay of $5e^{-4}$ and a learning rate of $3e^{-4}$. Data augmentation techniques such as flipping and rotation are applied during training. The performance of the proposed model has been assessed using various metrics, including S-measure (S_α), F-measure (F_β), Mean Absolute Error (MAE), Balance Error Rate

[1] https://labelbox.com/

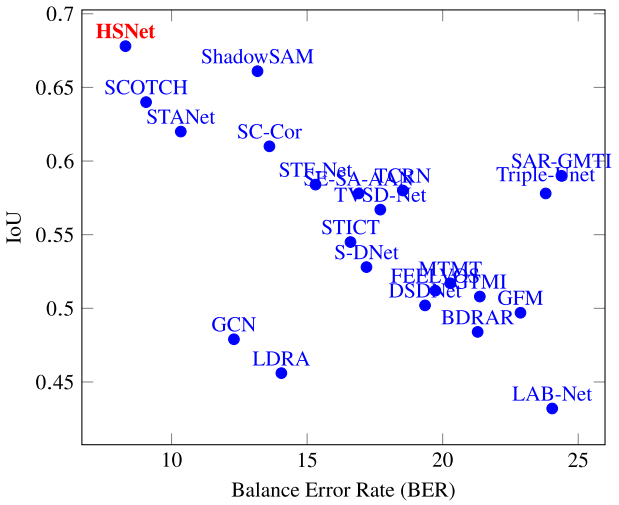


Fig. 9 Performance comparison between proposed HSNet model and SOTA models in terms of IoU and BER on ViSha dataset

(BER), and IoU, as described in [15, 25, 54, 69]. The computational complexities of various models are estimated using the number of network parameters in Millions (M), floating point operations in Gigabits (GFLOPS), training minutes per epoch (MPE), and latency speed in frames per second (FPS).

4.3 Training and Testing Performance

At first, the edge maps are generated using appearance and ground-truth frames via the proposed EDNet model. Next, the proposed HSNet model extracts attention-based, multi-scale, multi-level geometric spatiotemporal features from appearance frames and preserves the edge maps. The training dataset comprised 6656 frames (1870 frames from our new set of samples [25 videos] and 4786 frames from ViSha [50 videos]), with the remaining videos constituting the testing dataset. This training dataset is used to train the proposed model using the Adam optimizer to minimize the total L_{SRI} loss, which requires approximately 8 h to complete 100 epochs with a batch size of 8. Further, data augmentation and preprocessing techniques, including image cropping, resizing, rotation, normalization, and geometric transformation, are used. For evaluating the performance of the proposed HSNet, the three datasets: ViSha [7], VISAD [23], and our collected (new set of samples includes 35 videos) are used. Performance is evaluated in terms of S_α, F_β, and MAE, as well as computational metrics, including network parameters in millions (M), floating-point operations in gigaflops (GFLOPS), minutes per epoch (MPE), and latency speed in frames per second (FPS). The testing results are shown in Tables 1 and 3.

Table 3 Performance Comparison of the proposed HSNet with SOTA Edge-based models

Model Yr. Ref.	Backbone	# Param	GFLOPs	MPE	Inference Time	Our Samples			ViSha			VISAD		
	Network	(M)	(G)	(m)	FPS	S_α	F_β	MAE	S_α	F_β	MAE	S_α	F_β	MAE
$EGNet_{19}$ [49]	VGG-16	195.4	138.26	8.29	34.5	0.759	0.754	*0.035*	0.789	0.776	0.054	0.543	0.487	0.134
$PAGE\text{-}Net_{19}$ [56]	VGG-16	187.5	122.14	4.65	22.5	0.816	0.797	0.047	0.814	0.801	0.046	0.601	0.582	0.123
$MTMT_{20}$ [16]	ResNeXt-101	***125.9***	***94.34***	6.08	35.9	0.835	0.829	0.043	*0.904*	*0.890*	0.048	0.456	0.423	0.137
$ENFNet_{21}$ [47]	VGG-16	174.6	108.12	6.12	24.7	0.829	0.810	0.045	0.821	0.809	***0.038***	0.567	0.521	0.112
$EIEF_{22}$ [44]	ResNet-50	162.6	114.28	***4.15***	28.8	0.838	0.812	0.044	0.889	0.851	*0.037*	0.598	0.534	0.106
$STEG\text{-}Net_{22}$ [5]	ResNet-50	152.9	109.36	7.41	36.2	***0.889***	***0.856***	0.047	***0.901***	***0.886***	0.043	0.608	0.588	0.089
$EFDN_{22}$ [70]	EDBB	**0.36**	*14.70*	*4.04*	*59.7*	0.841	0.825	0.057	0.854	0.821	0.060	0.614	0.595	0.082
$SwinNet_{22}$ [45]	Swin-B	198.7	124.3	5.69	33.7	0.839	0.829	0.064	0.849	0.837	0.047	***0.623***	***0.600***	0.067
$D\text{-}ViShaDeRec_{22}$ [19]	D-ISDET	159.5	132.46	5.30	37.3	0.867	0.842	0.057	0.874	0.852	0.068	0.560	0.542	0.109
$ERPNet_{23}$ [46]	ResNet-34	172.6	113.72	4.54	***39.2***	0.837	0.829	0.056	0.846	0.821	0.052	0.618	0.599	*0.054*
$EDQNet_{23}$ [71]	VGG-16	168.6	123.98	6.62	33.6	*0.908*	*0.887*	***0.037***	0.899	0.867	0.039	*0.634*	*0.610*	***0.058***
HSNet	Modified Efficient	*7.2*	**6.70**	**2.28**	**90.0**	**0.921**	**0.900**	**0.026**	**0.915**	**0.902**	**0.028**	**0.723**	**0.689**	**0.056**

The top three results are highlighted in bold, italic, and bolditalic

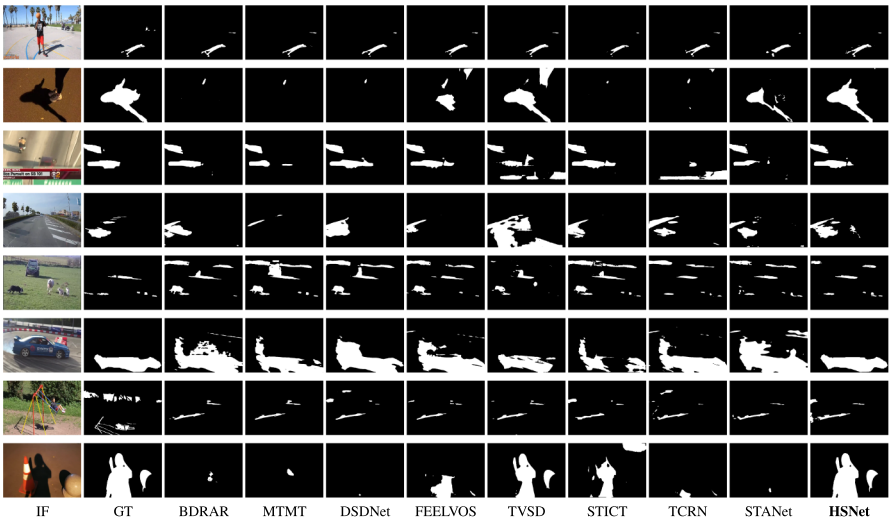


Fig. 10 Performance Comparison of HSNet with SOTA models. IF is the Input Frame, and GT is the corresponding annotation map followed by column-wise results from the BDRAR [58], MTMT [16], DSDNet [6], FEELVOS [64], TVSD [7], STICT [23], TCRN [54], STANet [25], and the proposed HSNet

4.4 Comparative Analysis

The test performance of the proposed model is compared with twenty SOTA models as shown in Table 1. From the table, it is evident that our proposed HSNet model outperforms all the twenty SOTA models in terms of the accuracy-specific metrics, i.e., S_α, F_β, and MAE. In terms of training speed, it also shows the best performance in comparison to the other models. Regarding GFLOPS and inference speed, it has achieved the second and third rank, respectively. Although, the LAB-Net [43] has the least $\#params$, GFLOPS followed by GCN-based [42] and LDRA [24], the % increase in S_α, F_β and MAE is significantly higher (Table 2). Thus, it can be claimed that the proposed model (HSNet) can maintain the trade-off between model complexity and accuracy for the VSD task. In a similar line, in Table 3, demonstrate that the proposed model outperforms with eleven edge-based models without increasing the complexity. Further, as the results are shown in Fig. 9, demonstrate that our proposed model outperforms almost all the measures (including Balance Error Rate (BER) and IoU) and generates the shadow saliency map accurately in much less time.

Additionally, the proposed model is visually compared with six exemplary SOTA methods in Fig. 10 under various difficult shadow situations. The first case is the challenge with motion blur on the top two rows of Fig. 10. Only our strategy in this scenario can effectively and prominently detect the shadow. The third and fourth row demonstrates the difficulty in detection due to deformed backgrounds. The case of partial occlusion, which is shown in the sixth and seventh rows, wrongly includes background or missing objects in their detection. Another challenging case, where objects move very far in the frames, is shown in the eight rows of Fig. 10. Besides our proposed HSNet model, almost every models fail to detect the shadow in these cases.

Table 4 Performance Comparison with Different Backbone Networks

Backbone Network	# Param (M)	GFLOPs (G)	MPE (m)	Inference Time (FPS)	Our Samples		ViSha		VISAD	
					S_α	MAE	S_α	MAE	S_α	MAE
MobileNet-V1 [9]	6.5	5.14	3.03	55	0.865	0.035	0.816	0.048	0.562	0.123
MobileNet-V2 [72]	**4.2**	**3.29**	2.89	69	0.862	0.036	0.832	0.044	0.587	0.112
EfficientNet-V1 (B4) [14]	21.6	17.19	3.01	30	0.884	0.033	0.893	0.041	0.601	0.101
ShuffleNet-V1 [73]	7.7	6.89	2.81	49	0.887	0.041	0.832	0.049	0.597	0.106
ShuffleNet-V2 [10]	9.1	7.92	2.50	45	0.889	0.040	0.847	0.047	0.584	0.110
VGG-16 [74]	5.5	4.8	2.67	49	0.879	0.031	0.876	0.042	0.619	0.098
ResNet-50 [75]	25.8	22.7	3.10	41	0.871	0.041	0.899	0.037	0.639	0.091
ViT [76]	5.6	4.7	2.52	64	0.877	0.039	0.900	0.036	0.650	0.084
EfficientNet-V2 (B4) [31]	7.2	6.7	**2.28**	**90**	**0.921**	**0.026**	**0.915**	**0.028**	**0.723**	**0.056**

4.5 Ablation Study

To study the contribution of each component of the proposed model towards the improvement of VSD performance and reduction of model complexity, we conducted the ablation study on all the datasets.

4.5.1 Comparison with Lightweight Backbone Networks

For a comprehensive qualitative comparison, we adopted various types of lightweight backbone networks, including MobileNet-V1 [9], MobileNet-V2 [72], ShuffleNet-V1 [73], ShuffleNet-V2 [10], EfficientNet-V1 [14], EfficientNet-V2 [31], ViT [76], VGG-16 [74], and ResNet-50 [75], in addition to the Modified EfficientNet-V2 B4 for the proposed HSNet model. The experiments are performed in the same training environments and performance results are shown in Table 4. From the Table 4 results, it is observed that Modified EfficientNet-V2 B4 on the proposed HSNet model gives better results than SOTA lightweight models for VSD tasks due to extraction of multi-scale, multi-level, attention-based geometric spatiotemporal information hierarchically.

4.5.2 Effectiveness of the Proposed Components

To illustrate the effectiveness of each module, experiments are carried out on various combinations and their combination results are shown in Table 5. From the table results, it is observed that the default setting means the baseline performance of the proposed model is very far from the individual performance of GAFM, GAIM, SGFM, GMEM, MFEM, and SGM components. The individual component does not increase the performance of video shadow detection. But, as the components are added to the proposed model the performance is increased slowly, which is shown in Table 7. The combination of component settings is shown in Table 5, which demonstrates the effectiveness of each component in the proposed model. Further, the effectiveness of the multi-scale blocks RFB, geometric multi-scale encoder modules GMEM, and multi-level feature extraction modules MFEM are shown in Table 5, and demonstrate that as each module increases in the proposed models the computational complexity, as well as the performance, are increased. When modules are more than three the performance has deteriorated, which shows the proposed model is the less constrained of the modules.

4.5.3 Effectiveness of EDNet with HSNet

By observing the shadow frames, we see that the boundaries of soft shadows may not be visible compared to the nearby non-shadow regions. Therefore, the edge information is used to improve the performance of shadow detection. From Table 7, we observe that the proposed EDNet can successfully maintain the long-term dependencies during the geometric low-level spatial and temporal information fusion and it helps to increase the quality of edge detection without increasing the network complexity. The Table 7 shows the effectiveness of the proposed EDNet model to enhance the model performance.

Table 5 Impact of Proposed Modules on Detection Performance

S.No.	Component Setting						Our Samples		ViSha		VISAD	
	GAFM	GMEM	MFEM	GAIM	SGFM	SGM	S_α	MAE	S_α	MAE	S_α	MAE
1	Default						0.834	0.047	0.754	0.038	0.678	0.099
2	✓				✓	✓	0.865	0.035	0.786	0.046	0.685	0.093
3		✓	✓	✓		✓	0.862	0.036	0.782	0.048	0.676	0.096
4	✓		✓	✓	✓		0.867	0.037	0.795	0.045	0.629	0.086
5	✓	✓		✓		✓	0.884	0.033	0.813	0.042	0.701	0.034
6		✓	✓	✓	✓	✓	0.887	0.030	0.832	0.039	0.688	0.037
7	✓	✓	✓	✓	✓		0.889	0.039	0.847	0.035	0.697	0.062
8	✓	✓	✓	✓	✓	✓	**0.921**	**0.026**	**0.915**	**0.028**	**0.723**	**0.056**

Bold values indicate the best values

Table 6 Ablation study on different components of HSNet

# Modules	Configuration Modules			# Param	# FLOPs	MPE	Our Samples		ViSha		VISAD	
	RFB	GMEM	MFEM	(M)	(G)	(m)	S_α	MAE	S_α	MAE	S_α	MAE
1	✓	✓	✓	4.5	3.9	1.50	0.839	0.048	0.759	0.054	0.686	0.096
2	✓	✓	✓	6.0	5.3	1.70	0.867	0.039	0.808	0.047	0.699	0.087
3	✓	✓	✓	**7.2**	**6.7**	**2.28**	**0.921**	**0.026**	**0.915**	**0.028**	**0.723**	**0.056**
4	✓	✓	✓	8.2	7.9	3.01	0.889	0.032	0.876	0.040	0.706	0.078
5	✓	✓	✓	9.7	8.2	4.17	0.900	0.030	0.890	0.037	0.712	0.066
6	✓	✓	✓	10.8	9.9	4.99	0.898	0.032	0.865	0.032	0.715	0.059

Bold values indicate the best values

Table 7 Impact of EDNet Module on Performance of HSNet

	Edge-based Model		Our Samples		ViSha		VISAD	
SNo.	With EDNet	Without EDNet	S_α	MAE	S_α	MAE	S_α	MAE
1		✓	0.895	0.035	0.906	0.032	0.692	0.092
2	✓		**0.921**	**0.026**	**0.915**	**0.028**	**0.723**	**0.056**

Bold values indicate the best values

Table 8 Ablation study on the different loss functions

	Our Samples		ViSha		VISAD	
Loss functions	S_α	MAE	S_α	MAE	S_α	MAE
BCE+IoU	0.861	0.039	0.854	0.042	0.678	0.089
wBCE+wIoU	0.872	0.036	0.862	0.039	0.682	0.072
BCE+IoU+L1	0.883	0.034	0.871	0.035	0.709	0.065
SRI ($\eta = 0.4$)	0.890	0.032	0.888	0.034	0.719	0.058
SRI ($\eta = 0.5$)	**0.921**	**0.026**	**0.915**	**0.028**	0.723	**0.056**
SRI ($\eta = 0.6$)	0.895	0.027	0.897	0.032	**0.732**	0.060
SRI ($\eta = 0.7$)	0.889	0.030	0.878	0.039	0.720	0.057

Bold values indicate the best values

4.5.4 Effectiveness of SRI Loss Function on HSNet

The proposed HSNet model has experimented with various combinations of loss functions and results are furnished in Table 8. The table shows that SRI loss gives more effective results than other loss functions due to the use of region intensity in the BCE, IoU, and L1 Loss functions. Moreover, when L1 loss is added to BCE and IoU, MAE decreases and preserves S_α. However, the SRI loss function decreased MAE compared to the others, and its performance is penalized with a higher η value. The higher η value assigns a high w value to pixels adjacent to straight or fine edges. Hence, we have fine-tuned $\eta = 0.5$ for better video shadow detection.

5 Conclusion

In this paper, the key contribution is a sustainable video shadow detection solution for dynamic scenes. Firstly, we proposed a Hierarchical Separable Network (HSNet), a novel lightweight model equipped with novel plug-ins like GAFM, GMEM, MFEM, SGFM, and SGM. These collectively extract multi-scale geometric high-level spatial and temporal information and generate saliency maps. Furthermore, we introduced EDNet, a novel solution for capturing geometric variations in edge information, a crucial aspect of sustainable video surveillance solutions for IoT devices. To reduce the dataset scarcity challenge in this domain, another contribution of this work is to further enrich the video shadow detection datasets containing various challenging

scenarios. With the help of extensive experimentation and a detailed ablation study, we demonstrate that our proposed HSNet model significantly reduces model complexity while enhancing video shadow detection performance compared to the SOTA models, making it the most suitable choice for sustainable video shadow detection in IoT devices. In the future, our work will continue to advance the field by addressing real-time shadow detection challenges deployed on IoT devices such as Nvidia Jetson Nano and Raspberry Pi, further contributing to a sustainable solution for AI-enabled video surveillance.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability The data used to support the findings of this study are available from the corresponding author upon request. We also confirm that the partial dataset used in this study is publicly available and has been properly cited in the paper.

Declarations

Conflict of interest The authors are certified that there are no affiliations or involvement in any organization for any financial interest such as educational grants, consultancies, patent-licensing arrangements, employment, participation in speakers' bureaus' membership, and other equity interest or non-financial interest such as affiliations, personal or professional relationships, knowledge in the material discussed or subject matter in this manuscript.

References

1. A. Amato, I. Huerta, M.G. Mozerov, F.X. Roca, J. Gonzalez, Moving cast shadows detection methods for video surveillance applications. In *Wide Area Surveillance: Real-time Motion Detection Systems*, Springer, pp 23–47 (2012)
2. J. Bao, X. Zhang, T. Zhang, T. Zeng, Z. Yang, X. Zhan, J. Shi, S. Wei, Shadow-enhanced self-attention and anchor-adaptive network for video sar moving target tracking. IEEE Trans. Geosci. Remote Sens. **61**, 1–13 (2023)
3. J. Bao, X. Zhang, T. Zhang, X. Xu, Shadowdenet: a moving target shadow detection network for video sar. Remote Sens. **14**(2), 320 (2022)
4. P.D. Barua, T. Tuncer, S. Dogan, C.P. Ooi, R.U. Acharya, Novel automated detection of sports activities using shadow videos. Multimed. Tools Appl. **83**(15), 44933–44954 (2024)
5. P.D. Barua, T. Tuncer, S. Dogan, C.P. Ooi, R.U. Acharya, Novel automated detection of sports activities using shadow videos. Multimed. Tools Appl. **83**(15), 44933–44954 (2024)
6. Q. Zheng, X. Qiao, Y. Cao, R.W. Lau, Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 5167–5176 (2019)
7. Z. Chen, L. Wan, L. Zhu, J. Shen, H. Fu, W. Liu, J. Qin, Triple-cooperative video shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 2715–2724 (2021)
8. X. Ding, J. Yang, X. Hu, X. Li, Learning shadow correspondence for video shadow detection. In *European Conference on Computer Vision*, Springer, pp 705–722 (2022)
9. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. N. Ma, X. Zhang, H. T. Zheng, J. Sun, Shufflenet v2: practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp 116–131 (2018)
11. Y. Piao, Y. Jiang, M. Zhang, J. Wang, H. Lu, Panet: patch-aware network for light field salient object detection. IEEE Trans. Cybern. (2021)

12. Z. Chen, L. Wan, Y. Xiao, L. Zhu, H. Fu, Learning physical-spatio-temporal features for video shadow removal. IEEE Trans. Circ. Syst. Video Technol. (2024)
13. Y. Sun, B. Xue, M. Zhang, G.G. Yen, J. Lv, Automatically designing cnn architectures using the genetic algorithm for image classification. IEEE Trans. Cybern. **50**(9), 3840–3854 (2020)
14. M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, PMLR, pp 6105–6114 (2019)
15. L. Liu, J. Prost, L. Zhu, N. Papadakis, P. Liò, C.B. Schönlieb, Aviles-Rivero AI Scotch and soda: A transformer video shadow detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 10449–10458 (2023)
16. Z. Chen, L. Zhu, L. Wan, S. Wang, W. Feng, P.A. Heng, A multi-task mean teacher for semi-supervised shadow detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp 5611–5620 (2020)
17. H. Zhang, Z. Liu, Moving target shadow detection based on deep learning in video sar. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE*, pp 4155–4158 (2021)
18. S. Sahoo, P.K. Nanda, Kde based simultaneous background model learning and entropy based fusion of cascaded features for video object segmentation with shadow removal. IEEE Access (2023)
19. Risnandar, D. Nabila, D-vishaderec: Double intensity of video shadow detection, removal, and recoloring in autonomous vehicle. In *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, pp 281–285 (2022)
20. D. Xu, Z. Wang, Shadow detection with attention feature block and multi-scale weight segmentation network. In *2020 the 6th International Conference on Communication and Information Processing*, pp 43–51 (2020)
21. S. Luo, H. Li, R. Zhu, Y. Gong, H. Shen, Espfnet: an edge-aware spatial pyramid fusion network for salient shadow detection in aerial remote sensing images. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **14**, 4633–4646 (2021)
22. L. Zhu, J. Chen, X. Hu, C.W. Fu, X. Xu, J. Qin, P.A. Heng, Aggregating attentional dilated features for salient object detection. IEEE Trans. Circ. Syst. Video Technol. **30**(10), 3358–3371 (2019)
23. X. Lu, Y. Cao, S. Liu, C. Long, Z. Chen, X. Zhou, Y. Yang, C. Xiao, Video shadow detection via spatio-temporal interpolation consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3116–3125 (2022)
24. M.K. Yücel, V. Dimaridou, B. Manganelli, M. Ozay, A. Drosou, A. Saa-Garriga, Lra&ldra: rethinking residual predictions for efficient shadow detection and removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp 4925–4935 (2023)
25. Z. Chen, X. Lu, L. Zhang, C. Xiao, Semi-supervised video shadow detection via image-assisted pseudo-label generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp 2700–2708 (2022)
26. M. Javaid, M. Maqsood, F. Aadil, J. Safdar, Y. Kim, An efficient method for underwater video summarization and object detection using yolov3. Intell. Autom. Soft Comput. **35**(2) (2023)
27. R. Ran, L.J. Deng, T.X. Jiang, J.F. Hu, J. Chanussot, G. Vivone, Guidednet: a general cnn fusion framework via high-resolution guidance for hyperspectral image super-resolution. IEEE Trans. Cybern. (2023)
28. H. Wei, G. Xing, J. Liao, Y. Zhang, Y. Liu, Structure-aware spatial-temporal interaction network for video shadow detection. In: Larson K (ed) *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, International Joint Conferences on Artificial Intelligence Organization*, pp 1425–1433, https://doi.org/10.24963/ijcai.2024/158, https://doi.org/10.24963/ijcai.2024/158, main Track (2024)
29. J. Lin, J. Shen, X. Yang, H. Fu, Q. Zhang, P. Li, B. Sheng, L. Wang, L. Zhu, Learning motion-guided multi-scale memory features for video shadow detection. IEEE Tran. Circ. Syst. Video Technol. (2024)
30. Y. Wang, S. Ji, Y. Zhang, A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **14**, 8100–8112 (2021)
31. M. Tan, Q. Le, Efficientnetv2: smaller models and faster training. In *International conference on machine learning, PMLR*, pp 10096–10106 (2021)
32. Z. Xing, T. Wang, X. Hu, H. Wu, C.W. Fu, P.A. Heng, Video instance shadow detection. arXiv preprint arXiv:2211.12827 (2022)
33. L. Chi, B. Jiang, Y. Mu, Fast fourier convolution. Adv. Neural Inform. Process. Syst. **33**, 4479–4488 (2020)

34. Y. Wang, S. Ji, Y. Zhang, A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **14**, 8100–8112 (2021)
35. V. Nguyen, T.F. Yago Vicente, M. Zhao, M. Hoai, D. Samaras, Shadow detection with conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp 4510–4518 (2017)
36. X. Hu, L. Zhu, C. W. Fu, J. Qin, P.A. Heng, Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 7454–7462 (2018)
37. Z. Wu, H. Xie, T. Gao, Y. Zhang, H. Liu, Moving target shadow detection method based on improved vibe in videosar images. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. (2024)
38. H. Zhou, H. Wang, T. Ye, Z. Xing, J. Ma, P. Li, Q. Wang, L. Zhu, Timeline and boundary guided diffusion network for video shadow detection. arXiv preprint arXiv:2408.11785 (2024)
39. L. Jie, H. Zhang, Rmlanet: random multi-level attention network for shadow detection and removal. IEEE Trans. Circ. Syst. Video Technol. **33**(12), 7819–7831 (2023)
40. W. Zhang, X. Zhang, X. Xu, Y. Xu, Z. Shao, J. Shi, S. Wei, T. Zeng, Gnn-jfl: graph neural network for video sar shadow tracking with joint motion-appearance feature learning. IEEE Trans. Geosci. Remote Sens. (2024)
41. J. Lin, L. Wang, Spatial-temporal fusion network for fast video shadow detection. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pp 1–5 (2022)
42. W. Wu, K. Zhou, X.D. Chen, J.H. Yong, Light-weight shadow detection via gcn-based annotation strategy and knowledge distillation. Comput. Vision Image Understand. **216**, 103341 (2022)
43. B. Hou, Y. Liu, N. Ling, L. Liu, Y. Ren, A fast lightweight 3d separable convolutional neural network with multi-input multi-output for moving object detection. IEEE Access **9**, 148433–148448 (2021)
44. Y. Shi, G. Qin, Y. Liang, X. Wang, J. Yan, Z. Zhang, Salient object detection based on edge-interior feature fusion. IET Image Process. **17**(2), 337–348 (2023)
45. Z. Liu, Y. Tan, Q. He, Y. Xiao, Swinnet: swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Trans. Circ. Syst. Video Technol. **32**(7), 4486–4497 (2021)
46. X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, C. Yan, Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. IEEE Trans. Cybern. **53**(1), 539–552 (2022)
47. Z. Tu, Y. Ma, C. Li, J. Tang, B. Luo, Edge-guided non-local fully convolutional network for salient object detection. IEEE Trans. Circ. Syst. Video Technol. **31**(2), 582–593 (2020)
48. X. Qin, D.P. Fan, C. Huang, C. Diagne, Z. Zhang, A.C. Sant'Anna, A. Suarez, M. Jagersand, L. Shao, Boundary-aware segmentation network for mobile and web applications. arXiv preprint arXiv:2101.04704 (2021)
49. J.X. Zhao, J.J. Liu, D.P. Fan, Y. Cao, J. Yang, M.M. Cheng, Egnet: edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp 8779–8788 (2019)
50. L. Jiao, L. Huo, C. Hu, P. Tang, Z. Zhang, Permutohedral refined unet: Bilateral feature-scalable segmentation network for edge-precise cloud and shadow detection. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. (2024)
51. B. Dong, G. Weng, Q. Bu, Z. Zhu, J. Ni, An active contour model based on shadow image and reflection edge for image segmentation. Expert Syst. Appl. **238**, 122330 (2024)
52. S. Liu, D. Huang, Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pp 385–400 (2018)
53. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Kaiser Ł, Polosukhin I, Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
54. K. Huang, C. Tian, J. Su, J.C.W. Lin, Transformer-based cross reference network for video salient object detection. Pattern Recog. Lett. (2022)
55. R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, X. Luo, Madnet: a fast and lightweight network for single-image super resolution. IEEE Trans. Cybern. **51**(3), 1443–1453 (2020)
56. W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1448–1457 (2019)

57. J. Wei, S. Wang, Q. Huang, F^3net: fusion, feedback and focus for salient object detection. Proc. AAAI Conf. Artif. Intell. **34**, 12321–12328 (2020)
58. L. Zhu, Z. Deng, X. Hu, C.W. Fu, X. Xu, J. Qin, P.A. Heng, Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 121–136 (2018)
59. X. Wu, M. Li, X. Lin, J. Wu, Y. Xi, X. Jin, Shallow triple unet for shadow detection. In *Twelfth International Conference on Digital Image Processing (ICDIP 2020), International Society for Optics and Photonics*, vol 11519, p 1151902 (2020)
60. C. Zhong, J. Ding, Y. Zhang, Joint tracking of moving target in single-channel video sar. IEEE Trans. Geosci. Remote Sens. (2021)
61. H. Singh, M. Verma, R. Cheruku, Vs-net: multiscale spatiotemporal features for lightweight video salient document detection. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, pp 1307–1311 (2022)
62. H. Singh, M. Verma, R. Cheruku, Dsnet: Efficient lightweight model for video salient object detection for iot and wot applications. Companion Proceedings of the ACM Web Conference **2023**, 1286–1295 (2023)
63. H. Singh, M. Verma, R. Cheruku, Dsfnet: video salient object detection using a novel lightweight deformable separable fusion network. IEEE Trans. Instrum. Measur. (2024)
64. P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, L.C. Chen, Feelvos: fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 9481–9490 (2019)
65. Z. Liu, D. An, X. Huang, Moving target shadow detection and global background reconstruction for videosar based on single-frame imagery. IEEE Access **7**, 42418–42425 (2019)
66. H. Shi, C. Liu, A new cast shadow detection method for traffic surveillance video analysis using color and statistical modeling. Image Vision Comput. **94**, 103863 (2020)
67. H. Singh, M. Verma, R. Cheruku, Novel dilated separable convolution networks for efficient video salient object detection in the wild. IEEE Trans, Instrum. Measur. (2023)
68. Y. Gu, L. Wang, Z. Wang, Y. Liu, M.M. Cheng, S.P. Lu, Pyramid constrained self-attention network for fast video salient object detection. Proc. AAAI Conf. Artif. Intell. **34**, 10869–10876 (2020)
69. M.M. Cheng, D.P. Fan, Structure-measure: a new way to evaluate foreground maps. Int. J. Comput. Vis. **129**(9), 2622–2638 (2021)
70. Y. Wang, Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 777–785 (2022)
71. K. Xu, J. Guo, A novel edge-inspired depth quality evaluation network for rgb-d salient object detection. J. Grid Comput. **21**(3) (2023)
72. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4510–4520 (2018)
73. X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: an extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 6848–6856 (2018)
74. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
75. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778 (2016)
76. C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, A. Yuille, Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 11998–12008 (2022)