International Journal of Multimedia Information Retrieval (2025) 14:12
https://doi.org/10.1007/s13735-025-00361-z

REGULAR PAPER

DMFNet: geometric multi-scale pixel-level contrastive learning for video salient object detection

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

Received: 23 July 2024 / Revised: 12 February 2025 / Accepted: 19 February 2025 / Published online: 10 March 2025

Abstract

For video salient object detection (VSOD) tasks, the geometric variations of object foregrounds and backgrounds across multiple scales pose significant challenges for deep learning models in extracting and integrating semantic features from video streams. Current deep learning approaches, such as recurrent neural networks and transformers, struggle to capture both short- and long-term temporal dependencies at a global level due to their fixed kernel structures. Additionally, these methods are computationally intensive, limiting their practical application. To address these challenges and achieve a balance between accuracy and computational efficiency, a novel lightweight Deformable Multi-scale Fusion Network is proposed, which extracts both attention-based multi-scale features and geometric features together to generate the efficient saliency map. Further, the Geometric Multi-Scale Pixel-level Contrastive Learning (GMPCL) approach, which enhances the geometric representation of features is proposed using GMPCL loss and separates the geometric representations of foreground and background features of objects at the pixel level. The performance evaluation is done on six benchmark datasets and compared with twenty-two state-of-the-art (SOTA) models. The main highlight of this work is that it performs well on most challenging datasets DAVSOD-Difficult as compared to SOTA models and has 6.2 million network parameters, 5.6 G FLOPS, and 90 FPS inference speed.

Keywords Video salient object detection · Deformable convolution · Multi-scale geometric feature · Deformable attention-based encoder module · Deformable atrous attention module · Deformable fusion network · Geometric multi-scale pixel-level contrastive learning

1 Introduction

Telecommunication networks empower the massive amount of video streaming data collected from multiple Industrial Internet of Things (IIoT) [29] devices and can be utilized for various computer vision (CV) tasks. Few application areas include autonomous cars [60], robotic manipulation [31], medical image segmentation [14], surveillance system [27], smart agriculture [57], smart traffic management [23], smart home [61], and many more. Video salient object detection (VSOD) is a crucial pre-processing component of computer vision systems, which extracts visually distinctive objects in a video stream. For instance, to extract suspicious activities in smart home systems, surveillance cameras employ VSOD components to detect and segment questionable or unusual objects. Deep learning models are proven to be SOTA in VSOD [6, 24, 42]. However, these models are not feasible to be applied in resource constraints environments, such as in Internet of Things (IoT)-based applications [29], because of the large requirement of computational resources, storage capacity, training data, poor camera quality, and power consumption (battery capacity is less).

To overcome the above challenges, designing new light weight deep learning models for multiple computer vision tasks is gaining more popularity. Recently, many lightweight SOD models [6, 19, 45] are proposed, which mainly focus on employing lightweight backbone networks, such

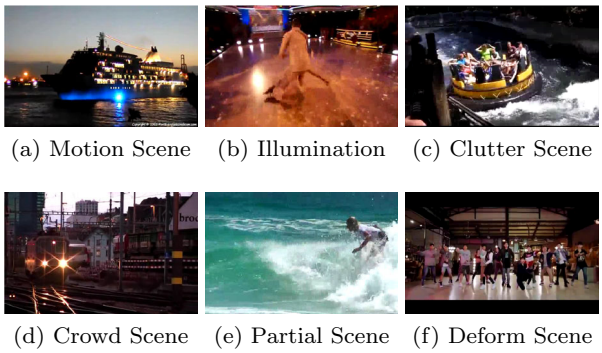
✉ Ramalingaswamy Cheruku
rmlswamy@nitw.ac.in

Hemraj Singh
720079@student.nitw.ac.in

Mridula Verma
vmridula@idrbt.ac.in

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

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



(a) Motion Scene (b) Illumination (c) Clutter Scene

(d) Crowd Scene (e) Partial Scene (f) Deform Scene

Fig. 1 Challenging scenarios from DAVSOD-Diff [18] and DAVSOD-Normal datasets

as MobileNets [57], VGG-16 [19, 58], EfficientNet [87], ResNet [21] and ShuffleNets [91] to capture salient objects. Other works focus on knowledge distillation [90, 96], quantization [13, 34], and model pruning [9, 33] to capture salient objects using lesser parameters. However, the area of designing lightweight models specific to the VSOD task, which is a more complex task due to the temporal dependencies among the image frames, has not been explored much in the literature.

In another direction, the conventional VSOD methods capture either spatial or temporal features separately [22, 38, 67], which leads to poor performance in challenging scenarios like low-light, camera motion, partial occlusion, object deformation, noise, crowd scene, and cluttered background as shown in Fig. 1. Thus, it is essential to use spatiotemporal modality. Multiple spatiotemporal fusion-based models [8, 10, 35, 36, 85] have been proposed, which fuse homogeneous spatiotemporal features for sharing the correlation patterns of spatial and temporal features of object perceptions, such as shape, movement, and structure. The limitation of these models is that they generate hefty model weights, which makes it difficult to deploy in resource-constrained environments, like edge devices and mobile applications.

To solve this issue and follow the current trend, new lightweight models are being designed for VSOD tasks. In [23], authors proposed a dual-stream network, which learns the appearance and motion features from two different modules, increases the storage space and decreases the latency. In [59], the VS-Net model is designed to extract the multi-scale spatiotemporal features using a long skip-connection between the encoder block and the decoder block. These models are unable to handle the geometric variations of spatial appearance and temporal locality. In [60], a deformable separable network is proposed, which extracts the geometric variation-based spatial and temporal features and overcomes the problem of long skip-connection and short skip-connection problems using a new module in the intermediate node.

Recently, in [25], authors designed a lightweight neural network using ShuffleNet-V2 [91] to extract the deep features. This model extracts combined multi-modal features with multi-scale spatial context. Authors in [66] designed a lightweight network using the concepts of knowledge distillation to transfer the knowledge of the teacher with the heavy model to the student network with the light model and check student learning capability using similarities calculation. Due to the lightweight design, a student fails to extract multi-scale features, which degrades the overall performance. Using a similar approach, authors in [24] designed a lightweight model using multiple heterogeneous decoders in the student network. However, due to the random initialization of the kernel order of the student network, this model is unable to recognize the informative patterns while reducing the overall model complexity. However, these models are not able to capture the local and global contextual features from spatial and temporal features. Hence are unable to balance the trade-off between accuracy and the number of network parameters. To overcome the above problems and balance the performance and number of network parameters, we propose a novel lightweight Deformable Multi-Scale Fusion Network *(DMFNet)*, which extracts efficient multi-scale geometric features at the pixel level without escalating network parameters. Hence, it is suitable for IoT applications. Our model is designed with the help of the Deformation Attention Encoder Module (DAEM), Deformable Atrous Attention Module (DAAM), Deformable Fusion Network (DFNet), and Receptive Field Blocks (RFB) [43] to extract efficient multi-scale geometric features across spatial and temporal features for efficient VSOD. The existing contrastive learning [3, 48, 78, 82, 95] methods utilize contrastive pre-training at pixel-level to optimize the dense spatial and temporal features and other methods use region mining algorithms [1, 81, 92] extracts region-based semantic features at image-level. These models pose two major challenges in dynamic scene environments. (1) It is unable to extract rich semantic features from the geometric variation changes of background/foreground and object in the video scene at random multi-scale and multi-object scenes. (2) The random view at multi-scale can be semantically inconsistent when the dynamic scene is geometrically changed and causes problems in finding out the correlation between the different objects or objects foreground and the background from the same and different frames. To overcome the above problems, the Geometric Multi-Scale Pixel-level Contrastive Learning (GMPCL) approach is proposed, which enhances the geometric representation of features and helps to differentiate the foreground and background of objects based on similarity match at each pixel level. In this process, the geometric representations of foreground region pairs within the same video are encouraged to attract each other, while the geometric representations of foreground-background region pairs are

pushed apart in the latent space using GMPCL loss. The proposed DMFNet model has 6.2 million network parameters, 5.6 Gigabyte floating point operation, and 90 FPS inference speed. Our work makes several key contributions, which are summarized as follows:

1. A novel, efficient, lightweight DMFNet model is proposed, which is equipped with an RFB [43], DAEM, DAAM, and DFNet to extract the attention-based multi-scale geometric spatiotemporal features and reduces the skip connection between two nodes.
   - To extract the attention-based multi-scale geometric spatiotemporal together, DAEM is proposed.
   - The DAAM extracts multi-scale geometric spatiotemporal features, preserves the short and long-term temporal dependency and enhances the localization of the objects.
   - Further, the DFNet is used, which predicts the saliency map efficiently and propagates depth-wise low-rank spatiotemporal features.
2. The proposed GMPCL technique helps the DMFNet model extract the salient geometric representation of multi-scale spatiotemporal features without increasing the network parameters using GMPCL loss.
3. Through extensive experiments, we demonstrate that the proposed DMFNet model surpasses the performance of six datasets across various evaluation metrics (including S_α, F_β, and MAE) and # network parameters, # FLOPs, and speed.

The upcoming section is described as follows: first, the related work; second, the proposed approach; third, the experiment result; and finally, the conclusion is discussed.

2 Related work

This section overviews the SOTA contrastive learning methods, deformable convolution methods, multi-scale feature methods, and lightweight architectures.

2.1 Contrastive learning

Recently, contrastive learning-based methods have gained popularity and are being explored in different frameworks, including self-, semi- or unsupervised learning frameworks. It learns the semantic features while differentiating the positive samples from several negative samples using similarity-matching techniques. In this, most of the methods [1, 81, 92] use the negative image pairs, and some are [48, 74, 80, 82, 95] at the pixel level or region level to differentiate the positive samples. Zhong et al. [95] use l2 loss and contrastive loss at a pixel level to address the false negative noise and computational issues. Wang et al. [80] propose a DenseCL to optimize the contrastive (dis)similarity loss of the image pair views at the pixel level. Pang et al. [48] introduce a PixCon framework, which uses two different matching similarity methods to learn the positive information and the negative information at the pixel level. Chen et al. [8] present a VSOD framework that employs a non-local self-attention mechanism and contrastive learning to enhance spatiotemporal feature representations. By integrating co-attention for multi-level feature fusion and intra/inter-frame contrastive losses to achieve temporal consistency and precise foreground-background separation. Tu et al. [70] present the self-supervised cross-view representation reconstruction (SCORER) network, which employs multi-head token-wise matching and contrastive alignment, enabling view-invariant image representations for stable caption generation. Furthermore, a backward reasoning cross-modal module refines the captions by modeling and aligning the 'before' and 'after' representations, improving their informativeness. Wu et al. [82] design pixel-wise contrastive learning (PCL) strategies to enhance the intra-pixel density and inter-pixel distinction for generating efficient feature maps. Tu et al. [71] introduce a Syntax-Calibrated Multi-Aspect Relation Transformer (SMART) to capture robust change features across diverse scenes and enable reliable cross-modal alignment for change captioning. It employs multi-aspect relation learning to disentangle fine-grained changes, ensure view-invariant representation, and integrate semantic change priors. Additionally, a Part-of-Speech (POS)-based visual switch dynamically calibrates the transformer decoder, enhancing syntax-aware alignment for generating linguistically rich change captions. Tu et al. [72] introduce the Context-Aware Difference Distilling (CARD) network, which decouples context features into common and difference contexts, applying consistency and independence constraints to effectively capture and distill all genuine changes between image pairs. This distilled change representation is then transformed into linguistic sentences using a transformer decoder. Tu et al. [73] present a self-supervised distractor-immune representation learning network that decorrelates and correlates image channels to stabilize representations under distractors. This enables enhanced interaction between representations for capturing reliable features essential for caption generation. Additionally, a cross-modal contrastive regularization is introduced to optimize alignment between attended difference features and generated words. Wang et al. [76] present a dual-branch dynamic selection-fusion network (DSFNet), integrating spatial saliency learning and dynamic spatiotemporal contrast via optical flow for VSOD. Advanced modules, including the Contrast Transformation Mdule (CTM), Contrast Analysis Module (CAM), and Selection Guidance Module (SGM), enable precise feature selection and refinement

for enhanced detection. However, these methods are unable to differentiate the multi-objects or object foreground and background when the object dynamically changes its position at a multi-scale pixel-level. To overcome this problem, we propose a multi-scale geometric pixel-level contrastive learning (GMPCL).

2.2 Deformable convolution-base approaches

Deformable convolution-based approaches [12, 14, 60, 79, 97] have been developed to adaptively capture geometric spatial features such as the structure of objects, edges, contour variations, shapes, angles, orientations, etc. Traditional Convolutional Neural Networks (CNNs) rely on fixed kernel structures to extract spatial features such as texture, colors, regions, and pixels. To address this limitation, Wang et al. [79] introduced the InternImage, a Vision Transformer (ViT)-based technique that creates a large effective receptive field, enhancing detection and segmentation tasks. Similarly, Deng et al. [14] proposed the Spatio-Temporal Deformable Convolution (STDC) to effectively capture and fuse motion features. Earlier, Dai et al. [12] presented a deformable convolution network that addresses geometric spatial structures using convolution offsets, but it primarily supports spatial structures and struggles with recognizing regions of interest. Zhu et al. [97] improved upon this with Deformable Convnets v2, incorporating an additional modulation mechanism to enhance region-level modeling capabilities. However, this approach still falls short in managing long-range dependencies of spatial and temporal features. To overcome these challenges, Singh et al. [60] developed DSNet, which extracts attention-based spatial and temporal features without increasing model parameters, thereby providing a more efficient and robust solution.

2.3 Multi-scale feature fusion approaches

The field of VSOD has seen significant advancements with the emergence of multi-scale feature extraction and fusion-based methodologies [5, 11, 32]. These approaches effectively capture rich information from video frames to identify salient objects while preserving module-specific details through layered architectures. However, a critical challenge faced by these models is the tendency to forget past module information, leading to increased computational complexity. To address these issues, Deng et al. [14] introduced the STDF network, which adeptly extracts multi-scale appearance and motion features. Meanwhile, Cong et al. [11] devised PSNet, a parallel multi-scale spatiotemporal extractor, and Ji et al. [32] proposed CASNet, which similarly focuses on spatiotemporal features extraction. Despite their efficacy, these models often incur high computational costs due to online optimization strategies. Moreover, Xu et al. [84] put forth a two-stream network designed to identify high-saliency locations and track salient objects across consecutive frames. However, these models typically overlook spatial and temporal distributions, resulting in imbalanced performance and model complexity. Yue et al. [89] introduce Multigrained Representation Aggregating Transformer (MURAT), a full-attentive network that effectively distinguishes viewpoint changes from actual changes by leveraging a Pair Encoder and Multi-grained Representation Aggregator (MRA) to construct a reliable difference representation. A Gating Cycle Mechanism ensures semantic consistency between different representation learning and language generation, bridging the gap between visual and textual features.

2.4 Lightweight video salient object detection

CNN-based models [23, 59, 81] for VSOD tasks leverage semantic information embedded within the network. These models typically utilize pre-trained backbones from ImageNet, yet face challenges such as information leakage and redundancy. To mitigate these issues, Hu et al. [23] introduced a dual-stream network that extracts appearance and motion representations for object detection. However, this approach struggles with the network burden imposed by sparse feature matrices. Tang et al. [66] proposed a lightweight network incorporating knowledge distillation and a saliency guidance feature embedding module to enhance spatial and temporal features. Singh et al. [59] developed VS-Net, tailored for detecting salient objects using multi-scale spatiotemporal features, but faces limitations due to long dependencies in skip connections. Hu et al. [24] introduced a lightweight model integrating multiple heterogeneous decoders via 3D convolutions to enhance accuracy, yet it does not explicitly address correlation features associated with object and background deformations. Cheng et al. [9] proposed a holistic lightweight model for extracting spatiotemporal features, yet struggles with detecting objects in cluttered backgrounds and scenes with deformation. Recently, DSNet [60] was introduced to address these challenges, demonstrating efficiency in training and testing times by leveraging separability and deformability concepts. Su et al. [63] introduce the Unified Framework for Group-based Segmentation (UFGS), which utilizes a transformer block to capture long-range dependencies between image patches, enhancing patch-structured similarities. An intra-MLP learning module generates self-masks to reduce partial activation, improving segmentation accuracy. Xu et al. [85] enhance video segmentation using the Segment Anything Model (SAM), incorporating SAM-guided edge information to refine labels and mitigate interference. A SAM-driven spatiotemporal network and a global-aware loss are introduced to capture global semantic relationships and improve

salient object detection. Zhao et al. [94] present a space-time memory (STM)-based network with an encoder-decoder architecture that efficiently extracts temporal features from adjacent frames and integrates spatial-temporal fusion for object detail enhancement and saliency map reconstruction. A motion-aware loss for multitask learning improves both video salient object detection (VSOD) and object motion prediction while maintaining object integrity. Huang et al. [26] introduce a lightweight VSOD architecture utilizing a ShuffleNet-V2 backbone for efficient feature extraction, combined with a Depth-wise Multi-scale Pooling Module (DMPM) for compact multi-scale context aggregation. A Shuffle-enhanced Multi-modal Fusion Module (SMFM) progressively fuses spatial and temporal information, achieving competitive accuracy with a significantly smaller model size.

3 Proposed methodology

3.1 Motivation

Current multi-scale spatial and temporal models [43, 44, 52, 65, 86, 89] often rely on fixed kernel structures and complex architectures, resulting in high computational overhead, slower inference speeds, and increased hardware demands. Deformable convolution-based approaches [12, 60, 79, 97], though efficient in capturing geometric spatiotemporal features at low cost, struggle with generalization on unseen data at multiple scales. To address these limitations, we propose a Deformable Multi-Scale Fusion Network (DMFNet), which extracts the attention-based geometric spatial and temporal features such as various size, shape, orientation, and deformation dynamically using the proposed modules Deformation Attention Encoder Module (DAEM), Deformable Atrous Attention Module (DAAM), and Deformable Fusion Network (DFNet) that improves accuracy and robustness by capturing diverse object patterns, including occlusions, scale variations, and rapid motion. Furthermore, existing contrastive learning methods [1, 8, 41, 70, 72] fail to differentiate geometric changes across foreground and background objects at multiple scales. To overcome this problem a Geometric Multi-Scale Pixel-wise Contrastive Learning (GMPCL) is proposed, which extracts geometric features such as shape, size, orientation, and angle, enabling the efficient differentiation of foreground and background object similarities across video clips. This novel approach enhances both accuracy and robustness while preserving the temporal locality of the objects.

3.2 Overview of proposed method

To capture the efficient location information and conserve boundary localization using prior context, a novel, efficient, lightweight Deformable Multi-scale Fusion Network (DMFNet) is proposed as shown in Fig. 2. Initially, the input frame is passed to the backbone VGG-16 [19, 58] network, which extracts the spatiotemporal information in the DMFNet. The top four blocks output of the backbone with dimensions (128, 256, 512, 512) are passed to the Receptive Field Blocks (RFBs) [43], which extract the multi-scale spatiotemporal information and generate the multi-scale spatiotemporal information with dimension (64). These multi-scale spatiotemporal information are passed to the encoder blocks as well as the decoder blocks to extract the multi-scale geometric spatiotemporal features. The output of encoder blocks is passed to the DAEM, which extracts the attention-based multi-scale geometric features. Further, the output of encoder blocks, decoder blocks, and DAEM are cross-multiplying together to enhance and generalize the feature quality. After that, these features are passed to DAAM to extract discriminative attention-based multi-scale geometric features and normalize them. At last, the DFNet is used to enhance the feature vector and generate the saliency maps. To efficiently learn the multi-scale geometric semantic information, the proposed Geometric Multi-Scale Contrastive Learning (GMPCL) technique is used, which differentiates the foreground and background information of salient objects. The main objective of GMPCL is to make similar foreground pixels of the feature maps close in the same video and far away from background pixels in the same or different video at the pixel level. By finding out the positive and negative similarities between the total video clips, where the positive similarity is between foreground-to-foreground and the negative similarity between foreground-to-background or background-to-background between the same or different video clips. Due to GMPCL, the proposed DMFNet is able to learn the geometric multi-scale pixel-level attention-based contrastive features and improves the temporal information. In the subsequent subsections, we will describe a detailed explanation of the DMFNet model and its components with the GMPCL loss and the process of producing the saliency map SM_k.

3.3 DMFNet network architecture

The DMFNet is a combination of four RFB blocks (RFB_i, i = 1, 2, 3, 4), each with different dimensions 128, 256, 512 and 512, four encoder blocks (E_i, i = 1, 2, 3, 4), four decoder blocks (D_i, i = 1, 2, 3, 4), four DAEM modules (DA_i, i = 1, 2, 3, 4) details is given in Sect. 3.5, four DAAM modules detail is given in Sect. 3.6, and one DFNet modules with 64 dimension detail is given in Sect. 3.7. The encoder blocks E_i is a combination of two depthwise convolutions (DWSConv), Batch Normalization (BN), and ReLU activation layers. The decoder blocks D_i is a combination of two DSConv layers, a transposed convolution layer (TC), and

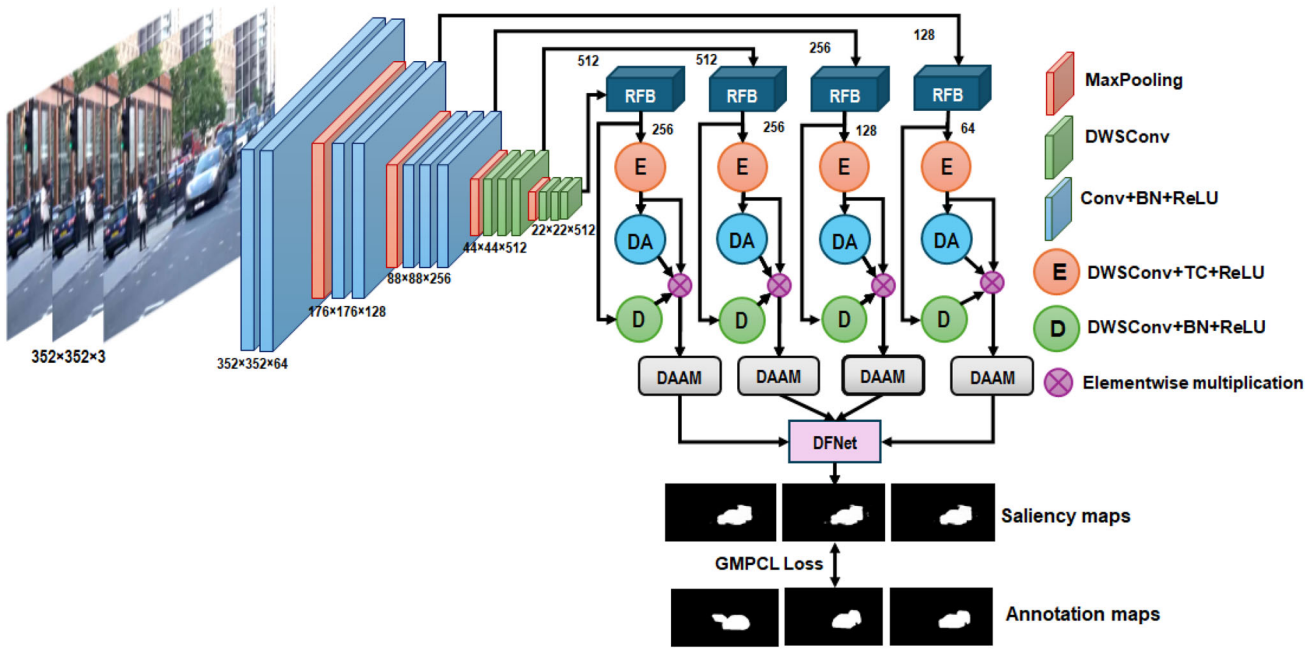


Fig. 2 Architecture of DMFNet. Where E is an encoder block, D is the decoder block, DA is deformable attention-based encoder module (DAEM), $\otimes$ is element-wise multiplication, DAAM is a deformable atrous attention module, and DFNet is a deformable fusion network

ReLU activation layers. The last four blocks of the backbone VGG-16 network [19, 58] are connected to the four RFB blocks with different dimensions (512, 512, 256, and 128), which generates the multi-scale spatiotemporal representation information in (256, 256, 128, and 64) dimensions. The RFB block output is passed to each encoder block as well as each decoder block. Each encoder block output is passed to each DAEM module to extract the attention-based multi-scale geometric spatiotemporal information. Further, the output of each encoder block, DAEM, and decoder block are multiplied to enhance the attention-based multi-scale geometric spatiotemporal information and give the generalized cross-attention-based multi-scale geometric spatiotemporal information. After that, DAAM is used to extract the discriminative cross-attention-based geometric features and normalize them. At last, these features are passed to the DFNet, which generalizes the features from high-level to low-level and generates enhanced saliency maps SM_k.

3.4 Multi-scale geometric attention feature extraction

Given a dataset comprising T video clips, each with k consecutive frames (where $k = 1, 2, \cdots, T$), the appearance frames are denoted as $(A_k)_{k=1}^{T}$, and the corresponding annotation maps are represented as $(GT_k)_{k=1}^{T}$. These frames are passed to the DMFNet, where at first, the backbone spatiotemporal features $X_k^{b_p}$'s for {p = 1, 2, 3, 4} with dimensions 128, 256, 512, and 512, respectively, are extracted using the backbone VGG-16 network. These backbone features are then forwarded to four RFB blocks with the help of different filters (1×1, 3×3, 5×5, 7×7) and dilation operations (1, 3, 5, and 7) to extract the multi-scale spatiotemporal information, $X_k^{m_q}$ for {q = 1, 2, 3, 4} with dimensions 256, 256, 128, and 64, respectively. These features are passed to the four encoder blocks E to encode the multi-scale geometric spatiotemporal features $X_k^{e_q}$. These multi-scale geometric encoder spatiotemporal features are passed to the four DAEM modules DA to extract the attention-based multi-scale geometric spatiotemporal information and generate the attention-based multi-scale geometric spatiotemporal information ($X_k^{da_q}$). The process is given in Eq. 1.

$$X_k^{da_q} = DA(X_k^{e_q}), \text{ for } q = 1, 2, 3, 4 \tag{1}$$

The multi-scale spatiotemporal features $X_k^{m_q}$'s are passed to the four decoder blocks D_q's to extract the multi-scale geometric decoder spatiotemporal features ($X_k^{d_q}$). The procedure is given in Eq. 2.

$$X_k^{d_q} = D(X_k^{m_q}), \text{ for } q = 1, 2, 3, 4 \tag{2}$$

where D's are the decoder blocks. Further, the cross mutation of feature attention is performed between encoder blocks output ($X_k^{e_q}$), decoder blocks output ($X_k^{d_q}$), and DAEM modules output ($X_k^{da_q}$) for {q = 1, 2, 3, 4}, using element-wise multiplication operation $\otimes$ and generate the generalized

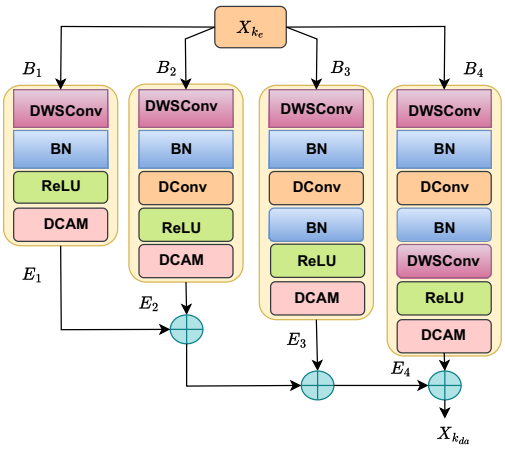


Fig. 3 Architecture of deformable attention encoder module (DAEM). Where DConv is a deformable convolution layer with 1×1 filter, $\oplus$ is element-wise addition operation, BN is batch normalization, DWSConv is the depth-wise separable convolution

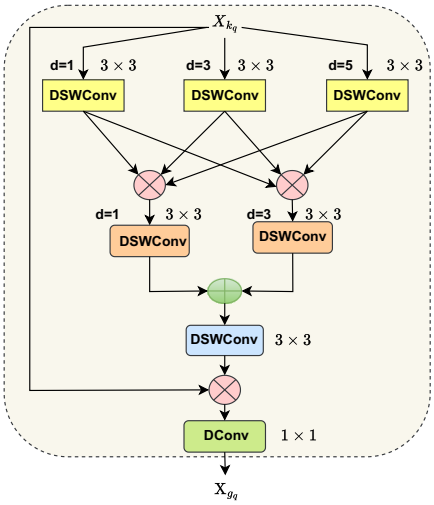


Fig. 4 Architecture of deformable atrous attention module (DAAM). Where DConv is a deformable convolution layer with 1×1 filter, $\oplus$ is element-wise addition operation, d is the dilation rate, BN is batch normalization, DWSConv is the depth-wise separable convolution and σ sigmoid operation

representation of attention-based multi-scale geometric spatiotemporal information using Eq. 3.

$$X_k^q = X_k^{d_q} \otimes X_k^{da_q} \otimes X_k^{e_q}, \text{ for } q = 1, 2, 3, 4 \tag{3}$$

Next, the cross-mutated representation of attention-based multi-scale geometric spatiotemporal features (X_k^q) is given as input to the DAAM modules, which extract the cross representation of attention-based multi-scale geometric spatiotemporal features and enhance the representation of these features.

$$X_k^{g_q} = \text{DAAM}(X_k^q), \text{ for } q = 1, 2, 3, 4 \tag{4}$$

At last, the cross representation of enhanced attention-based multi-scale geometric spatiotemporal features ($X_k^{g_q}$) is passed to the DFNet, which converts the high-level resolution spatiotemporal features to low-level attention-based multi-scale geometric features and generates the enhanced saliency maps SM_k.

$$\text{SM}_k = \text{DFNet}(X_k^{g_q}), \text{ for } q = 1, 2, 3, 4 \tag{5}$$

To provide the supervision to the proposed model (DMFNet), the GMPCL loss followed by Binary Cross Entropy (BCE) and Intersection over Union (IoU) is used, which distinguishes the foreground and background hard pixels locally from the same video clips or different clips and soft pixels globally (Fig. 3).

3.5 Deformable attention encoder module (DAEM)

The localization of salient objects is a challenging problem in VSOD. Generally handling the localization problem, recent methods [10, 59] only use multi-scale spatiotemporal features, which is insufficient to detect deformable scenes and objects. So, some contemporary methods use the motion map [31, 55] as an additional modality to detect salient objects. However, these methods still failed to detect the geometric variation of background and foreground scenes and increased the computational complexity, storage, and network parameters. So to overcome these problems, the Deformable Attention-based Encoder Module (DAEM) is proposed, which solves the localization problem with the help of deformable convolution layer (DConv), DWSConv layers, and deformable cross-attention module (DCAM) [7]. The DAEM is used to bridge the correlation gap between two adjacent frames. It is learning the weight map to give localization information from the feature map and play a more significant role in low-light scenes. DAEM has four blocks (B_1, B_2, B_3, B_4), and the B_1 block has one DWSConv layer with 3×3 filter, BN, non-linear ReLU activation function followed by DCAM module and generate the features E_1. B_2 has one DWSConv layer with 3×3 filters, BN, DConv, and ReLU followed by DCAM and give the features E_2 and fuse with E_1. The B_3 has one DWSConv with 3×3 filter, two BN, DConv with 1×1 filter, ReLU followed by DCAM and fuse with the fusion of E_1 and E_2. The B_4 has two DWSConv with 3×3 filter, two BN, DConv with 1×1 filter, ReLU followed by DCAM to extract the multi-scale geometric spatiotemporal features and fuse using element-wise addition operation ($\oplus$) with the fusion of E_1, E_2, and E_3. At last, the attention-based multi-scale geometric spatiotemporal features is generated $X_{k_{da}}$.

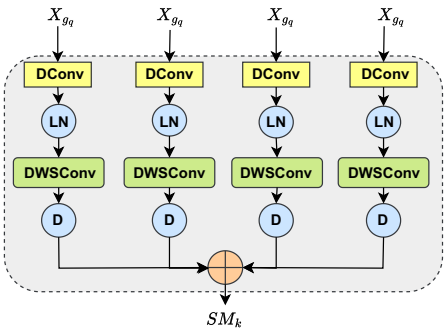


Fig. 5 Architecture of deformable fusion (DFNet). DWSConv is a depth-wise separable convolution layer, D is a dropout layer, DConv is a deformable convolution layer, LN is layer normalization, and, at last, element-wise (⊕) addition operation

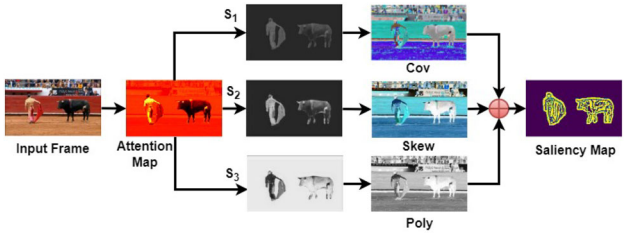


Fig. 6 Architecture of geometric multi-scale pixel-level attention contrastive learning (GMPCL). AT is the attention map, S_1, S_2, $and S_3$ are the multi-scale attention feature maps, and ⊕ is the element-wise addition operation

3.6 Deformable atrous attention module (DAAM)

A DAAM module, which autonomously fuses diverse features, is proposed to explore the correlation between cross-level and multi-scale features. Initially, the distinctions in cross-level features arise from the focus of shallow layers on spatial texture cues and deeper layers on semantic context information. It extracts the multi-scale and cross-level features using the DWSConv layer and DConv layer in a densely connected manner and strengthens the multi-scale geometric spatiotemporal cues. As shown in Fig. 4, the first three DWSConv layers with 3 × 3 filters and dilation rates (1, 3, 5, 7) are used to extract the features at different scales. Next, these layers' output is densely cross-mutated using element-wise multiplication ⊗ and passed to two DWSConv layers, which have 3 × 3 filters with dilation rates (1, 3) to extract the more discriminative cross-level multi-scale features and fuse together using element-wise addition ⊕ operation. Further, a DWSConv layer is used with 3 × 3 filters to generalize the cross-fused features and multiplied using element-wise multiplication ⊗ with the directly passed features to preserve the features inconsistency. Lastly, the DConv layer with 1 × 1 filter is used to extract the geometric multi-scale spatiotemporal representation and adapt the short- and long-term temporal dependency dynamically.

3.7 Deformable fusion network (DFNet)

The current existing methods [31, 39, 50] use the saliency head U-Net-based architecture, which converts high-level strong features to low-level weak features and generates inconsistency at the feature level. These models are unable to differentiate the foreground and background of saliency objects in geometric variation changes at the pixel-level. To overcome these problems, the Deformable Fusion Network (DFNet) is proposed, which has four blocks, and each block is a combination of a DWSConv with 3 × 3 filter, Layer normalization (LN), dropout (D) layer, DConv layer as shown in Fig. 5. The DFNet has two properties: (i) it gives depth-wise low-rank geometric features, and (ii) it preserves temporal order. The generalized representation of attention-based multi-scale geometric features X_{g_2}, X_{g_3}, X_{g_4}, X_{g_5} are given input to the four DConv layer to extract low-rank geometric patterns and captures essential properties of objects, which summarizes the object's position and orientation without increasing the network parameters and floating point operation. Next, LN is used to normalize the activation of the neurons within a layer and reduce internal covariate shifts to the parameter updates during feature extraction. Further, four DWSConv layers are used to learn the depth-wise and point-wise sparse representation of the object's foreground and background features. The four dropout layer (D) is used to drop the randomly 50% connection of the network to reduce the overburden and overfitting of the network parameters. At last, all the block features are fused together using element-wise addition ⊕ operation, and the saliency maps SM_k are generated.

3.8 Geometric multi-scale pixel-level contrastive learning

The recent existing VSOD methods [8, 31, 59, 60, 93] utilize the Binary Cross Entropy (BCE) and Intersection Over Union (IOU) loss function during training. These methods are unable to capture the geometric variation of intra-pixel semantic features and inter-pixel discrimination features at multiple scales. To overcome these challenges, the Geometric Multi-Scale Pixel-level Contrastive Learning (GMPCL) technique is proposed as shown in Fig. 6, which performs two operations. (1) Attention map augmentation is performed using the proposed Spatial Channel Attention Pooling (SCAP) Module and generates multi-scale views of attention while sharing important views of features at the pixel level. (2) The attention map view is distributed into two regions (foreground and background) using ground-truth annotation maps at pixel-level. Then the Geometric Multi-Scale Pixel-wise Contrastive learning (GMPCL) loss

is performed to find the similarity between foreground and background regions at each pixel from the same or the different video clips as shown in Fig. 7 where similar foreground attention features pull together and dissimilar foreground attention regions and background attention regions far at pixel-level representation space.

Spatial channel attention pooling (SCAP) SCAP calculates the attention maps in two ways using saliency maps (SM_k), first using 3×3 convolution layer (Conv2d), channel attention (CA), and sigmoid (σ) activation layer, which gives more attention to each channel of the feature map. Second, Conv2d, spatial attention (SA), and σ give more attention to each and every pixel-level change of the object's spatial structure, and sigmoid activation is used to parameterize the spatial and channel attention maps on the activation priority. The process is given in below Eq. 6.

$$\begin{aligned} AM_1 &= \sigma(\text{Conv2d}(\text{CA}(SM_k))) \\ AM_2 &= \sigma(\text{Conv2d}(\text{SA}(SM_k))) \end{aligned} \tag{6}$$

Further, these attention maps (AM_1, AM_2) are fused at the pixel level together to generate the total attention maps (AM_k). The process is given in below Eq. 7.

$$AM_k = \frac{1}{|C|} \times \left(\sum_{i \in C} AM_1 \right) + AM_2 \tag{7}$$

where C is the number of channels. Next, the attention maps (AM_k) are normalized using the l2-norm and denoted as $A \in \mathbb{R}^{h \times w \times 3}$ as attention maps. N_c represents the count of pixels with class $c \in \{0, 1\}$ in the annotation map G, while N^G represents the total number of pixels in G, a_p^A denotes a d-dimensional attention map vector, which is extracted from A at pixel p. Suppose $\mathbb{1}_{pq}^{GG} = \mathbb{1}\left[y_p^G = y_q^G, p \neq q\right]$ and $e^{F(a_p^A, a_q^A)} = exp(F(a_p^A/\tau, a_q^A/\tau))$, where $F(.)$ is the functions such as covariance, skewness, and polynomial kernel, y_p^G and y_q^G are annotation maps of pixel p and q in G, while r is the pixel $r \in (p, q)$ in G, and τ is a temperature hyper-parameter [8, 82]. The process of calculating GMPCL loss is shown in Eq. 8.

$$\text{GMPCL} = \sum_{k=1}^{K} -\frac{1}{N_G} \sum_{p=1}^{N^G} \frac{1}{N_{y_p^G}^G} \sum_{r=1}^{N_0+N_1} \mathbb{1}_{pg}^{GG} log(\text{GMPCL}_w) \tag{8}$$

where GMPCL_w is GMPCL weight, which is calculated using the covariance [2, 4], skewness [1], and polynomial kernels [28] at the pixel-level from the attention maps. As shown in Figs. 6 and 7 the covariance calculates the similarity between foreground and background pixels from geometric changes in attention map resolution, while skewness calculates the similarity between foreground and background pixels from geometric changes of attention map contrast, and the polynomial kernel calculates the similarity between outer and inner bound of the foreground and background pixels from attention maps at pixel-level. Positiveness is defined as foreground or foreground-to-foreground from the same video frames, while negativeness is defined as foreground-to-background or background-to-background from the same or different video frames. The similarity is found using the exponential operation on covariance, skewness, and polynomial kernel. The process is given in Eq. 9.

$$\begin{aligned} \text{GMPCL}_w = &\frac{e^{\text{Cov}(a_p^A, a_q^A)}}{\sum_{r=1}^{N_0+N_1} e^{\text{Cov}(a_p^A, a_r^A)}} + \frac{e^{\text{Skew}(a_p^A, a_q^A)}}{\sum_{r=1}^{N_0+N_1} e^{\text{Skew}(a_p^A, a_r^A)}} \\ &+ \frac{e^{\text{Polyk}(a_p^A, a_q^A)}}{\sum_{r=1}^{N_0+N_1} e^{\text{Polyk}(a_p^A, a_r^A)}} \end{aligned} \tag{9}$$

where + is the element-wise addition operation. $\text{Cov}(a_p^A, a_q^A)$ is the covariance, $\text{Skew}(a_p^A, a_q^A)$ is the skewness, and $\text{Polyk}(a_p^A, a_q^A)$ is the polynomial kernel between two pixels p and q of attention map.

Covariance The covariance is used to calculate the similarity of the attention map inside a region of contrast at pixel-level using below Eq.10.

$$\text{Cov}(a_p^A, a_q^A) = \frac{1}{N-1} \sum_{i=1}^{N} (a_{p_i}^A - \bar{a_p^A})^T (a_{q_i}^A - \bar{a_q^A}) \tag{10}$$

where N is the total number of attention maps.

Skewness The skewness is used to balance the distribution mismatch at high order and efficiently discriminate the positive and negative foreground and background pixels at pixel-level at low variance using Eq.11.

$$\text{Skew}(a_p^A, a_q^A) = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(a_{p_i}^A - \bar{a_p^A})^T (a_{q_i}^A - \bar{a_q^A})}{S^3} \tag{11}$$

Where S^3 is the standard deviation, $\bar{a_p^A}$ is the mean of the pixels.

Polynomial Kernel The previous polynomial kernel-based contrastive learning method [92] calculates positiveness and negativeness from the feature map at the image-level. Due to that, it contains incomplete or false positive and negative pixels of salient objects. To overcome this problem, the polynomial kernel, which has two Gaussian kernels, is used to find out the mean discrepancy between foreground and background pixels. The first kernel observes similar foreground pixel contrast, pixel positions, and pixel intensities, while

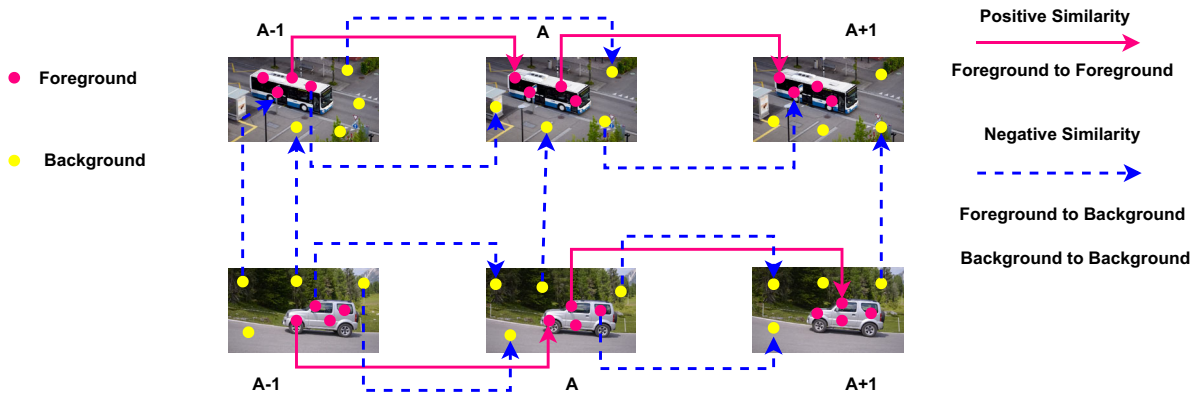


Fig. 7 Illustration of the positive and negative similarity pixel-wise from same video frames and different frames. The pink dot is the foreground, and the yellow dot is the background. The foreground-to-foreground similarity is positive, while foreground-to-background and background-to-background similarity is negative from the same video frames and different video frames

the second kernel is used to control the scale of the Gaussian function to isolate foreground and background pixels and enhance the smoothness of attention regions. The process is given in Eq. 12.

$$\begin{aligned} \text{Polyk}(a_p^A, a_q^A) &= \frac{1}{N-1}\sum_{i=1}^{N}(a_{p_i}^A - a_{q_i}^A)^2 \\ &= (a_{p_i}^A)^2 + (a_{q_i}^A)^2 - 2a_{p_i}^A a_{q_i}^A \end{aligned} \tag{12}$$

The polynomial kernel (Polyk) increases the contrast between foreground and background pixels. The Polyk generates the contrast 0 when $a_{p_i}^A = a_{q_i}^A$ and $a_{p_i}^A > a_{q_i}^A$ it generates the minimum gap between $a_{p_i}^A$ and $a_{q_i}^A$ pixels contrast.

Total loss Further, to train the proposed model, the above proposed GMPCL loss is combined with Binary Cross Entropy (BCE) and Intersection Over Union (IoU) loss, which efficiently and effectively guides the network to learn the more geometric variation changes of objects. The process is given in Eq. 13.

$$\begin{aligned} T_{loss} &= \text{BCE}(\text{SM}_k, \text{GT}_k) + \text{IoU}(\text{SM}_k, \text{GT}_k) \\ &\quad + \text{GMPCL}(\text{SM}_k, \text{GT}_k) \end{aligned} \tag{13}$$

Where T_{loss} is the total loss, SM_k is saliency maps, and GT_k is annotation maps.

4 Experiments and result analysis

This section comprehensively details the experimental setup, encompassing datasets, training and testing performance, computational complexity measures, a comparative analysis against benchmarks, and an ablation study.

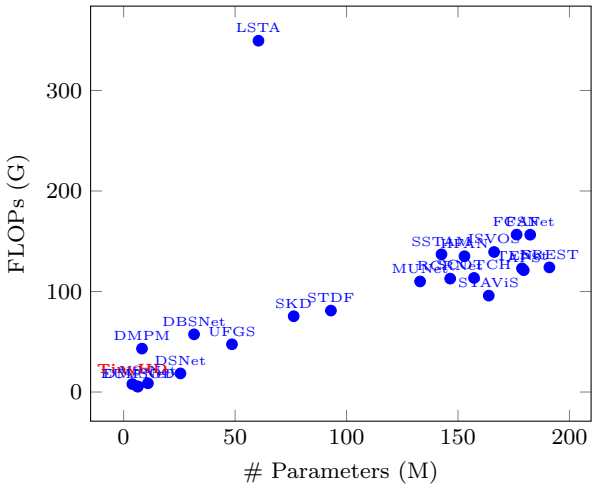


Fig. 8 Performance evaluation of our proposed DMFNet model and SOTA models in terms of # parameters and # FLOPs

4.1 Experimental setup

The experiment is conducted on a 64-bit Ubuntu 18.04 system, which has 32 GB RAM, 1 TB hard disk, and a 200 MB Solid State Drive (SSD). The GPU P5000/PCIe/SSE2 configuration is 16 GB with 490 NVIDIA Driver. Anaconda 3.8 and PyTorch [49] version 1.12.0 with CUDA 11.2 is installed on the GPU machine. All the input frame sizes are 352×352. To optimize the total loss (GMPCL + BCE + IOU) function, a weighted Adam optimizer is used in the proposed model (DMFNet) with the learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, and multi-scale training is performed on weight (1, 0.75, 0.50, 0.25).

Table 1 Comparative analysis of the proposed DMFNet model, seventeen SOTA VSOD models, and eight lightweight models across six datasets. The top three results are visually highlighted in bold, italic, and bolditalic

Model Yr. Ref.	# Param	FLOPs	Speed	DAVIS			FBMS			DAVSOD			SegTrack-V2			MCL			DAVSOD-Diff		
	(M)	(G)	(FPS)	S_α	F_β	MAE	S_α	F_β	MAE	S_α	F_β	MAE	S_α	F_β	MAE	S_α	F_β	MAE	S_α	F_β	MAE
L-VSOD Models																					
$EUVSOD_{22}$ [23]	***6.4***	**5.4**	32.5	0.920	0.894	*0.015*	0.765	0.754	0.067	0.774	0.762	0.076	0.790	0.764	0.062	0.734	0.712	0.077	0.339	0.312	**0.085**
$DBSNet_{22}$ [17]	31.6	57.4	64.0	0.867	0.846	0.026	0.888	0.873	0.030	0.760	0.690	0.062	0.751	0.678	0.049	0.859	0.837	0.032	0.429	0.338	0.152
$UFGS_{23}$ [63]	48.6	47.5	52.6	***0.921***	0.907	**0.013**	0.899	0.893	***0.026***	0.764	0.692	0.075	***0.901***	0.867	**0.012**	0.853	0.833	0.047	0.445	0.421	0.148
$VS\text{-}Net_{23}$ [59]	***10.94***	***8.7***	66.0	0.900	0.883	0.019	0.774	0.731	0.088	0.709	0.665	0.102	0.734	0.703	0.035	0.720	0.688	0.045	0.495	0.438	0.135
$TinyHD_{23}$ [24]	**3.9**	***7.9***	16.0	0.866	0.843	0.049	0.853	0.829	0.064	0.751	0.725	0.088	0.841	0.812	0.053	0.697	0.667	0.100	0.447	0.428	0.144
$DSNet_{23}$ [60]	25.5	18.5	***80.0***	*0.931*	*0.927*	***0.016***	0.898	0.887	*0.025*	0.799	0.762	***0.056***	0.897	***0.878***	***0.015***	0.860	0.835	0.023	0.519	***0.498***	0.098
$DMPM_{24}$ [26]	8.24	43.2	49.2	0.889	0.845	0.030	0.878	0.860	0.049	0.785	0.749	0.063	0.844	0.780	0.037	0.836	0.816	0.045	0.436	0.420	0.149
$DSFNet_{24}$ [62]	23.4	20.0	*84.0*	**0.940**	**0.930**	*0.015*	***0.901***	*0.896*	**0.024**	***0.816***	***0.805***	**0.054**	**0.912**	**0.902**	*0.014*	**0.867**	**0.843**	***0.021***	***0.543***	*0.508*	***0.097***
Large VSOD Models																					
$TENet_{20}$ [56]	178.7	123.0	34.0	0.916	0.904	0.019	**0.915**	**0.897**	***0.026***	0.780	0.664	0.074	0.837	0.821	0.038	0.844	0.830	0.042	0.478	0.459	0.147
$STAViS_{20}$ [69]	163.8	95.9	33.0	0.884	0.834	0.041	0.834	0.812	0.087	0.758	0.745	0.087	0.859	0.849	0.085	0.834	0.829	0.079	0.410	0.399	0.157
$EREST_{21}$ [6]	191.0	124.0	70.0	0.892	0.865	0.023	0.872	0.856	0.038	0.746	0.651	0.086	0.891	0.860	0.017	0.763	0.769	0.056	0.403	0.363	0.163
$FSNet_{21}$ [31]	182.4	156.5	13.0	0.920	0.907	0.020	0.890	0.888	0.041	0.773	0.685	0.072	0.833	0.698	0.038	***0.864***	0.821	0.023	**0.662**	0.487	0.099
APS_{21} [93]	179.5	121.2	11.0	0.894	0.884	0.039	0.803	0.796	0.059	0.715	0.695	0.092	0.815	0.725	0.056	0.675	0.597	0.094	0.461	0.415	0.102
$RCRNet_{22}$ [53]	146.5	112.8	26.0	0.914	0.900	0.019	0.899	0.892	0.027	0.768	0.694	0.059	0.844	0.778	0.027	0.860	0.822	**0.018**	0.449	0.417	0.136
$HFAN_{22}$ [50]	152.9	135.1	37.0	0.900	0.895	0.023	0.878	0.863	0.031	0.745	0.737	0.067	0.834	0.818	0.039	0.846	0.821	0.039	0.448	0.410	0.129
SKD_{22} [66]	76.3	75.4	36.0	0.893	0.883	0.022	0.850	0.831	0.055	0.624	0.612	0.084	0.860	0.847	0.025	0.726	0.711	0.079	0.348	0.323	0.109
$SSTAM_{23}$ [40]	142.6	137.0	28.0	0.913	***0.908***	0.031	0.897	0.884	0.030	0.778	0.723	0.085	0.784	0.762	0.066	0.850	***0.838***	0.048	0.446	0.419	0.130
$ISVOS_{23}$ [75]	166.2	139.3	23.6	0.902	0.895	0.049	0.832	0.814	0.062	0.738	0.707	0.093	0.835	0.767	0.059	0.697	0.657	0.098	0.459	0.437	0.146
$FCAF_{23}$ [55]	176.3	156.6	39.0	0.897	0.889	0.035	0.876	0.845	0.056	0.754	0.737	0.103	0.881	0.867	0.039	0.847	0.828	0.051	0.489	0.466	0.140
$SCOTCH_{23}$ [41]	157.2	113.5	34.0	0.889	0.859	0.039	0.897	0.888	0.054	0.776	0.763	0.073	0.847	0.831	0.040	0.841	0.829	0.053	0.460	0.442	0.153
$MUNet_{23}$ [64]	133.0	110.0	40.0	0.917	0.901	0.017	0.894	0.865	0.040	*0.839*	*0.818*	0.065	0.889	0.840	0.019	0.732	0.725	0.089	0.409	0.387	0.146
$STDF_{24}$ [14]	93.0	81.0	68.0	0.915	0.900	0.021	0.810	0.794	0.074	0.650	0.600	0.123	0.787	0.711	0.033	0.732	0.695	0.048	0.501	0.440	0.142
$LSTA_{24}$ [39]	60.5	349.5	34.0	0.884	0.867	0.026	0.773	0.762	0.097	0.623	0.485	0.135	0.719	0.601	0.054	0.654	0.596	0.068	0.580	0.413	0.162
STM_{24} [94]	100.8	83.5	100.0	0.897	0.877	0.020	0.894	0.883	0.032	0.777	0.708	0.065	0.886	0.850	0.014	0.790	0.769	0.064	*0.622*	0.418	*0.089*
SAM_{24} [85]	120.0	101.6	30.8	0.873	0.883	0.028	0.872	0.842	0.036	0.770	0.682	0.071	0.884	0.841	0.014	0.799	0.778	0.061	0.520	0.469	0.109
DMFNet	*6.2*	*5.6*	**90.0**	0.910	0.890	0.024	*0.907*	***0.895***	**0.024**	**0.840**	**0.821**	*0.055*	**0.905**	*0.887*	**0.012**	*0.865*	*0.842*	*0.020*	0.528	**0.515**	***0.097***

4.2 Datasets and performance metrics

The experiment of the proposed DMFNet model is conducted on six well-established VSOD benchmark datasets: (1) DAVIS-16 [6][1]: Renowned for comprises 50 high-quality video sequences, where 30 videos allocate for training and 20 for testing. (2) MCL [31][2] Consists of nine testing video clips. (3) FBMS [56] [3] encompasses 59 video clips with 29 videos designated for training and 30 for testing. (4) SegTrack-V2 [37] [4] comprises 13 testing video clips. (5) $DAVSOD_{19}$ [15] [5] consist of 142 video sequences, which includes 61 video clips for training and 81 for testing. (6) DAVSOD-Difficult-20 [15] contains 20 testing video clips. The evaluation performance of the proposed DMFNet model is measured in structure-measure (S_α) [56], Mean Absolute Error (MAE) [75], and F-measure (F_β) [31].

4.3 Training performance

The proposed DMFNet model has two training steps. (1) Pre-training is performed using the DUTS [31] images dataset to overcome overfitting of the model and the loss function (BCE + IOU) is used. (2) The fine-tuning operation is performed on 5189 frame samples, which is the combination of DAVIS with 2373 frames (30 videos), FBMS 600 frames (29 videos), and remaining DAVSOD with 2216 frames (35 videos) datasets. DMFNet extracts the multi-scale geometric attention-based spatial features from each and every frame, while the temporal features are extracted from adjacent frames based on the batch sizes. Fine-tuning is performed on training datasets and optimizes the total loss (BCE + IOU + GMPCL) loss using the Weighted Adam Optimizer. For controlling the vanishing gradient problem, l1 and l2 regularization [68] functions are used during the training of the model. The proposed DMFNet model takes approximately 8 h to perform the fine-tuning on 35 epochs with 8 batch sizes on a single NVIDIA GPU.

4.4 Testing performance

The testing performance of the proposed DMFNet model is evaluated on a test dataset of $DAVIS_{16}$ with 20 videos, FBMS with 30 videos, DAVSOD (Easy, Difficult) with 81 and 20 videos, MCL with nine videos, and SegTrack-V2 with 13 videos. The data augmentation techniques (motion blur, l1 and l2 regularization, rotation, and geometric transformation) are used to preprocess the feature quality. The testing

Table 2 Comparison of the different types of lightweight backbone network with DMFNet

Backbone	# Param (M)	FLOPs (G)	Speed (FPS)	Memory footprint (MB)	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
					S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
LightViT [27]	6.8	5.73	90.2	23	0.890	0.034	0.829	0.032	0.878	0.028	0.869	0.035	0.820	**0.054**	0.439	0.110
MobileNet2.0 [57]	**5.5**	**4.7**	80.9	30	0.901	0.029	0.826	0.040	0.870	0.031	0.824	0.049	0.772	0.057	0.453	0.107
EfficientNet [23]	14.2	13.4	68.2	36	0.885	0.042	0.860	0.030	0.865	0.035	0.882	0.029	0.780	0.045	0.460	0.104
PeleeNet [77]	7.9	6.8	49.7	45	0.872	0.039	0.859	0.033	0.831	0.034	0.863	0.049	0.765	0.057	0.449	0.111
SqueezeNet [30]	7.5	6.1	60.4	56	0.890	0.039	0.856	0.036	0.842	0.038	0.860	0.039	0.762	0.060	0.433	0.115
ResNet-50 [21]	29.2	27.6	40.2	59	0.911	0.027	**0.865**	0.028	0.857	0.048	0.868	0.047	0.769	0.066	0.466	0.100
ResNeXt-50 [83]	32.6	28.8	44.5	48	0.895	0.030	0.846	0.034	0.832	0.055	0.868	0.038	0.776	0.059	0.429	0.117
VGG-16 [58]	6.2	5.6	**92.0**	**20**	**0.910**	0.024	0.865	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

[1] https://davischallenge.org/.

[2] http://mcl.usc.edu/mcl-jcv-dataset/.

[3] https://lmb.informatik.uni-freiburg.de/resources/datasets/.

[4] https://web.engr.oregonstate.edu/~lif/SegTrack2/.

[5] https://github.com/DengPingFan/DAVSOD/.

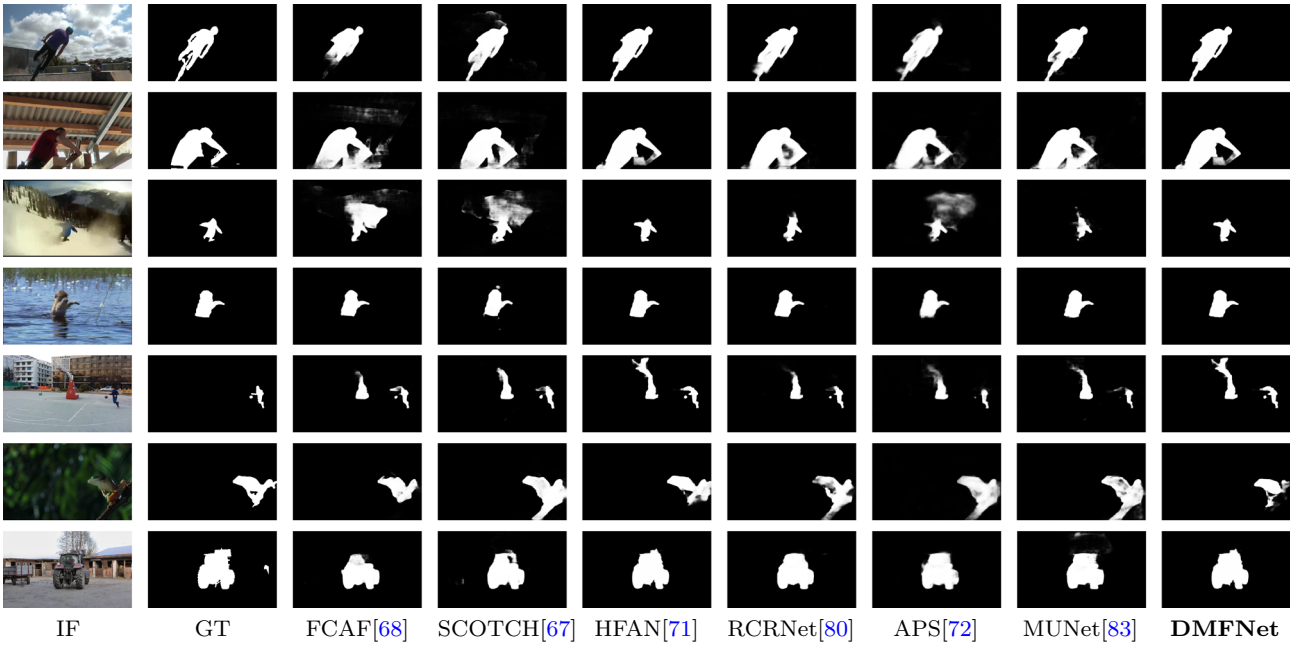


Fig. 9 The visual performance comparison of the proposed DMFNet model and SOTA models on the most difficult scenario of the DAVSOD-Difficult dataset

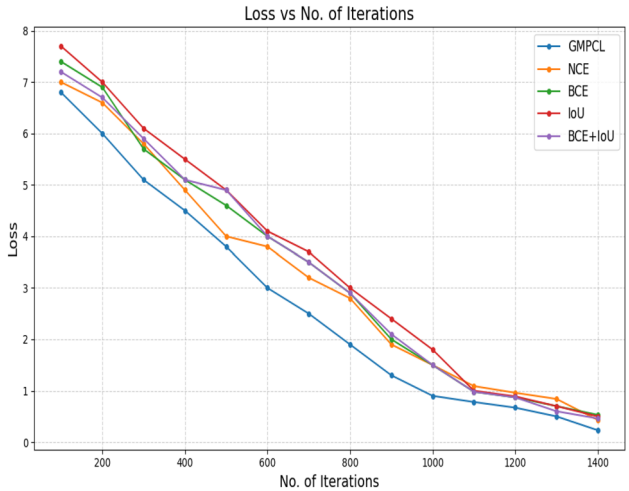


Fig. 10 Performance comparison between loss (MAE) and # of iteration during training of our proposed DMFNet model

performance of the proposed DMFNet model is generated on S_α, F_β, and MAE. Next, the complexity and computational performance of the proposed DMFNet model are measured in the # network parameters, # FLOPs, and the speed and the results are illustrated in Table 1.

4.5 Comparative analysis

The test performance of the proposed model (DMFNet) is compared with fifteen SOTA VSOD models and seven lightweight models in terms of S_α, F_β, and MAE. The Table 1 shows that the proposed (DMFNet) model demonstrates superior performance on challenging datasets such as DAVSOD-Difficult, its performance on other datasets shows some variability, especially on the DAVIS dataset. The reason behind this underperformance could be the extremely fine-grained annotations, and higher diversity in motion patterns in the DAVIS dataset [16]. Note that the issue on DAVIS-16 doesn't necessarily challenge the generalization ability of the proposed model. Instead, it reflects the dataset's unique characteristics. These concerns will be addressed in future work. From Table 1, we observe that the proposed DMFNet model demonstrates superior performance compared to fifteen SOTA models and seven lightweight VSOD models, almost on DAVSOD, MCL, SegTrack-V2, and DAVSOD-Difficult datasets, as measured by metrics such as S_α, F_β, and MAE. Additionally, the DMFNet model outperforms fifteen SOTA models in terms of complexity and computational efficiency and is second best to seven lightweight VSOD models in # parameters, Flops, and first in speed. The saliency maps are efficiently generated by the DMFNet model in less time in comparison to the SOTA models. The Table 1 shows that the DMFNet model takes fewer parameters, FLOPs, and significantly faster speed in comparison to SOTA models to get better testing accuracy. So, the DMFNet model performs well, and it is able to balance the performance and # parameters, FLOPs, and speed. The comparison of # parameters and FLOPs are given in given in the Fig. 8. Additionally, the DMFNet is compared with different backbone networks, and the results are illustrated in Table 2.

Table 3 Ablation studies for the components setting of DMFNet. Where * denotes the single scale results of all modules

No	Component setting				DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	RFB	DAEM	DAAM	DFNet	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
1	Baseline				0.823	0.049	0.668	0.070	0.790	0.041	0.768	0.067	0.733	0.066	0.354	0.153
2	✓				0.857	0.038	0.798	0.053	0.785	0.042	0.781	0.060	0.766	0.060	0.390	0.140
3		✓			0.869	0.048	0.768	0.059	0.783	0.047	0.780	0.063	0.763	0.058	0.388	0.129
4			✓		0.866	0.045	0.804	0.050	0.799	0.040	0.783	0.056	0.770	0.056	0.397	0.130
5				✓	0.878	0.031	0.829	0.039	0.805	0.038	0.793	0.049	0.757	0.054	0.373	0.120
6	✓		✓		0.865	0.044	0.799	0.049	0.823	0.030	0.758	0.075	0.786	0.049	0.368	0.133
7		✓		✓	0.895	0.038	0.818	0.029	0.863	0.035	0.816	0.032	0.766	0.053	0.435	0.114
8			✓	✓	0.873	0.032	0.832	0.024	0.861	0.037	0.807	0.035	0.763	0.055	0.432	0.116
9	✓	✓	✓		0.908	0.030	0.830	0.028	0.856	0.040	0.848	0.022	0.752	0.058	0.478	0.108
10		✓	✓	✓	0.905	0.026	0.837	0.026	0.860	0.038	0.855	0.020	0.760	0.056	0.499	0.102
11	✓		✓	✓	0.902	0.029	0.845	0.025	0.869	0.033	0.868	0.018	0.770	0.054	0.519	0.098
12	✓	✓		✓	0.900	0.038	0.858	0.022	0.863	0.035	0.896	0.016	**0.846**	**0.045**	0.515	0.100
13	✓	✓	✓	✓	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	0.840	0.055	**0.528**	**0.097**
14*	✓	✓	✓	✓	0.899	0.043	0.854	0.024	0.850	0.045	0.890	0.014	0.834	0.049	0.500	0.105

Table 4 Ablation studies for the design choice of DMFNet. Where E_i & D_i are the encoder and decoder i=1, 2, 3, 4

Module	Parameters			DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	# Params(M)	FLOPs(G)	Speed(FPS)	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
VGG-16	28.3	26.7	36	0.876	0.037	0.699	0.070	0.773	0.060	0.789	0.064	0.776	0.066	0.385	0.134
E_i & D_i	25.7	24.0	50	0.885	0.034	0.789	0.066	0.806	0.059	0.793	0.059	0.782	0.063	0.380	0.137
DAEM	20.1	18.9	65	0.892	0.031	0.826	0.032	0.827	0.054	0.799	0.051	0.789	0.059	0.389	0.130
DFNet	18.9	16.8	70	0.896	0.029	0.792	0.060	0.837	0.051	0.810	0.050	0.796	0.057	0.395	0.125
DAAM	14.3	13.5	76	0.902	0.027	0.819	0.053	0.849	0.045	0.826	0.042	0.832	**0.052**	0.499	0.102
All	**6.2**	**5.6**	**90**	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

4.5.1 Qualitative comparison

The qualitative comparison of the DMFNet is compared with six SOTA models, which are given in Fig. 9 as a form of saliency map in various difficult scenarios. The DMFNet model can distinguish the salient object from the background with clear borders in various difficult scenarios, such as illumination with motion blur (2nd, 6th rows), partial occlusion with motion blur (5th and 7th), illumination with low-light vision (1st, 2nd, 4th, and 6th rows), clutter background with noise (3rd and 8th rows), and illumination with deformable object/background (1st, 2nd, 3rd, 4th, and 5th rows). From Fig. 9, we observed that our DMFNet model is able to detect almost all the difficult scenarios on the DAVSOD-Difficult dataset, which shows their capability to use in real-time VSOD applications on mobile and edge devices.

4.6 Failure cases and future works

The proposed DMFNet model is compared with four SOTA models in terms of the failure cases as shown in Fig. 11, which illustrates that it faces problems in detecting the shadow with low light contrast in row 1, and deforming objects with different lighting conditions in rows 3 and 4. Row 2 shows a large scale with deformation, which causes problems in distinguishing the border between an object and its background and is unable to detect the proposed as well as SOTA models. To overcome these problems, knowledge distillation and multi-domain contrastive learning approaches will be considered in the future.

4.7 Ablation study

The ablation study of the proposed DMFNet model is illustrated in Table 3, which shows the effectiveness of each component setting of the DMFNet model. It is designed parallelly, which is densely connected to VGG-16 Backbone Network, four Encoder ($E_i, i = 1, 2, 3, 4$) blocks, and Decoder ($D_i, i = 1, 2, 3, 4$) blocks with DAEM, DAAM, and DFNet. The parallel connection of the DMFNet modules is more effective and can outperform when the visual connection processing is used hierarchically. So, we use the attention-based multi-scale geometric spatiotemporal features extraction mechanism in our DMFNet model. The results of the DMFNet model show that using the attention-based multi-scale geometric spatiotemporal features outperformed visual perception learning hierarchically. The Table 3 shows that as the component of the model is added into the framework, the performance is gradually increased. In addition, the comparison between No. 1 to No. 13 shows the inferiority of the proposed solution, which is compared with the baseline. The Table 4 shows that as the components are added to the proposed model, # parameters

Table 5 Ablation studies for the effectiveness of proposed modules in DMFNet

Module	Parameters			DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	# Params(M)	FLOPs(G)	Speed(FPS)	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
DAEM (W/O)	6.12	5.7	77	0.879	0.045	0.783	0.067	0.819	0.058	0.800	0.056	0.790	0.059	0.390	0.128
DAAM (W/O)	6.10	5.5	79	0.873	0.049	0.800	0.058	0.830	0.049	0.814	0.052	0.769	0.062	0.396	0.123
DFNet (W/O)	6.17	5.8	84	0.890	0.044	0.818	0.054	0.828	0.052	0.820	0.050	0.750	0.066	0.409	0.119
DAEM (W)	6.14	5.6	87	0.887	0.033	0.820	0.062	0.827	0.055	0.822	0.047	0.826	0.057	0.489	0.104
DAAM (W)	**5.90**	**5.3**	82	0.892	0.036	0.836	0.058	0.806	0.059	0.813	0.049	0.831	0.049	0.499	0.100
DFNet (W)	6.11	5.9	85	0.899	0.030	0.845	0.033	0.828	0.054	0.799	0.051	0.820	0.058	0.508	0.099
All	6.20	5.6	**90**	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

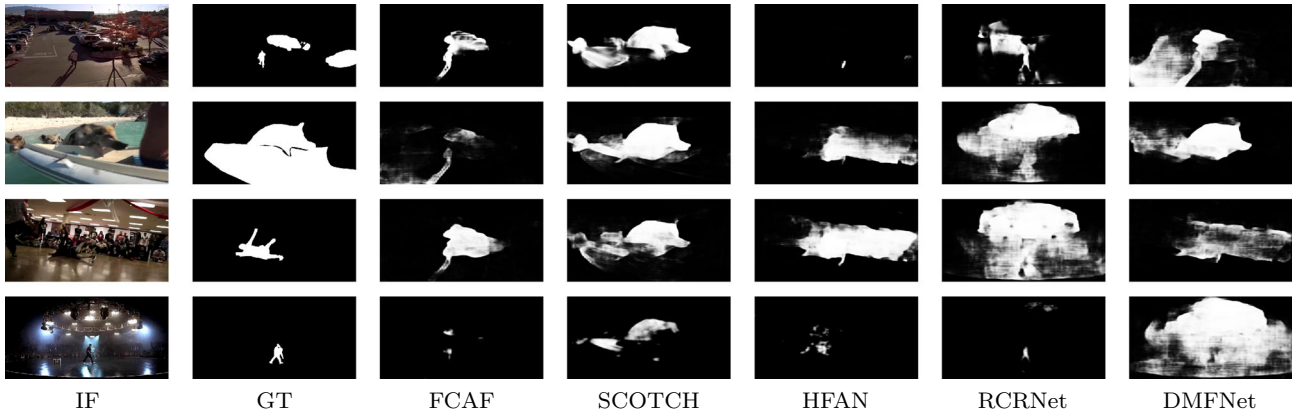


Fig. 11 The failure case of DMFNet and SOTA models on the DAVSOD-Difficult dataset. IF is the input frame, and GT is the annotation map, SPGO [54], TMO [10], CFAM [8], HCPN [51], and DMFNet

Table 6 Effectiveness of attention module with different variations of GMPCL. Where CL is contrastive learning, AM is attention module, PCL is pixel-level contrastive learning, MPCL is multi-scale pixel-level contrastive learning, and GMPCL is geometric multi-scale pixel-level contrastive learning

Different combination	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
CL + AM	0.894	0.033	0.821	0.062	0.869	0.052	0.797	0.051	0.789	0.056	0.389	0.130
PCL + AM	0.897	0.030	0.844	0.056	0.877	0.040	0.821	0.040	0.790	0.052	0.398	0.122
MPCL + AM	0.905	0.026	0.859	0.036	0.889	0.035	0.863	0.028	0.832	**0.049**	0.497	0.102
GMPCL	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

The bold signifies the best result value

Table 7 Comparison between proposed GMPCL and different contrastive learning methods

Contrastive learning	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
SimpCL [8]	0.880	0.047	0.810	0.067	0.846	0.063	0.779	0.075	0.730	0.079	0.380	0.135
FineCo [81]	0.871	0.056	0.819	0.062	0.838	0.067	0.763	0.077	0.737	0.074	0.406	0.124
BYOL [20]	0.886	0.043	0.826	0.057	0.827	0.069	0.772	0.069	0.746	0.069	0.396	0.117
MIL-NCE [47]	0.891	0.039	0.839	0.049	0.851	0.058	0.797	0.048	0.764	0.062	0.445	0.110
MaskContrast [74]	0.896	0.034	0.835	0.045	0.863	0.050	0.854	0.039	0.828	0.059	0.390	0.120
PCL [82]	0.900	0.030	0.844	0.026	0.879	0.037	0.880	0.020	0.835	**0.052**	0.398	0.122
GMPCL	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	**0.024**	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

The bold signifies the best result value

and FLOPs are decreased because of the reduction in the dimension of the feature maps. Additionally, the network complexity and performance are compared with the number of the filters size, encoder modules, decoder modules, DAEM modules, and DAAM modules as shown in Table 8, which shows that as the number of filters and proposed modules increase, the performance and network complexity increase, but at filters 128 the performance is saturated and network complexity increases due to losing the spatial and temporal locality at higher scales. Further, in Table 6, we compared the results of the proposed GMPCL learning how the GMPCL concept is framed. First, the proposed attention module (AM) with simple Contrastive Learning (CL+AM) together is used to perform the fine-tuning, which is not so effective. Second, AM with Pixel-wise Contrastive Learning (PCL+AM) together is used to finetune the network and give better results than (CL+AM) due to the efficiently differentiated attention map pixel-wise. Third, the PCL is finetuned at multiple scales (MPCL+AM) with AM giving better results in comparison to (PCL+AM) and (CL+AM). Fourth, when (MPCL+Geometric) together is called Geometric Multi-Scale Pixel-level Contrastive Learning, which

Table 8 Ablation studies for configuration of modules in proposed (DMFNet) model

Configuration of modules									DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
# Filters	# RFB	# E	# D	# DAEM	# DAAM	# Params(M)	FLOPs(G)	Speed(FPS)	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
8	1	1	1	1	1	4.10	3.6	100	0.870	0.048	0.780	0.068	0.810	0.065	0.780	0.060	0.776	0.070	0.381	0.130
16	2	2	2	2	2	4.60	3.8	99	0.872	0.046	0.776	0.070	0.816	0.060	0.789	0.057	0.765	0.074	0.392	0.126
32	3	3	3	3	3	5.00	4.3	95	0.880	0.043	0.819	0.058	0.822	0.055	0.811	0.053	0.776	0.068	0.405	0.117
64	4	4	4	4	4	6.20	5.6	**90**	0.910	0.024	**0.865**	**0.020**	**0.907**	**0.024**	0.905	0.012	**0.840**	**0.055**	**0.528**	**0.097**
128	5	5	5	5	5	6.50	6.3	85	**0.919**	**0.019**	0.857	0.022	0.898	0.027	**0.909**	**0.010**	0.838	0.057	0.519	0.099

The bold signifies the best result value

performs better than other combinations efficiently because it is able to differentiate the foreground and background of objects at multiple scales due to dividing the objects into two regions (foreground and background) and calculate the similarity between each region with help of covariance, skewness, and non-linear feature interactions as shown in Table 6. The covariance and skewness help to capture both the global distribution and local structure of the attention features at various scales, allowing for a more nuanced separation of foreground and background regions in the pixel-level representation space. Using the polynomial kernel, the GMPCL loss enhances the model's ability to discriminate similar foreground features and dissimilar foreground-background pairs across different video clips. This approach goes beyond the traditional contrastive losses by using multi-scale geometric insights, which allows for improved feature alignment and stronger generalization, particularly in challenging video scenarios.

4.7.1 Effectiveness of the proposed modules in DMFNet

To evaluate the impact of each component in the Deformable Multi-Scale Fusion Network (DMFNet), we performed ablation studies on the Deformable Attention Encoder Module (DAEM), Deformable Atrous Attentive Module (DAAM), and Deformable Fusion Network (DFNet). Table 5 presents the quantitative comparison of the proposed DMFNet with and without these modules. The results reveal a significant performance degradation when any of these modules is omitted from the baseline framework. Specifically, the exclusion of DAEM reduces the model's ability to capture long-range dependencies and fine-grained contextual relationships, resulting in diminished spatial-temporal accuracy. Similarly, removing DAAM impairs the network's capacity to leverage multi-scale feature extraction with deformable attention, leading to suboptimal saliency prediction in complex scenarios. Finally, the absence of DFNet compromises the effective fusion of spatial and temporal features, critical for detecting salient objects under challenging conditions such as fast motion, deformation, illumination, and occlusion. In contrast, the inclusion of all these modules (DAEM, DAAM, and DFNet) synergistically boosts the performance of DMFNet while balancing the network complexity as shown in Table 8. DAEM enhances feature discrimination through deformable attention mechanisms, DAAM ensures efficient multi-scale feature fusion with minimal computational overhead, and DFNet significantly extracts and fuses the geometric multi-scale spatial and temporal features to generate efficient saliency maps. The integration of these modules collectively enables the network to effectively adapt geometric variation of objects, capture salient features at multiple scales, and robustly fuse spatial and temporal to achieve

Table 9 Comparison between proposed GMPCL Loss and different loss

Loss function	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
BCE Loss [46]	0.886	0.049	0.815	0.065	0.849	0.053	0.775	0.077	0.734	0.076	0.384	0.133
IoU Loss [88]	0.873	0.054	0.822	0.063	0.838	0.068	0.778	0.072	0.757	0.070	0.403	0.126
BCE + IoU Loss [5]	0.888	0.045	0.829	0.059	0.853	**0.022**	0.770	0.079	0.769	0.066	0.497	0.107
NCE Loss [47]	0.895	0.037	0.843	0.046	0.876	0.030	0.823	0.041	0.764	0.062	0.445	0.110
GMPCL Loss	**0.910**	**0.024**	**0.865**	**0.020**	**0.907**	0.024	**0.905**	**0.012**	**0.840**	0.055	**0.528**	**0.097**

The bold signifies the best result value

substantial gains in accuracy, efficiency, and robustness compared to the baseline network.

4.7.2 Effectiveness of GMPCL

To illustrate the effectiveness of our proposed GMPCL using Attention Module (AM), the proposed model is trained with different combinations and tested to check the performance such as the Attention Module (AM) with baseline Contrastive Learning (CL), AM with Pixel-level Contrastive Learning (PCL), AM with Multi-Scale Pixel-level contrastive learning (MPCL), and GMPCL as shown in Table 6, which demonstrate that the proposed GMPCL performed better than the SOTA models due to extract the geometric multi-scale spatial and temporal feature and differentiate the foreground and background similarity of objects in complex scenarios such as deformation, occlusion, illumination, and rapid motion efficiently while preserving the temporal locality than other combination.

The effectiveness of the GMPCL loss is compared with SOTA loss functions such as Binary Cross Entropy, Intersection over Union (IoU), BCE+IoU, and Noise Contrastive Estimation (NCE) as shown in Table 9, which shows that the proposed GMPCL loss is performed better than the BCE, IoU, BCE+IoU, and NCE loss because these loss functions are captured local and global, noise based feature, but fail to differentiate the more complex scenario such as deformation, various illumination conditions, occlusion, and clutter background at multiple scale variations. The GMPCL is rooted in its ability to enforce fine-grained pixel-level contrast between foreground and background features across video clips. It operates by utilizing both geometric priors and multi-scale feature representations to better capture spatial and temporal correlations within foreground attention features, while simultaneously ensuring that dissimilar features (i.e., foreground vs. background) are pushed apart in the feature space. This pixel-level contrastive framework allows the model to better learn discriminative, multi-scale representations that enhance the separability of foreground objects, which is particularly beneficial in complex, dynamic video scenes. This mechanism improves feature alignment and promotes more robust object detection across varying video content. As shown in Table 7, the comparison results of the SOTA and proposed GMPCL are presented, which show that the SOTA models are downgraded the performance while the proposed increase continuously (Tables 8, 9). Additionally, we compared the performance in terms of the loss (MAE) and # iteration as shown in Fig. 10, which describes that as the iteration is increasing the loss of decreasing (Fig. 11).

5 Conclusion

In this paper, a novel, efficient, fast, Deformable Multi-Scale Fusion Network (DMFNet) is proposed, which fully utilizes the attention-based multi-scale geometric spatiotemporal features to detect the salient object in videos. It is designed using the encoder and decoder concepts with a Deformable Fusion Network (DFNet), a Deformable Atrous Attention Module (DAAM), and a Deformable Attention Encoder Module (DAEM), which balance the network parameters and performance. The DAEM extracts the attention-based multi-scale geometric features. The DAAM enhances the contrast of the attention-based geometric features of the objects. At last, the DFNet generates efficient saliency maps without increasing the network parameters and preserves the multi-scale geometric spatiotemporal features. Further, the proposed Geometric Multi-Scale Pixel-level Contrastive Learning (GMPCL) enhances the geometric representation of features using the GMPCL loss function. With the help of extensive experiments, each module of the DMFNet model is validated on six benchmark datasets and shows significant improvement in the unified solution to advance the research in VSOD. In future work, we will plan to extend our research to federated settings, particularly for real-time VSOD tasks.

Author contributions The manuscript is prepared by all the authros and review by all the authors..

Funding This research did not receive any specific grant from 1001 funding agencies in the public, commercial, or not-for-profit sectors.

Data availability We hereby confirm that the dataset used in this study is publicly available and has been cited in the paper.

Declarations

Conflict of interest The authors state that they do not have any competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

1. Acharya K, Ghoshal D (2020) Contrast enhancement of images through skewness and mode based bi-histogram equalization. Int J Image Graphics Signal Process 12(5):13–27
2. Ahn WJ, Yang GY, Choi HD, Lim MT (2024) Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning. arXiv preprint arXiv:2403.06122
3. Alonso I, Sabater A, Ferstl D, Montesano L, Murillo AC (2021) Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8219–8228
4. Bardes A, Ponce J, LeCun Y (2021) Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906
5. Bi HB, Lu D, Zhu HH, Yang LN, Guan HP (2021) STA-Net: spatial-temporal attention network for video salient object detection. Appl Intell 51(6):3450–3459
6. Chen C, Wang G, Peng C, Fang Y, Zhang D, Qin H (2021) Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. IEEE Transact Image Process 30:3995–4007
7. Chen H, Du Y, Fu Y, Zhu J, Zeng H (2023) DCAM-Net: a rapid detection network for strip steel surface defects based on deformable convolution and attention mechanism. IEEE Transact Instrum Meas 72:1–12
8. Chen YW, Jin X, Shen X, Yang MH (2022) Video salient object detection via contrastive features and attention modules. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1320–1329
9. Cheng MM, Gao SH, Borji A, Tan YQ, Lin Z, Wang M (2021) A highly efficient model to study the semantics of salient object detection. IEEE Transact Pattern Anal Mach Intell 44(11):8006–8021
10. Cho S, Lee M, Lee S, Park C, Kim D, Lee S (2023) Treating motion as option to reduce motion dependency in unsupervised video object segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 5140–5149
11. Cong R, Song W, Lei J, Yue G, Zhao Y, Kwong S (2022) Psnet: Parallel symmetric network for video salient object detection. IEEE Transact Emerg Top Comput Intell, vol 5
12. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773
13. Dai Y, Xue C, Zhou L (2022) Visual saliency guided perceptual adaptive quantization based on HEVC intra-coding for planetary images. Plos one 17(2):e0263729
14. Deng J, Dong S, Chen L, Hu J, Zhuo C (2024) Stdf: Spatio-temporal deformable fusion for video quality enhancement on embedded platforms. ACM Transactions on Embedded Computing Systems
15. Fan DP, Wang W, Cheng MM, Shen J (2019) Shifting more attention to video salient object detection. In: IEEE CVPR, IEEE, Long Beach, CA, vol 32
16. Fan DP, Wang W, Cheng MM, Shen J (2019) Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, Hawaii, pp 8554–8564
17. Fan J, Su T, Zhang K, Liu Q (2022) Bidirectionally learning dense spatio-temporal feature propagation network for unsupervised video object segmentation. In: Proceedings of the 30th ACM international conference on multimedia, pp 3646–3655
18. Fu K, Gu IYH, Yang J (2017) Saliency detection by fully learning a continuous conditional random field. IEEE Transact Multimed 19(7):1531–1544
19. GongyangLi Z, Bai Z, Lin W, Ling H (2022) Lightweight salient object detection in optical remote sensing images via feature correlation. IEEE Trans Geosci Remote Sens 60:5617712
20. Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21271–21284
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
22. Heo Y, Jun Koh Y, Kim CS (2020) Interactive video object segmentation using global and local transfer modules. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, vol 16, pp 297–313
23. Hu C, Zhu L (2022) Efficient unsupervised video object segmentation network based on motion guidance. arXiv preprint arXiv:2211.05364 10
24. Hu F, Palazzo S, Salanitri FP, Bellitto G, Moradi M, Spampinato C, McGuinness K (2023) Tinyhd: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2051–2060
25. Huang K, Xu Z (2023) Lightweight video salient object detection via channel-shuffle enhanced multi-modal fusion network. Multimedia Tools and Applications pp 1–15
26. Huang K, Xu Z (2024) Lightweight video salient object detection via channel-shuffle enhanced multi-modal fusion network. Multimed Tools Appl 83(1):1025–1039
27. Huang T, Huang L, You S, Wang F, Qian C, Xu C (2022) Lightvit: Towards light-weight convolution-free vision transformers. arXiv preprint arXiv:2207.05557
28. Huang Z, Wang N (2017) Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219
29. Hussain T, Muhammad K, Del Ser J, Baik SW, de Albuquerque VHC (2019) Intelligent embedded vision for summarization of multiview videos in IIoT. IEEE Transact Ind Inf 16(4):2592–2602
30. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360 34
31. Ji GP, Fu K, Wu Z, Fan DP, Shen J, Shao L (2021) Full-duplex strategy for video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, IEEE, ICCV, China 30:4922–4933
32. Ji Y, Zhang H, Jie Z, Ma L, Wu QJ (2020) Casnet: a cross-attention siamese network for video salient object detection. IEEE Transact Neural Netw Learn Syst 32(6):2676–2690
33. Jia F, Wang X, Guan J, Li H, Qiu C, Qi S (2021) Wrgpruner: a new model pruning solution for tiny salient object detection. Image Vis Comput 109:104143
34. Khan A, Kuribayashi M, Wong K, Monn Baskaran V (2023) Hdr image watermarking using saliency detection and quantization index modulation. arXiv e-prints pp arXiv–2302
35. Kumain SC, Singh M, Awasthi LK (2024) Dbtsf-vsod: a decision-based two-stage framework for video salient object detection. Int J Multimed Inf Retr 13(4):38

36. Lee M, Cho S, Lee S, Park C, Lee S (2023) Unsupervised video object segmentation via prototype memory network. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 5924–5934
37. Li F, Kim T, Humayun A, Tsai D, Rehg JM (2013) Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision, IEEE, Portland, OR, USA 26:2192–2199
38. Li F, Wu B, Yi K, Zhao Z (2016) Wander join: Online aggregation via random walks. In: Proceedings of the 2016 international conference on management of data, ACM, pp 615–629
39. Li P, Zhang Y, Yuan L, Xiao H, Lin B, Xu X (2024) Efficient long-short temporal attention network for unsupervised video object segmentation. Pattern Recognit 146:110078
40. Lin L, Zheng Y, Chen W, Lan C, Zhao T (2023) Saliency-aware spatio-temporal artifact detection for compressed video quality assessment. arXiv preprint arXiv:2301.01069
41. Liu L, Prost J, Zhu L, Papadakis N, Liò P, Schönlieb CB, Aviles-Rivero AI (2023) 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
42. Liu N, Nan K, Zhao W, Yao X, Han J (2023) Learning complementary spatial–temporal transformer for video salient object detection. IEEE Transactions on Neural Networks and Learning Systems
43. Liu S, Huang D, et al. (2018) Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 385–400
44. Liu X, Wang L (2024) Msrmnet: multi-scale skip residual and multi-mixed features network for salient object detection. Neural Netw 173:106144
45. Liu Y, Gu YC, Zhang XY, Wang W, Cheng MM (2020) Lightweight salient object detection via hierarchical visual perception learning. IEEE Transact Cybern 51(9):4439–4449
46. Mannor S, Peleg D, Rubinstein R (2005) The cross entropy method for classification. In: Proceedings of the 22nd international conference on Machine learning, pp 561–568
47. Miech A, Alayrac JB, Smaira L, Laptev I, Sivic J, Zisserman A (2020) End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9879–9889
48. Pang Z, Nakashima Y, Otani M, Nagahara H (2024) Revisiting pixel-level contrastive pre-training on scene images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1784–1793
49. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8026–8037
50. Pei G, Yao Y, Xie GS, Shen F, Tang Z, Tang J (2022) Hierarchical feature alignment network for unsupervised video object segmentation. arXiv preprint arXiv:2207.08485 12(3-4):223–240
51. Pei G, Yao Y, Shen F, Huang D, Huang X, Shen HT (2023) Hierarchical co-attention propagation network for zero-shot video object segmentation. IEEE Transactions on Image Processing
52. Peng D, Zhou W, Pan J, Wang D (2024) Msednet: multi-scale fusion and edge-supervised network for rgb-t salient object detection. Neural Netw 171:410–422
53. Piao Y, Lu C, Zhang M, Lu H (2022) Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. Adv Neural Inf Process Syst 35:5614–5627
54. Ponimatkin G, Samet N, Xiao Y, Du Y, Marlet R, Lepetit V (2023) A simple and powerful global optimization for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 5892–5903
55. Qin Z, Lu X, Nie X, Liu D, Yin Y, Wang W (2023) Coarse-to-fine video instance segmentation with factorized conditional appearance flows. IEEE/CAA J Automatica Sin 10(5):1192–1208
56. Ren S, Han C, Yang X, Han G, He S (2020) Tenet: Triple excitation network for video salient object detection. European conference on computer vision, Springer, China 16:212–228
57. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv 2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, CVPR, Salt Lake City 31:4510–4520
58. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
59. Singh H, Verma M, Cheruku R (2022) 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
60. Singh H, Verma M, Cheruku R (2023) Dsnet: efficient lightweight model for video salient object detection for iot and wot applications. Companion Proc ACM Web Conf 2023:1286–1295
61. Singh H, Verma M, Cheruku R (2023) Novel dilated separable convolution networks for efficient video salient object detection in the wild. IEEE Transactions on Instrumentation and Measurement.
62. Singh H, Verma M, Cheruku R (2024) Dsfnet: Video salient object detection using a novel lightweight deformable separable fusion network. IEEE Transactions on Instrumentation and Measurement
63. Su Y, Deng J, Sun R, Lin G, Wu Q (2022) A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. arXiv preprint arXiv:2203.04708 14
64. Sun J, Mao Y, Dai Y, Zhong Y, Wang J (2023) Munet: motion uncertainty-aware semi-supervised video object segmentation. Pattern Recognit 138:109399
65. Tang Y, Zou W, Jin Z, Li X (2018) Multi-scale spatiotemporal conv-lstm network for video saliency detection. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval, pp 362–369
66. Tang Y, Li Y, Zou W (2020) Fast video salient object detection via spatiotemporal knowledge distillation. arXiv preprint arXiv:2010.10027
67. Tokmakov P, Alahari K, Schmid C (2017) Learning video object segmentation with visual memory. Proceedings of the IEEE International Conference on Computer Vision, IEEE, Hawaii 30:4481–4490
68. Tran VN, Liu SH, Huang CE, Aslam MS, Yang KL, Li YH, Wang JC (2024) Hapiclr: heuristic attention pixel-level contrastive loss representation learning for self-supervised pretraining. The Visual Computer pp 1–16
69. Tsiami A, Koutras P, Maragos P (2020) Stavis: Spatio-temporal audiovisual saliency network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4766–4776
70. Tu Y, Li L, Su L, Zha ZJ, Yan C, Huang Q (2023) Self-supervised cross-view representation reconstruction for change captioning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2805–2815
71. Tu Y, Li L, Su L, Zha ZJ, Huang Q (2024) Smart: syntax-calibrated multi-aspect relation transformer for change captioning. IEEE Transact Pattern Anal Mach Intell 46:4926
72. Tu Y, Li L, Su L, Zha ZJ, Yan C, Huang Q (2024) Context-aware difference distilling for multi-change captioning. arXiv preprint arXiv:2405.20810
73. Tu Y, Li L, Su L, Yan C, Huang Q (2025) Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In: European conference on computer vision, Springer, pp 311–328

74. Van Gansbeke W, Vandenhende S, Georgoulis S, Van Gool L (2021) Unsupervised semantic segmentation by contrasting object mask proposals. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10052–10062
75. Wang J, Chen D, Wu Z, Luo C, Tang C, Dai X, Zhao Y, Xie Y, Yuan L, Jiang YG (2023) Look before you match: Instance understanding matters in video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2268–2278
76. Wang J, Huang Z, Huang Z, Zhang M, Ren X (2024) Dsfnet: dynamic selection-fusion networks for video salient object detection. Multimed Tools Appl 83(17):53139–53164
77. Wang RJ, Li X, Ling CX (2018) Pelee: A real-time object detection system on mobile devices. Advances in neural information processing systems 31
78. Wang W, Zhou T, Yu F, Dai J, Konukoglu E, Van Gool L (2021) Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7303–7313
79. Wang W, Dai J, Chen Z, Huang Z, Li Z, Zhu X, Hu X, Lu T, Lu L, Li H, et al. (2023) Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14408–14419
80. Wang X, Zhang R, Shen C, Kong T, Li L (2021) Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3024–3033
81. Wang Z, Zhong Y, Miao Y, Ma L, Specia L (2022) Contrastive video-language learning with fine-grained frame sampling. arXiv preprint arXiv:2210.05039 2
82. Wu J, Hao F, Liang W, Xu J (2024) Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection. IEEE Transact Multimed 26:1011–1026. https://doi.org/10.1109/TMM.2023.3275308
83. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500
84. Xu B, Liang H, Ni W, Gong W, Liang R, Chen P (2022) Learning video salient object detection progressively from unlabeled videos. arXiv preprint arXiv:2204.02008
85. Xu B, Jiang Q, Zhao X, Lu C, Liang H, Liang R (2024) Multi-dimensional exploration of segment anything model for weakly supervised video salient object detection. IEEE Transactions on circuits and systems for video technology
86. Xu M, Fu P, Liu B, Li J (2021) Multi-stream attention-aware graph convolution network for video salient object detection. IEEE Transact Image Process 30:4183–4197
87. Xu Y, Song D, Hoogs A (2014) An efficient online hierarchical supervoxel segmentation algorithm for time-critical applications. In: BMVC, Citeseer, San Francisco, CA, vol 23
88. Yu J, Jiang Y, Wang Z, Cao Z, Huang T (2016) Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on multimedia, pp 516–520
89. Yue S, Tu Y, Li L, Gao S, Yu Z (2024) Multi-grained representation aggregating transformer with gating cycle for change captioning. ACM Transact Multimed Comput, Commun Appl 20:1
90. Zhang J, Liang Q, Shi Y (2022) Kd-scfnet: Towards more accurate and efficient salient object detection via knowledge distillation. arXiv preprint arXiv:2208.02178
91. Zhang X, Zhou X, Lin M, Sun J (2018) 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
92. Zhao JX, Cao Y, Fan DP, Cheng MM, Li XY, Zhang L (2019) Contrast prior and fluid pyramid integration for rgbd salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3927–3936
93. Zhao X, Pang Y, Yang J, Zhang L, Lu H (2021) Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia, ACM MM, China, vol 29, pp 2645–2653
94. Zhao X, Liang H, Li P, Sun G, Zhao D, Liang R, He X (2024) Motion-aware memory network for fast video salient object detection. IEEE Transactions on Image Processing
95. Zhong Y, Yuan B, Wu H, Yuan Z, Peng J, Wang YX (2021) Pixel contrastive-consistent semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 7273–7282
96. Zhou W, Sun F, Jiang Q, Cong R, Hwang JN (2023) Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection. IEEE Transactions on Image Processing
97. Zhu X, Hu H, Lin S, Dai J (2019) Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9308–9316