Arabian Journal for Science and Engineering (2025) 50:19673–19689
https://doi.org/10.1007/s13369-025-10301-y

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

DAMFNet: Breaking Computational Barriers in Video Salient Object Detection with Multi-Scale Deformable Appearance–Motion Fusion

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

Received: 9 October 2024 / Accepted: 4 May 2025 / Published online: 23 May 2025

Abstract

Video salient object detection (VSOD) aims to identify and segment the most visually prominent objects in videos by leveraging both appearance and motion cues. Existing VSOD models struggle with challenges such as geometric variations, occlusion, cluttered backgrounds, and complex lighting conditions while also requiring significant computational resources. In this paper, we propose a novel architecture that directly addresses these limitations: a deformable appearance–motion fusion network (DAMFNet). Our approach introduces a multi-scale deformable fusion mechanism that effectively captures both appearance and motion information, allowing precise object detection even in resource-constrained environments. DAMFNet utilizes deformable convolution (DConv) layers, depth-wise separable convolutions (DSConv), and transposed convolutions to balance performance and computational efficiency. Additionally, we propose a novel appearance–motion transfer learning (AMTL) strategy, guided by the AMTLoss function, which further enhances the model's capability to generalize across various video datasets. Our method outperforms 19 state-of-the-art VSOD models across six benchmark datasets, achieving superior accuracy with only 11.0 million parameters, 10.7 GFLOPs, and an inference speed of 75 FPS. These results position DAMFNet as a highly efficient and scalable solution for real-time video salient object detection. Our code and pre-trained models will be released to encourage further research.

Keywords VSOD · Deformable convolution · Depth-wise convolution · Multi-scale appearance and motion information · Geometric information

1 Introduction

Video salient object detection (VSOD) aims to detect and segment the objects in dynamic video scenes that capture the most visual attention, which is essential for understanding the human visual systems (HVS) and enhancing various high-level computer vision applications, including video object recognition [1], video object segmentation [2], video shadow detection [3], classification [4], autonomous vehicles [5], and some of the real-time applications such as medical image processing [6], robotic manipulation [7], surveillance systems [8], traffic management [9], drones [10], and smart homes [11], drops detection [12, 13], many more. Despite its significance, traditional VSOD methods rely on hand-crafted features such as color and heuristic priors like background and center priors, which struggle to produce accurate saliency maps in complex scenes. Further, to overcome the above, machine learning-based approaches [14–16] have been designed to address these challenges. These approaches utilize simple image pre-processing techniques such as image transformation, rotation, zoom in and out, translation, etc., and low-level handcrafted features and combine low-level features to generate the saliency maps. However, they often fail to preserve fine object details, particularly when the salient object overlaps with the image boundary or blends with the background, as low-level features are inherently limited in such scenarios. Recent advancements have leveraged convolutional neural networks (CNNs) [17–19], which offer robust visual representation capabilities to address these chal-

✉ 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, Warangal, Telangana, India

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

lenges and categories in three parts: (1) appearance-based methods, (2) motion-based methods, and (3) appearance and motion-based methods. The motion-based VSOD method transforms into a moving object segmentation (MOS) task due to motion information, which loses the target object information in slow-moving and stationary objects. The appearance-based VSOD methods use sophisticated image segmentation techniques to capture the detailed target information.

However, these approaches often suffer from mis-detection due to the absence of prior object-specific knowledge. The motion-based and appearance-based methods [21–23] address the above limitations by integrating both modality features. These approaches enhance the semantic representation by combining both modalities, where appearance features provide detailed object descriptions while motion cues help identify potential candidate regions. However, the appearance- and motion-based VSOD methods enhance the detection and segmentation performance by integrating the optical flow maps as motion. Nevertheless, optical flow-based VSOD methods face some issues. (1) The optical flow captures motion information between two frames and fails to align motion features accurately with appearance features, leading to the loss of boundary information. (2) It faces challenges in various unconstrained scenarios, including occlusion, motion blur, low-light, deformation, and clutter (as illustrated in *Fig.* 1) in detecting and segmenting the object at multiple scales.

To overcome the above issues, (1) the first approach is extracting geometric features related to object pose, scale, part, and viewpoint transformations. (2) other methods are either constructing extensive artificial training datasets incorporating a range of possible variations and performing data augmentation (e.g., affine transformations), which increases model complexity and parameter count, or employing transformation-invariant techniques, such as sliding window approach [24, 25], structure from motion (SFM) [26], and scale-invariant feature transform (SIFT) [27]. (3) It reduces the dimension of the feature vector using decomposition-based techniques [13, 28] such as the wavelet transformation, dynamic mode decomposition (DMD) and proper orthogonal decomposition (POD). These decomposition-based methods are effective in simple scenarios (e.g., motion blur, well-defined scene views); they often require preprocessing steps that add complexity and are less suited for dynamic, real-time applications. Additionally, these approaches may struggle with unknown or complex geometric transformations and rely heavily on handcrafted modules (e.g., max-pooling for translation invariance), which can increase model size and limit generalization capabilities. Some recent methods have introduced bi-directional modality transmission schemes inspired by computer network principles [29–32]. These schemes capture the shape, semantic structure, and motion during appearance and motion feature extraction and fusion. Despite their promise, these approaches often result in substantial model sizes, posing significant challenges for deployment in resource-constrained environments such as surveillance cameras and video KYC.

In recent advancements, there has been a growing emphasis on developing lightweight models for VSOD tasks. For instance, [9] introduced a dual-stream architecture that separately processes appearance and motion features, optimizing storage efficiency and reducing latency. Similarly, [33] proposed VS-Net, which leverages long skip-connections between encoder and decoder blocks to capture multi-scale spatiotemporal features. However, these approaches struggle to address geometric variations in spatial appearance and temporal locality. To mitigate this, [1] introduced a deformable separable network that extracts spatial and temporal features based on geometric variations, resolving issues related to skip-connections through an innovative intermediate module. More recently, [34] utilized ShuffleNet-V2 [35] to design a lightweight neural network capable of extracting deep, multi-modal features with multi-scale spatial context. In a different approach, [6] employed knowledge distillation to transfer knowledge from a computationally heavy teacher model to a lightweight student model, assessing the student's learning through similarity metrics. However, the lightweight nature of the student model limits its ability to capture multi-scale features, impacting overall performance. Extending this idea, [5] developed a lightweight framework with multiple heterogeneous decoders in the student network. Despite this, random initialization of the kernel order hinders the model's ability to identify informative patterns while maintaining reduced complexity. A common limitation across these models is their inability to effectively balance local and global contextual features from spatial and temporal data, resulting in a trade-off between accuracy and parameter efficiency.

To address the above challenges and balance the performance and network complexity, we propose a lightweight and efficient deformable appearance and motion fusion network (DAMFNet), which incorporates deformable convolution (DConv) layers [1, 36], depth-wise convolution (DSConv) [1, 37], and transposed convolution (TC2d) layers to simultaneously extract multi-scale geometric variations of objects locally in appearance and motion-based features while fusing globally. Furthermore, we introduce a novel multi-scale appearance and motion transfer learning (AMTL) mechanism, utilizing an AMTLoss function to transfer appearance and motion information across multiple scales, thereby enhancing the model's performance. By combining local geometric feature extraction, global fusion, and transfer learning, our approach achieves state-of-the-art results in VSOD tasks.

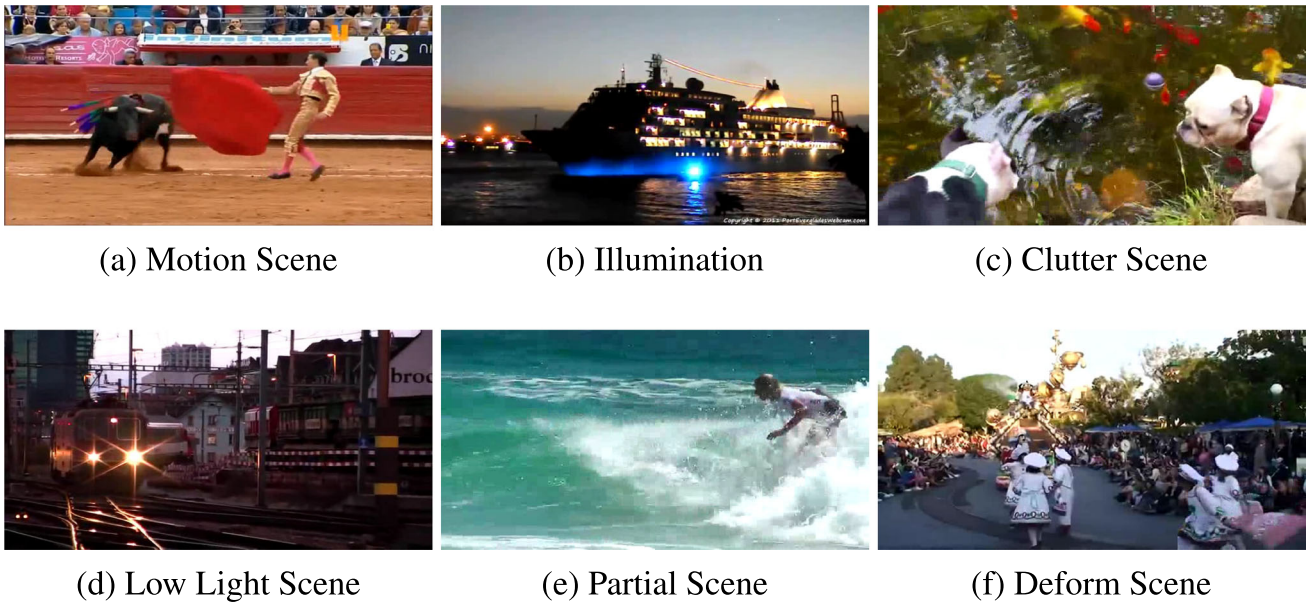


Fig. 1 Challenging scenarios from DAVSOD [20] dataset

The primary contributions of this work include:

1. A novel, efficient, lightweight DAMFNet model is designed, which uses encoder and decoder concepts with the help of DConv, DSConv, and TC2d layers to extract the geometric appearance and motion information.
2. To enhance the performance of DAMFNet, a novel multi-scale appearance and motion transfer learning (AMTL) is proposed using the AMTLoss function.
3. Further, the convolution layer (Conv2d) with 1×1 filter, Batch Normalization (BN) followed by the nonlinear activation ReLU is used after fusion of all decoder output to generate the Saliency Map.
4. Extensive experiments are carried out on the DAVSOD-Difficult dataset, revealing that DAMFNet outperforms in terms of accuracy in terms of SOTA models with less number of parameters, floating-point operations, and increasing speed.

The upcoming section and subsection are arranged in this way, related work, proposed method, experiment work, and ablation study.

2 Related Work

This section explains appearance and motion-based models, unsupervised video object segmentation models, and multi-scale, deformable, and lightweight approaches.

2.1 Appearance and Motion-based Unsupervised VSOD Models

Several appearance and motion-based approaches have tackled various VSOD tasks recently, as discussed in [21, 38]. These methods leverage low-level handcrafted features such as optical flow [29], or super-pixels [21], object proposals [39], saliency priors [40], long sparse point orientation [41], for speculative detection inference. However, these conventional models face limitations in handling complex and dynamic scenarios due to the lack of semantic representation and high-level content learning. In contrast, the emergence of the recurrent neural network (RNN) model [42] has effectively handled the long-term temporal dependencies. A motion pattern-based model proposed in [38], utilizing motion patterns from video, faces difficulties in segmenting objects across two adjacent frames, despite guiding optical flow correctly. To address these issues, various works, such as those presented in [39, 41], have proposed solutions by fusing spatiotemporal information using parallel networks. Additionally, multi-stage processing methods introduced by Li et al. [43] offer motion-based consistent features for object detection. Seo et al. [44] proposed a network that integrates comprehensive language expression to detect objects across the entire video frame. Gu et al. [39] propose a constrained self-attention (CSA) module specifically designed to extract motion information in anticipation of object movement. Mao et al. [45] introduce a deep learning method to extract multi-scale spatiotemporal cues, while dual boundary feature branches enhance focus on salient object

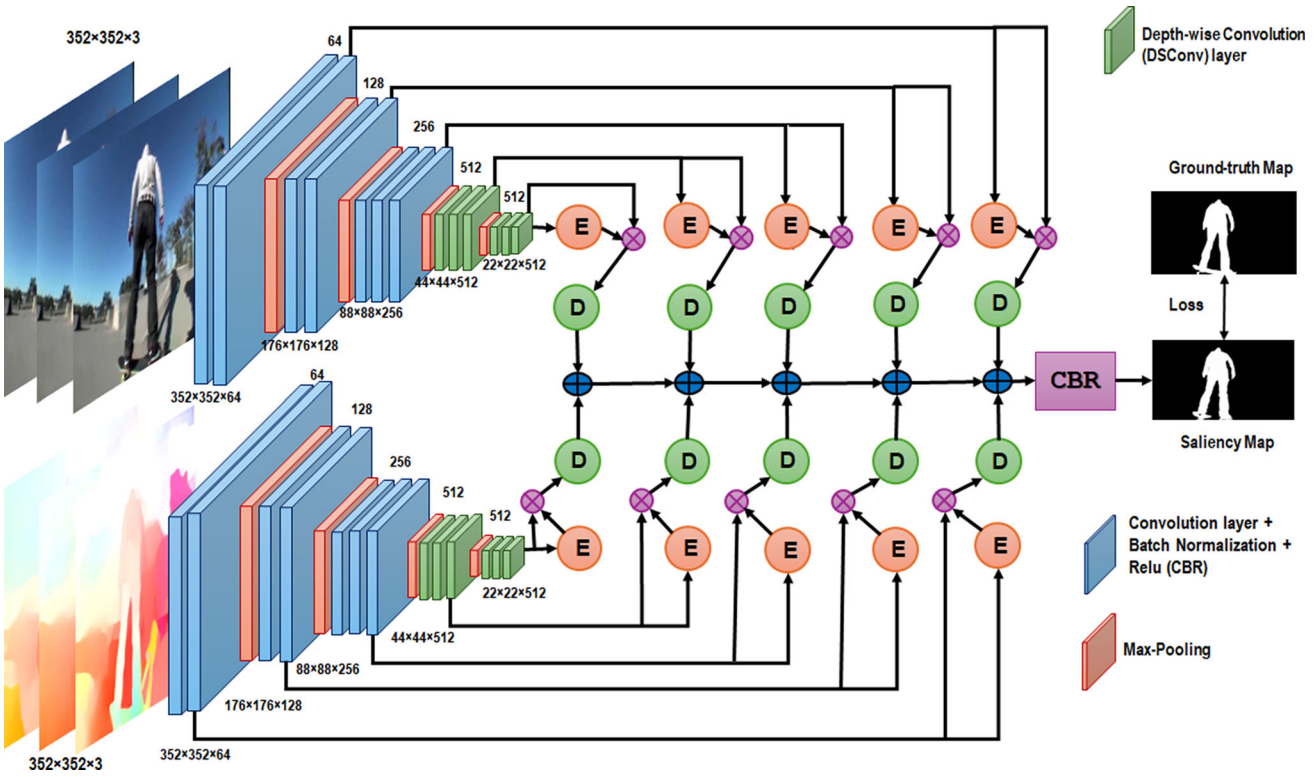


Fig. 2 Architecture of DAMFNet. Where $E_i = (i = 1, 2, 3, 4, 5)$ is the encoder blocks, $D_i = (i = 1, 2, 3, 4, 5)$ is the decoder blocks, $\otimes$ is element-wise multiplication, and $\oplus$ is element-wise addition operation

boundaries, and a feature alignment branch integrates and aligns multi-scale features across internal and external pathways.

Unsupervised VSOD is closely linked to attention-based UVSOD tasks, aiming to derive attention-aware information from video clips. Traditional models, as seen in [46, 47], calculate saliency by employing handcrafted information, and consistency across consecutive frames has been maintained by spatiotemporal optimization. Additionally, approaches like [48, 49] focus on extracting highly semantic spatiotemporal information for end-to-end object detection. Several deep learning models, including [50–52], extract motion information from optical flow or information from two consecutive frames. Ballas et al. [53] introduced the key-frame method to identify the high-quality video frames for saliency object categorization. Another approach in [54] detects salient objects by extracting spatial–temporal information from high-quality frames. Addressing challenges such as analyzing relative saliency and limited labeled data [48] [55] in VSOD, researchers have developed techniques that enhance the quality of temporal information. For instance, Fan et al. [56] present a Shift-Aware ConvLSTM to extract spatial and temporal features from high-quality annotations VSOD datasets. Zhang et al. [30] employed deep learning techniques to extract spatial and temporal feature similarities across consecutive frames. Han et al. [57] introduce OVSNet, an open-vocabulary saliency-guided progressive refinement network for Unsupervised Video Object Segmentation (UVOS), which integrates saliency cues from foundation models like CLIP and CLIPSeg with motion cues to generate an OVS attention map. These cues, combined with a fine-tuned Segment Anything Model (SAM) via lightweight adapters, progressively refine object representations in the appearance branch, yielding precise segmentation masks.

2.2 Deformable Multi-scale Feature Extraction Techniques

Most existing multi-scale feature extraction approaches, such as those by Singh et al. [58], Zhang et al. [59], and others [60–62], have relied on multi-encoders, ASPP modules, and varying kernel dilation rates. For instance, Singh et al. employed the ASPP module to extract multi-scale features, while Zhang et al. introduced a multi-scale information enhancement (MIE) module to enrich shared information by converting RGB features into point features across scales. Other works, like those of Zhang et al. [60], leveraged multi-scale graph neural networks to capture spatial and disparity correlations, and Liu et al. [61] proposed a multi-

scale deformation module (MSDM) for adapting to varying shapes of salient objects. Peng et al. [62] further utilized a multi-scale encoder–decoder network for semantic feature extraction. However, these state-of-the-art methods often suffer from increased computational complexity and information leakage. Sharma et al. [32] introduce a bidirectional multi-scale spatiotemporal network (BMST-Net) for salient video object detection, leveraging an encoder–decoder architecture to learn spatiotemporal feature representations. Singh et al. [63] introduce a lightweight deformable multi-scale fusion network that jointly extracts attention-guided multi-scale features and geometric features to produce highly efficient saliency maps. Additionally, geometric multi-scale pixel-level contrastive learning (GMPCL) uses GMPCL loss to enhance geometric feature representations and distinctly separate foreground and background features at the pixel level. To address these issues, we propose a multi-scale encoder, which reduces computational demands by employing deformable and depth-wise separable convolutions for a more efficient fusion of multi-scale and geometric information.

In recent years, deformable convolution-based methods [1, 36, 64, 65] have emerged as a powerful solution to adaptively capture geometric spatial structures in objects, addressing limitations in traditional convolutional neural networks (CNNs), which rely on fixed kernel structures. Dai et al. [36] introduced the pioneering deformable convolution network, which utilizes convolution offsets to adapt to geometric variations. However, while effective in capturing spatial structures, it falls short in recognizing regions of interest. Building on this, Zhu et al. [64] proposed Deformable Convnets v2, incorporating an additional modulation mechanism to enhance region-level modeling. Despite these advancements, the challenge of handling long-range dependencies in spatial and temporal information persisted. To address this, Wang et al. [65] developed the InternImage ViT-based technique, which generates a large effective receptive field for improved detection and segmentation tasks. Further advancements were made by Deng et al. [66], who introduced spatiotemporal deformable convolution (STDC) for effective motion information extraction and fusion. Finally, Singh et al. [1] presented DSNet, which leverages attention mechanisms to extract spatial and temporal information without significantly increasing model parameters. Singh et al. [67] introduce a deformable separable fusion network (DSFNet) that dynamically captures multi-scale geometric spatiotemporal variations while maintaining computational efficiency. Additionally, a swarm-enhanced Adam (SEAdam) optimizer was introduced, which adaptively balances local and global gradient exploration and exploitation, significantly accelerating convergence. The problem with this model is over-parameter learning.

2.3 Lightweight VSOD Approaches

CNN-based VSOD models [9, 68–70] often leverage semantic information from pre-trained ImageNet backbones. However, these backbones can suffer from information leakage and redundancy. Cheng et al. [71] addressed this with a highly lightweight model trained from scratch, which, while effective at feature extraction, struggles with cluttered backgrounds and deformations. Hu et al. [5] introduced a lightweight model with heterogeneous decoders and 3D convolutions to enhance accuracy, but it does not address training and inference time constraints and falls short in handling deformations. Singh et al. [33] developed the VS-Net model to leverage multi-scale spatiotemporal features for salient document detection but faces limitations due to long feature dependencies. In response, Hu et al. [9] proposed a dual-stream network for appearance and motion representations, yet it does not fully address the issues of feature sparsity. Singh et al. [1] recently introduced DSNet, which improves upon previous models by minimizing training and testing times through separability and deformability concepts. Su et al. [72] introduce the Unified Framework for Group-based Segmentation (UFGS) that leverages transformer blocks to model long-range dependencies among image patches and enhances the structural similarity of patches. The intra-MLP learning module incorporates and generates self-masks to mitigate partial activation issues, leading to improved segmentation precision. Xu et al. [73] integrated the Segment Anything Model (SAM) into their pipeline for video segmentation tasks and utilized edge information to refine segmentation labels and reduce noise interference. Additionally, a global-aware loss function introduces to capture global semantic relationships, significantly enhancing salient object detection but draping the feature geometric structure. Zhao et al. [74] incorporated a space-time memory (STM)-based network featuring an encoder–decoder architecture to extract temporal features from consecutive frames and employed spatial–temporal fusion to enhance object details and reconstruct saliency maps. A motion-aware loss function is introduced to facilitate multitask learning, simultaneously improving VSOD and object motion prediction while preserving object integrity while lacking to preserve the long-term temporal prediction. Huang et al. [75] presented a lightweight VSOD architecture that utilizes a ShuffleNet-V2 backbone for efficient feature extraction. The architecture is augmented with a depth-wise multi-scale pooling module (DMPM) to aggregate multi-scale contextual information compactly. Furthermore, a shuffle-enhanced multi-modal fusion module (SMFM) is employed to progressively fuse spatial and temporal information, achieving state-of-the-art accuracy with a substantially reduced model size. However, these models face problems in balancing the network complexity and performance.

3 Proposed Methodology

The appearance and motion frames are passed into the two parallel Backbone Networks VGG-16 [76], which extract the backbone appearance and motion information individually as shown in *Fig.* 2. This information is passed to encoder blocks, which encode the geometric appearance and motion information at multiple scales and then cross-multiply with an output of VGG-16 blocks. Next, decoder blocks decode the multi-scale geometric appearance and motion information. Further, the output of all the decoders is fused together and transferred to CBR blocks (Convolution, batch normalization, ReLU) layers to generate the saliency maps. In the subsequent subsections, a detailed explanation of the DAMFNet model will be provided including its associated loss function and elucidates the process of generating the saliency map.

3.1 DAMFNet Architecture

The previous SOTA appearance and motion-based methods [4, 29, 55] extracted appearance and motion information and fused together but failed to capture the geometric variation of localized information of object background and foreground dynamically. Motivated by this, a novel, efficient, lightweight deformable appearance–motion fusion network (DAMFNet) is designed to enhance the capture of precise location information and preserve location boundaries effectively. It has two branches to extract the geometric appearance and motion features separately. The DAMFNet has used the VGG-16 [76] backbone network, which has five blocks, each block dimension is 64, 128, 256, 512, and 512, to extract the backbone appearance information and motion information independently. The proposed architecture has five encoder blocks (E_i= i=1,2,3,4,5) and five decoder blocks (D_i, i=1,2,3,4,5), which are connected with backbone network blocks. Each backbone network block is connected with each encoder block and skip-connected to multiply before passing to each decoder block. The Encoder Blocks (E_i) is the combination of two DConv, two DSConv, MaxPooling, and PReLU. The Decoder Blocks (D_i) is the combination of two DConv, two DSConv, and TC2d followed by a ReLU activation function. The DConv and DSConv configuration is given in [1]. The DConv layers extract the geometric appearance and motion information dynamically. The DSConv layers extract the appearance and motion information with fewer network parameters during encoding and decoding. Next, an element-wise multiplication operation is performed between each VGG-16 [76] backbone network block (BN_i, i=1, 2, 3, 4, 5) outputs (BO_i, i=1, 2, 3, 4, 5) and Encoder Block (E_i, i= 1, 2, 3, 4, 5) outputs (EO_i, i=1, 2, 3, 4, 5) node feature vectors. After that passed to the decoder blocks, which decode the geometric appearance and motion features separately. Further, these appearance and motion features are fused together using element-wise addition operation ($\oplus$) to generate the generalized latent representation of multi-scale geometric appearance and motion information. Next, all blocks fused multi-scale geometric appearance and motion information is fused together to enhance the representation of feature quality. The fused multi-scale geometric appearance and motion information is passed to the Conv2d with 1 × 1 filter, BN, followed by ReLU, which generates the saliency map (SM_k).

Algorithm 1: Deformable Appearance Motion Fusion Net (DAMFNet).

Input: A_k: Appearance frames, M_k: Motion frames, and GT_k: Annotation frames
Output: SM_k: Saliency map.

1. The A_k, M_k, and GT_k are given to the proposed DAMFNet and passed to two branches of VGG-16 Network parallelly.
2. The backbone VGG-16 network extracts the backbone appearance and motion and generates backbone appearance and motion information via *Eq.* 1.
3. The Encoder blocks extract geometric appearance and motion information and enhance representation after applying element-wise multiplication operation via skip connection using *Eq.* 2 and 3.
4. The decoder appearance and motion information are fused together via *Eq.* 4.
5. The Adam optimizer is used to update the DAMFNet weight parameter and minimizes the MSG loss function via *Eq.* 6, 7 and 8.
6. At last, the saliency map is generated by Conv2d, BN followed by ReLU using *Eq.* 5.

3.2 Appearance and Motion Feature Extraction

Consider a dataset with T video clips, each comprising k consecutive frames (where $k = 1, 2, \cdots, T$), which includes appearance frames $(A_k)_{k=1}^{T}$, motion frames $(M_k)_{k=1}^{T}$, and corresponding annotation maps $(GT_k)_{k=1}^{T}$. The motion frames are generated using BSCNet [58]. These appearance and motion frames are passed DAMFNet, where at first, the first backbone appearance $X_k^{b_p}$ and motion $Y_k^{b_p}$ for (p = 1, 2, 3, 4, 5) information are extracted using the Backbone Network VGG-16 [76], which has five blocks with dimensions (64, 128, 256, 512, 512), respectively. The process is given in *Eq.* 1.

$$\begin{aligned} X_k^{b_p} &= \text{VGG-16}(A_k), \text{ for } p = 1, 2, 3, 4, 5 \\ Y_k^{b_p} &= \text{VGG-16}(M_k), \text{ for } p = 1, 2, 3, 4, 5 \end{aligned} \tag{1}$$

These backbone appearances and motion information are passed to five encoder blocks, which extract the geometric appearance and motion information. The backbone output is cross-multiplied to encoder output to extract the inconsistent

representation of geometric appearance and motion information at multiple scales. The process is given in *Eq.* 2.

$$\begin{aligned} X_k^{e_q} &= X_k^{b_p} \otimes E(X_k^{b_p}), \text{ for } p = 1, 2, 3, 4, 5 \\ Y_k^{e_q} &= Y_k^{b_p} \otimes E(Y_k^{b_p}), \text{ for } p = 1, 2, 3, 4, 5 \end{aligned} \tag{2}$$

These cross-multiplied multi-scale geometric appearance and motion information are decoded via decoder blocks and enhance the representation of information. The process is given in *Eq.* 3

$$\begin{aligned} X_k^{d_r} &= D(X_k^{e_q}), \text{ for } q = 1, 2, 3, 4, 5 \\ Y_k^{d_r} &= D(Y_k^{e_q}), \text{ for } q = 1, 2, 3, 4, 5 \end{aligned} \tag{3}$$

Next, these decoded geometric information at multiple scales are fused together to generalize the discriminative representation of information. The process is given in *Eq.* 4.

$$f_k^t = X_k^{d_r} \oplus Y_k^{d_r}, \text{ for } r = 1, 2, 3, 4, 5 \tag{4}$$

At last, the fused appearance and motion information f_k^t at time t is passed to the Conv2d layer, which has 1×1 filter to convert high-level information to low-level information, then applying the BN followed by ReLU to normalize the appearance and motion information representation and produce the saliency map. The procedure is shown in *Eq.* 5.

$$SM_k^t = ReLU(BN(Conv2d(f_k^t))) \tag{5}$$

The multi-scale global (MSG) loss is optimized using an ADAM optimizer during the training.

3.3 Appearance and Motion Transfer Learning (AMTL)

The AMTL is a novel approach to transferring knowledge from one modality to another through the integration of multi-scale appearance and motion transfer learning. The AMTL addresses the inherent challenges in detecting salient objects across varying scales and motion contexts within video sequences, which are often exacerbated by dynamic backgrounds and complex object interactions. To capture geometric multi-scale features effectively, we employ a hierarchical encoder and decoder architecture that leverages geometric multiple levels of feature extraction and preserves fine-grained details at different scales, which is crucial for accurately identifying salient objects regardless of their size or spatial context. To transfer the knowledge, we proposed multi-scale appearance and motion transfer loss (AMTLoss), which effectively calculates the similarity transfer by each modality to fuse features and guide the network efficiently in the right direction. Basically, let A $\in A^{h\times h\times 3}$ as an appearance feature, and M $\in M^{h\times h\times 3}$ denotes a motion feature. Let F $\in R^{h\times h\times 3}$ denote the fused appearance and motion feature generated after applying element-wise addition on the parallel decoder output. The multi-scale MATLoss is derived as follows in *Eq.*6 using the KL-divergence loss [77] with average multiple scales between appearance, motion, and fuse feature.

$$\begin{aligned} &AMTLoss(F_k, A_k, M_k) \\ &= \frac{1}{N}\sum_{k=1}^{K} \alpha \\ &\quad \times [F_k \times log(F_k) - F_k \times log(A_k)] + (1-\alpha) \\ &\quad \times [F_k \times log(F_k) - F_k \times log(M_k)] \end{aligned} \tag{6}$$

where F_k is the fused feature, A_k is appearance features, M_k is motion features at k video clips, α is the learning parameter which is fixed = 0.6, N is the number of samples, and K is the multi-scale kernel (5, 10, 15, 20). This MATLoss not only enhances the model's ability to handle varying scales and motion patterns but also improves its performance in real-world scenarios where traditional methods often struggle. It is used in fine-tuning the proposed DAMFNet model.

3.4 Multi-Scale Global (MSG) Loss

The binary cross-entropy loss (BCE) is employed to independently calculate the loss of each pixel, providing a pixel-level constraint on the network. To address its limitation of neglecting global structural information, [88] introduced the intersection over union loss (IoU) to focus on the global structure and impose a global constraint on the network. However, these losses treat all pixels equally, disregarding potential differences between them. Building upon this, [89] enhanced the aforementioned losses by introducing the weighted binary cross-entropy loss (BCE) and IoU loss (IoU). In this approach, each pixel is assigned a different weight based on calculating the difference between the center pixel and its surrounding environment. This weighting mechanism aims to provide more attention to challenging pixels. However, it is unable to recognize the hardness and softness of the pixel value in geometric variations of objects in multiple scales. To overcome this, the multi-scale global loss is included with weight using *Eq.* 7.

$$w = 1 + \frac{1}{K}\left(\sum_{i=1}^{K} SM_{k_i} - GT_{k_i}\right)/GT_k \tag{7}$$

where K is the different kernel sizes such as 3, 5, 10, SM_{k_i} and GT_{k_i} are the saliency maps and annotation maps at i^{th} scale. The MSG loss is calculated as follows, shown in *Eq.*8.

Table 1 Performance comparison of DAMFNet with nineteen existing VSOD models across six datasets. Top three results highlighted in bolditalic, bold, and italic

		Backbone	# Param	FLOPs	Speed	DAVIS			FBMS			DAVSOD			SegTrack-V2			MCL			DAVSOD-Diff		
	$\text{Model}_{Yr.}[Ref.]$	Network	(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	DefED-Net_{21} [78]	ResNet-101	14.5	222.4	39.5	0.853	0.830	0.032	0.847	0.834	0.050	0.675	0.600	0.106	0.848	0.807	0.028	0.698	0.678	0.041	0.491	0.432	0.148
	EUVSOD_{22} [9]	MobileNetv2	**6.40**	***5.40***	32.5	0.920	0.894	0.150	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***
	VS-Net_{23} [33]	ResNet-50	*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} [5]	HD2S	***3.94***	**7.95**	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
	InternImage_{23} [65]	M3I	49.0	270.0	32.0	0.890	0.880	0.030	0.775	0.765	0.086	0.654	0.552	0.130	0.751	0.676	0.038	0.721	0.688	0.041	0.527	0.445	0.150
	DSNet_{23} [1]	ResNet-50	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*
	DAMFNet	VGG-16	11.0	10.7	**75.0**	**0.925**	**0.916**	0.018	**0.914**	**0.896**	*0.027*	***0.811***	***0.782***	***0.054***	***0.899***	***0.879***	***0.015***	***0.861***	***0.839***	***0.022***	*0.549*	***0.501***	**0.096**
Large VSOD Models	EREST_{21} [79]	ResNeXt101	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} [29]	ResNet-50	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
	MSDM_{21} [80]	ResNet-50	361.2	397.4	24.0	0.899	0.884	0.028	0.788	0.773	0.092	0.629	0.515	0.139	0.760	0.664	0.045	0.683	0.647	0.068	0.512	0.476	0.172
	CFAM_{22} [81]	DeepLabv3	59.3	425.7	30.0	0.918	*0.909*	**0.015**	***0.915***	***0.900***	**0.026**	0.753	0.662	0.083	0.890	0.857	**0.015**	0.838	0.804	*0.033*	0.459	0.399	0.110
	SKD_{22} [6]	ResNet-50	76.32	75.40	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
	HCPN_{23} [82]	ResNet-101	181.1	126.3	33.0	0.849	0.798	0.039	0.813	0.793	0.074	0.719	0.663	0.101	0.821	0.789	0.034	0.797	0.767	0.075	0.457	0.439	0.151
	TMO_{23} [2]	ResNet-101	172.4	246.8	43.2	0.900	0.879	0.025	0.801	0.761	0.094	0.697	0.644	0.114	0.727	0.689	0.053	0.718	0.680	0.042	0.490	0.421	0.140
	PMN_{23} [83]	VGG-16	160.2	93.8	37.2	0.881	0.864	0.044	0.843	0.811	0.088	0.765	0.741	0.087	0.865	0.851	0.060	0.851	*0.834*	0.083	0.417	0.397	0.156
	SPGO_{23} [84]	DINO/MoCo	154.0	111.7	24.0	0.798	0.775	0.049	0.726	0.700	0.062	0.756	*0.749*	0.093	0.831	0.779	0.067	0.789	0.738	0.065	0.459	0.437	0.130
	PACNet_{23} [85]	ResNet-50	127.9	147.2	49.2	0.912	0.904	*0.016*	0.891	0.880	0.032	*0.801*	0.732	*0.060*	**0.908**	**0.884**	0.020	0.849	0.830	0.051	0.486	0.466	0.137
	CoSTFormer_{23} [86]	ResNet-50	333.4	287.4	49.9	*0.921*	0.903	***0.014***	0.889	0.856	0.046	**0.806**	0.731	0.061	*0.904*	0.870	*0.016*	0.783	0.752	0.057	0.479	0.459	0.141
	STDF_{24} [66]	DCNv2	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} [87]	DeepLabv3+	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

$$MSG_{loss} = w \otimes l_{bce}(SM_k, G_k) + w \otimes l_{IoU}(SM_k, GT_k) + w \otimes AMTLoss(SM_k, A_k, M_k) \quad (8)$$

where SM_k is the predicted saliency maps, GM_k is the annotation maps, $\otimes$ is the element-wise multiplication, A_k is the multi-scale appearance features, M_k is the motion features, and + is element-wise addition operation.

4 Experiments and Result Analysis

4.1 Experimental Setup

The experimental evaluation of the proposed DAMFNet model is conducted on a 64-bit Ubuntu 18.04 system equipped with 32 GB RAM and a 1 TB Hard Disk. The system is powered by a 16 GB P5000 NVIDIA GPU, configured using the 460 NVIDIA Driver, CUDA 11.2, and CuDNN 8.5. Anaconda 3.9 and PyTorch version 1.12.0, along with OpenCV, are installed on the GPU machine. For consistency, the input frames, motion maps, and corresponding annotation maps are resized to dimensions of 352×352. To optimize the MSG loss and AMTLoss function with multi-scale training weight (0.75, 1.0, 1.25, 1.50), the weighted Adam optimizer is employed, which has a $1e^{-3}$ learning rate and $5e^{-4}$ weight decay.

4.2 Datasets and Evaluation Metrics

The proposed DAMFNet model undergoes experiments across six benchmark datasets: DAVIS-16, MCL, FBMS, SegTrack-V2, DAVSOD-19, and DAVSOD-Difficult. DAVIS-16 [90] stands out 50 video clips, with 30 for training and 20 for testing. MCL [91] consists of 9 videos, while FBMS [92] features 59 videos of natural scenes, divided into 29 training and 30 testing videos. SegTrack-V2 [43] has 13 testing video sequences. Additionally, DAVSOD-19 [93] includes 61 training video sequences and 81 testing sequences, and DAVSOD-Difficult-20 [93] comprises 20 video sequences specifically for testing purposes. The proposed DAMFNet model is assessed on the test datasets using the F-measure, S-measure, and mean absolute error (MAE) [90]. Additionally, computational cost metrics, including the network parameters (# Param) in a million (M), the floating-point operations (FLOPs) in Giga bytes (G), and latency performance measured in frames per second (FPS), are utilized to gauge the efficiency. The detailed results are presented in Table 1.

4.3 Training Performance

To train the proposed network, we follow two distinct approaches (1) Pretrain-RGB: is conducted on the DUTS dataset [29], which consists of 10,567 images. (2) Finetune: is performed on 2,973 appearance frames, which include 2,373 frames from DAVIS (30 videos) and 600 frames from FBMS (29 videos). The pre-trained weights are fixed and utilized throughout the fine-tuning process. The appearance and motion transfer learning (AMTL) is used in both processes, which update the weight and transfer from each block of the backbone to the decoder using the AMTLoss function. The weighted Adam Optimizer is employed to train DAMFNet on these datasets, minimizing the MSG loss, and information extraction is given in *Algorithm* 1. The vanishing gradient, geometric variation, and overfitting challenges are tackled using batch normalization, deformable convolution layer followed by ReLU activation function during the training of each spatial branch. The fine-tuning process takes nearly 6 hours, encompassing 25 epochs with 8 batch sizes.

4.4 Testing Analysis

The effectiveness of the proposed DAMFNet model is evaluated on six test datasets, including DAVIS16 [90] with 20 videos, FBMS [92] with 30 videos, DAVSOD [29] (Easy and Difficult) with 81 and 20 videos, MCL [91] with nine videos, and SegTrack-V2 [43] with 13 videos, as outlined in Table 1. The performance of DAMFNet is assessed using metrics, (i) S_α, (ii) F_β, and (iii) MAE. Additionally, the model's computational complexity is measured with all the lightweight VSOD and heavyweight VSOD, which shows that DAMFNet performs better than the heavyweight VSOD model in performance as well as complexity, also, but compared to lightweight, it is four and performs better than all.

4.5 Comparative Analysis

The performance comparison of the proposed DAMFNet model is evaluated across six datasets with nineteen SOTA models, including DefED-Net [78], VS-Net [33], TinyHD [5], EUVSOD [9], DSNet [1], InternImage [65], FSNet [29], TMO [2], SPGO [84], EREST [79], MSDM [80], CFAM [81], SKD [6], HCPN [82], PMN [83], PACNet [86], CoSTFormer [86], STDF [66], LSTA [87] in terms of S_α, F_β, MAE, and network complexity. The comparative results are presented in *Table* 1, which demonstrate consistent outperformance of DAMFNet in comparison of twenty SOTA models across DAVIS, DAVSOD, MCL, and DAVSOD-Difficult datasets. The above VSOD methods rely on simple feature extraction strategies, such as image texture, color features, and spatiotemporal priors, but struggle in complex scenarios due to the lack of high-level semantic understanding. Implicit motion encoding methods enhance performance by using learning-based architectures that integrate motion cues through concatenation, addition, LSTM structures, and cross-frame attention. While these models outperform tradi-

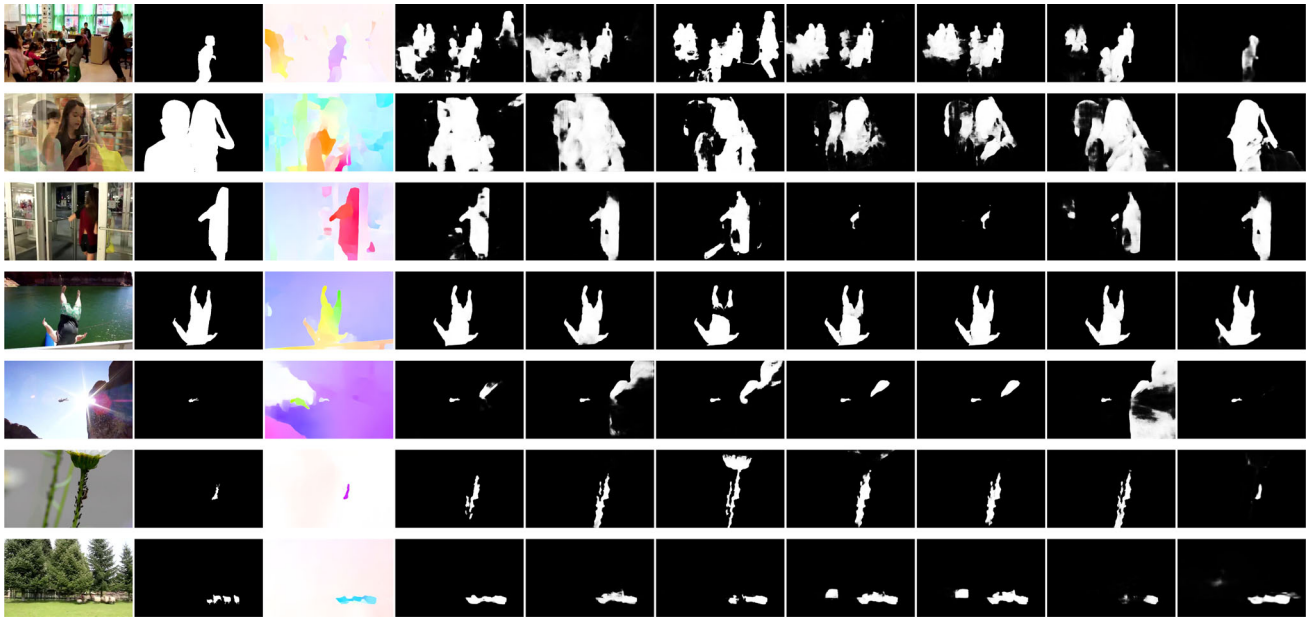


Fig. 3 Comparison of DAMFNet model and six SOTA Models on more challenging scenarios of DAVSOD and DAVSOD-Difficult datasets. where IF denotes input frames, GT denotes annotation frames, and MT denotes motion frames

Fig. 4 Performance evaluation of proposed DAMFNet model and SOTA Models in terms of # network parameter and FLOPs

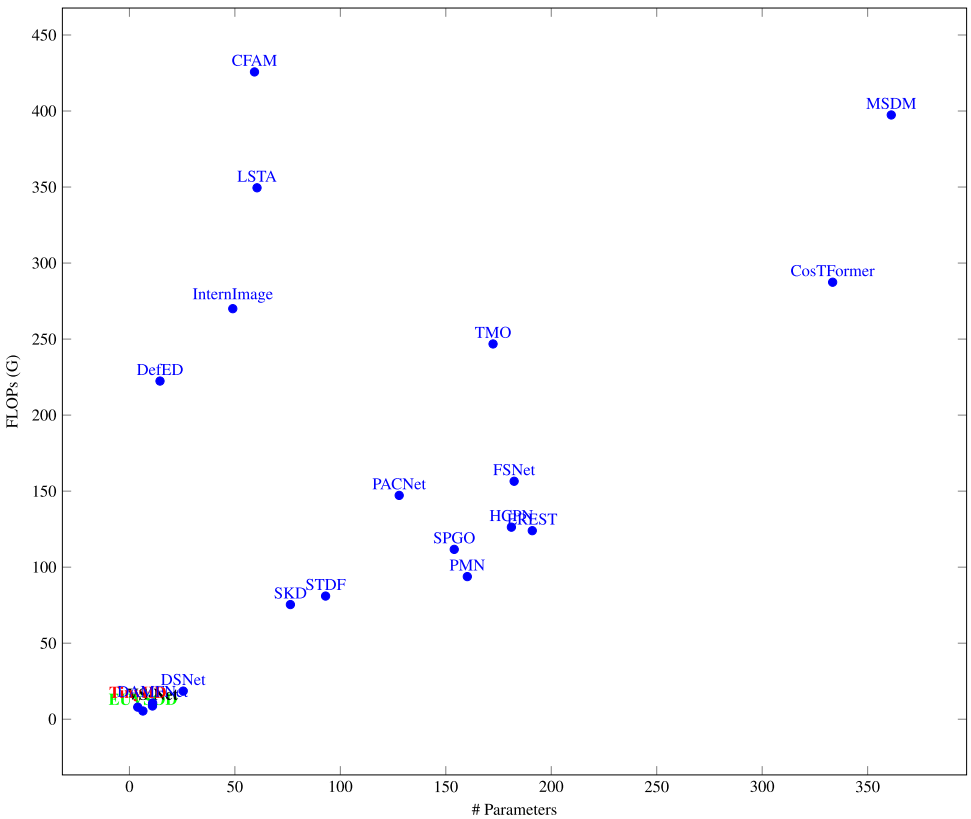
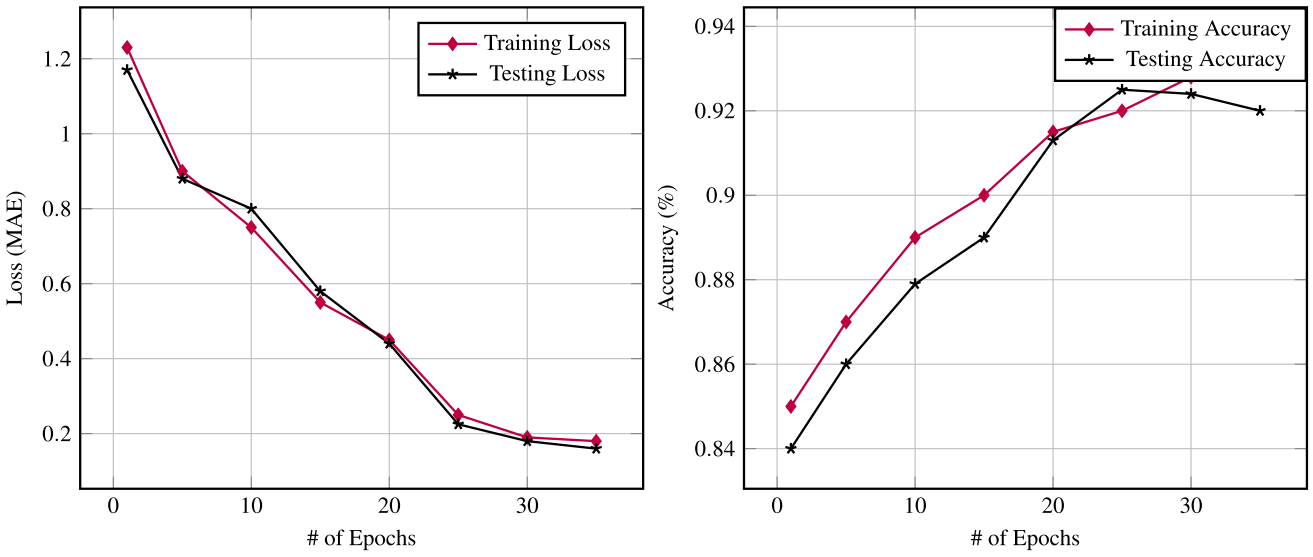


Table 2 Comprehensive analysis of hyperparameter tuning at DAMFNet

	Hyperparameter		DAVIS		FBMS		MCL		SegTrack-V2		DAVSOD		DAVSOD-Diff	
S.No.	Learning Rate	Weight Decay	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
1.	$1e^{-1}$	$5e^{-2}$	0.879	0.037	0.841	0.042	0.779	0.053	0.790	0.050	0.710	0.069	0.428	0.121
2.	$1e^{-2}$	$5e^{-3}$	0.901	0.026	0.867	0.036	0.829	0.048	0.799	0.045	0.729	0.066	0.449	0.117
4.	$1e^{-3}$	$5e^{-4}$	**0.925**	**0.017**	**0.914**	**0.027**	**0.861**	**0.022**	**0.879**	**0.015**	0.811	**0.054**	**0.549**	**0.096**
3.	$1e^{-4}$	$5e^{-5}$	0.916	0.020	0.877	0.034	0.849	0.038	0.856	0.033	**0.817**	0.056	0.540	0.100
5.	$1e^{-5}$	$5e^{-6}$	0.910	0.023	0.890	0.029	0.840	0.039	0.848	0.036	0.804	0.058	0.534	0.103

Table 3 Comprehensive hyperparameter analysis in terms of # of Batch Size and # of Epochs to fine-tune the DAMFNet

S.No.	Hyperparameters			DAVIS		FBMS		MCL		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	# Batch Size	# Epochs	Fine-tuning Time (H)	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
1.	4	15	4.10	0.878	0.034	0.843	0.038	0.789	0.054	0.787	0.053	0.710	0.069	0.428	0.122
2.	6	20	5.20	0.893	0.029	0.859	0.031	0.842	0.035	0.838	0.048	0.769	0.063	0.459	0.115
3.	**8**	**25**	**6.37**	**0.925**	**0.017**	**0.914**	**0.027**	**0.861**	**0.022**	**0.879**	**0.015**	**0.811**	**0.054**	**0.549**	**0.096**
4.	10	30	9.10	0.917	0.023	0.890	0.031	0.853	0.030	0.863	0.021	0.803	0.060	0.530	0.101
5.	12	40	12.02	0.913	0.026	0.880	0.027	0.850	0.035	0.868	0.025	0.797	0.063	0.534	0.099



(a) Loss vs. # of Epochs (b) Accuracy Vs. # of Epochs

Fig. 5 Performance comparison between loss (MAE) vs. # of Epochs and accuracy vs. # of Epochs of our proposed DAMFNet model

tional approaches, they remain susceptible to background clutter, leading to performance degradation. Our method addresses these challenges by effectively capturing motion and appearance features while leveraging cross-element-wise multiplication to enhance architectural efficiency. It achieves top performance across multiple datasets while maintaining network efficiency. Notably, on the DAVSOD-Difficult dataset, our model DAMFNet outperforms by 4.1% in S_α and 4.5% in F-measure, demonstrating superior robustness. Unlike DSNet [1], EREST [79], LSTA [87], and STDF [66], which rely on extensive training with the challenging DAVSOD-Difficult dataset, our method achieves strong generalization without additional augmentation, proving its effectiveness in real-world VSOD tasks. This comparison highlights the superior accuracy of DAMFNet in efficiently generating saliency maps, while the computational comparison results are given in *Fig.* 4, showing that the proposed model is able to detect the salient object in difficult scenarios, such as partial occlusion, deformation, small scene, and illumination scene, due to efficiently extracting and fusing

geometric features at multiple scales. Additionally, we conducted experiments on various Operating Systems (OS) to evaluate the effectiveness of performance. The results, presented in *Table* 4, demonstrate that the performance of the proposed DAMFNet does not significantly vary across different platforms.

Hyperparameter tuning To compare the model performance with various hyperparameters, we perform the experiments with various values of hyperparameters such as learning rate, weight decay, number of epochs, and number of batch size, which is shown in *Table.* 2 and *Table.* 3. From the tables, we see that as these parameters increase, the performance of the proposed models increases, but at some point it starts to downgrade the performance of the model and increases the fine-tuning times (in terms of Hours (H)), which impacts the network convergence and generalization. Additionally, the performance of the proposed model (DAMFNet) is illustrated in *Fig.* 5 in terms of number of epochs vs. Loss and number of epochs vs. accuracy, which demonstrates that as the number of epochs increases, the proposed model converges the loss efficiently and increases the performance due to the use of multi-scale global (MSG) loss.

Qualitative comparison In *Fig.* 3, DAMFNet is systematically compared with six SOTA methods across various challenging scenarios. The model demonstrates the ability to discern salient objects with coherent boundaries in challenging situations, including instances of noise and occlusion between foreground and background (1st, 3rd, and 5th rows), cluttered backgrounds with low light (2nd and 5th rows), deformed objects with motion blur (1st and 4th rows), motion blur combined with illumination scenarios (2nd and 4th rows), and small objects with deformation (3rd and 6th rows). *Fig.* 3 clearly illustrates the efficient object detection capabilities of our proposed DAMFNet in these diverse and challenging scenes. The 1st row shows the crowd scene, the 3rd-row partial occlusion, and the 7th row shows the object that is at a very long distance; all these scenarios show that the proposed model and the SOTA models face difficulty in detecting them efficiently, which has not been explored in future works. Apart from that, the proposed model faces the challenge of balancing the proper long and short-range motion dependency, which will be explored in future.

4.6 Failure Cases and Future Works

The proposed DAMFNet model demonstrates significant advancements and exhibits certain limitations when compared to four state-of-the-art (SOTA) models, as illustrated in *Fig.*6. Specifically, the model struggles with detecting shadows under low-light contrast conditions, as highlighted in row 1, and encounters challenges in handling deforming objects under varying lighting conditions, as shown in rows 3 and 4. Additionally, row 2 reveals difficulties in distinguishing object boundaries from the background at large scales with deformation, a limitation shared by both the proposed model and existing SOTA approaches. These failure cases underscore the need for further refinement in handling complex visual scenarios, particularly those involving illumination variability, object deformation, and scale diversity. To address these challenges, future work will explore the integration of knowledge distillation techniques to enhance the model's ability to generalize across diverse conditions by leveraging insights from more robust teacher networks. Furthermore, multi-domain physics informed-based contrastive learning will be investigated to improve feature representation learning, enabling the model to better discriminate between objects and their backgrounds under challenging conditions. These advancements aim to push the boundaries of current capabilities, paving the way for more robust and adaptable models in salient object detection and related tasks.

4.7 Ablation Analysis

To evaluate the efficacy of the proposed model (DAMFNet), we conduct an ablation analysis to dissect the contributions and performance impacts of its individual components and parameter configurations. The ablation results demonstrated in *Table* 5 have an impact on achieving effective multi-scale geometric learning from appearance and motion information. The results demonstrate a clear enhancement in performance with the progressive integration of each component, as shown by the superior performance of the full model compared to the baseline configuration. Specifically, the comparative analysis between configurations No. 1 and No. 8 illustrates a significant performance improvement attributable solely to the added component in DAMFNet. Additional findings are presented in *Table* 6, which shows the design of the proposed network. From *Table* 6, the first row shows default means simple autoencoder concepts, and in other rows, two to four individual components are added and compared in their complexity and performance. The last row combines all component configurations, which we have underscores the robustness of the proposed DAMFNet.

4.7.1 Effectiveness of the DSConv, DConv, and Conv2d

The proposed DAMFNet network leverages the strengths of DSConv, DConv, and Conv2D to achieve efficient and robust performance for VSOD tasks. DSConv enhances computational efficiency by decomposing standard convolutions into depth-wise and pointwise operations, reducing redundancy while preserving spatial information. DConv improves dynamic feature adaptation by modulating kernel weights based on input variations, enabling effective motion-aware representations. Conv2D serves as the foundation for capturing local spatial details, ensuring stable feature

Table 4 Comparative study of various OS systems of DAMFNet

OS	Parameters			DAVIS		FBMS		MCL		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	# Training Time (m)	Testing Time (m)	Speed(FPS)	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
Windows (64 bit) (Python)	2.5	4.0	72	0.914	0.023	0.878	0.029	0.830	0.039	0.839	0.035	0.753	0.060	0.489	0.108
Windows (64 bit) (Matlab)	2.9	4.5	70	0.917	0.021	0.889	0.028	0.850	0.036	0.868	0.025	0.754	0.060	0.532	0.102
Linux (64 bit) (Python)	**2.1**	**2.9**	**75**	**0.925**	**0.017**	**0.914**	**0.027**	**0.861**	**0.022**	**0.879**	**0.015**	**0.811**	**0.054**	0.549	0.096
Linux (64 bit) (Matlab)	2.6	3.3	74	0.924	0.019	**0.914**	**0.027**	0.859	0.023	0.879	0.015	0.810	0.055	**0.550**	**0.095**

Table 5 Comprehensive ablation analysis on the component configuration in the DAMFNet

No.	Component Setting			DAVIS		FBMS		MCL		SegTrack-V2		DAVSOD		DAVSOD-Diff	
	DSConv	DConv	Conv2d	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
1	Default			0.835	0.044	0.823	0.047	0.790	0.055	0.748	0.067	0.733	0.062	0.354	0.153
2	✓			0.857	0.038	0.835	0.042	0.823	0.053	0.756	0.064	0.766	0.060	0.390	0.140
3		✓		0.877	0.033	0.841	0.038	0.827	0.050	0.789	0.060	0.769	0.059	0.398	0.132
4			✓	0.889	0.030	0.865	0.035	0.846	0.044	0.792	0.059	0.777	0.058	0.417	0.125
5	✓	✓		0.893	0.026	0.881	0.032	0.848	0.036	0.799	0.050	0.786	0.057	0.433	0.118
6	✓		✓	0.898	0.024	0.890	0.030	0.851	0.030	0.821	0.043	0.789	0.056	0.468	0.103
7		✓	✓	0.915	0.020	0.908	0.028	0.857	0.029	0.856	0.036	0.794	0.055	0.478	0.099
8	✓	✓	✓	**0.925**	**0.018**	**0.914**	**0.027**	**0.861**	**0.022**	**0.899**	**0.015**	**0.811**	**0.054**	**0.549**	**0.096**

Table 6 Comprehensive ablation analysis of design choice in DAMFNet

Module	Parameters			DAVIS		FBMS		MCL		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
Default	20.3	18.5	32	0.876	0.035	0.839	0.040	0.776	0.055	0.789	0.054	0.699	0.070	0.425	0.120
Conv2d	18.7	17.3	45	0.889	0.028	0.853	0.032	0.809	0.048	0.799	0.045	0.729	0.066	0.449	0.117
DConv	14.5	13.2	50	0.904	0.021	0.869	0.030	0.823	0.042	0.834	0.038	0.731	0.065	0.485	0.110
DSConv	12.9	12.8	60	0.917	0.018	0.889	0.028	0.850	0.036	0.868	0.025	0.754	0.060	0.532	0.102
All	**11.0**	**10.7**	**75**	**0.925**	**0.017**	**0.914**	**0.027**	**0.861**	**0.022**	**0.879**	**0.015**	**0.811**	**0.054**	**0.549**	**0.096**

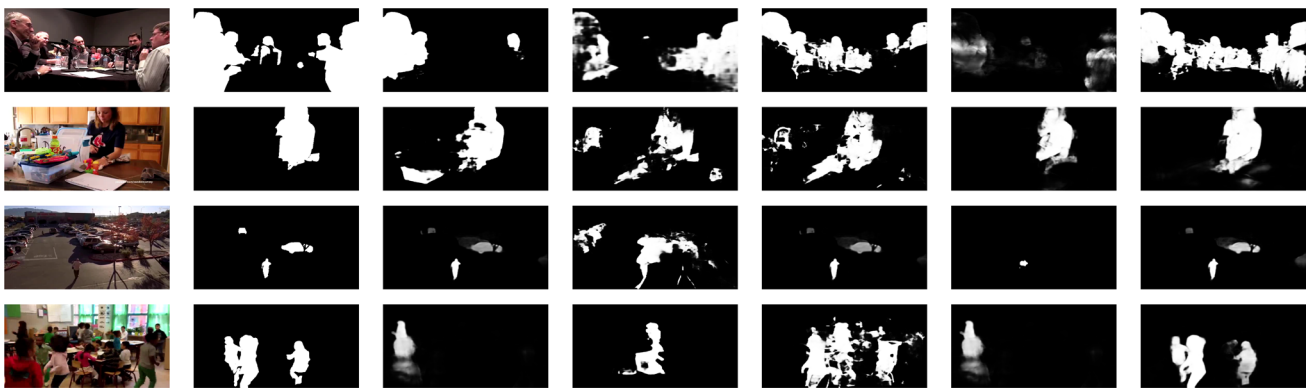


Fig. 6 The failure case of DAMFNet and SOTA models on the DAVSOD-Difficult dataset. IF is the input frame, and GT is the annotation map

extraction. By integrating these convolutional mechanisms, DAMFNet effectively balances motion and appearance modeling while maintaining network efficiency. This design allows DAMFNet to outperform existing methods on complex VSOD tasks, demonstrating superior generalization without excessive computational overhead.

5 Conclusion

This paper introduces a novel and efficient, lightweight DAMFNet model to leverage geometric appearance and motion features for rapid and effective video salient object detection. The proposed model incorporates depth-wise convolution (DSConv) layers and deformable convolution (DConv) layers to extract crucial geometric appearance and motion information. The residual skip connection between the backbone and encoder output generates the enhanced geometric representation of appearance and motion information. Then, the fusion of geometric appearance and motion information gives meaningful information. For enhancing the performance of DAMFNet, appearance and motion transfer learning (AMTL) is proposed using the AMTLoss function. Further, the Conv2d, BN followed by ReLU is used to generate the saliency map. The proposed model is validated through extensive experiments, positioning itself as a unified solution that advances research in VSOD. In future, we will propose tiny deep learning models to detect objects in multi-scale and multi-domain.

Funding This research did not receive any specific grant from 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

Ethical Approval The authors declare no affiliations or financial interests related to educational grants, consultancies, patent licensing, employment, or other equity interests. They also disclose no personal or professional relationships or knowledge impacting the material discussed in this manuscript.

References

1. Singh, H.; Verma, M.; Cheruku, R.: Dsnet: Efficient lightweight model for video salient object detection for iot and wot applications. Companion Proceedings of the ACM Web Conference **2023**, 1286–1295 (2023)
2. Cho, S.; Lee, M.; Lee, S.; Park, C.; Kim, D.; Lee, S.: 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 (2023)
3. Singh, H.; Verma, M.; Cheruku, R.: Hsnet: A novel edge-preserving hierarchical separable network for video shadow detection. Circuits, Systems, and Signal Processing pp 1–30 (2025)
4. Bi, H.B.; Lu, D.; Zhu, H.H.; Yang, L.N.; Guan, H.P.: Sta-net: spatial-temporal attention network for video salient object detection. Applied Intelligence **51**(6), 3450–3459 (2021)
5. Hu, F.; Palazzo, S.; Salanitri, F.P.; Bellitto, G.; Moradi, M.; Spampinato, C.; McGuinness, K.: 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 (2023)
6. Tang, Y.; Li, Y.; Zou, W.: Fast video salient object detection via spatiotemporal knowledge distillation. arXiv preprint arXiv:2010.10027 (2020)
7. Galab, M.K.; Taha, A.; Zayed, H.H.: Adaptive technique for brightness enhancement of automated knife detection in surveillance video with deep learning. Arabian Journal for Science and Engineering **46**(4), 4049–4058 (2021)
8. Ahmadi, M.; Ouarda, W.; Alimi, A.M.: Efficient and fast objects detection technique for intelligent video surveillance using transfer learning and fine-tuning. Arabian Journal for Science and Engineering **45**(3), 1421–1433 (2020)
9. Hu, C.; Zhu, L.: Efficient unsupervised video object segmentation network based on motion guidance. arXiv preprint arXiv:2211.05364 (2022)
10. Xu, T.; Zhao, W.; Duan, Z.: Bdfgnet: A lightweight salient object detection network based on background denoising and feature generation. Arabian Journal for Science and Engineering **49**(3), 4365–4381 (2024)
11. Alhichri, H.; Bazi, Y.; Alajlan, N.: Assisting the visually impaired in multi-object scene description using owa-based fusion of cnn models. Arabian Journal for Science and Engineering **45**(12), 10511–10527 (2020)
12. Kasmaiee, S.; Tadjfar, M.: Elliptical pressure swirl jet issuing into stagnant air. Physics of Fluids 36(7), (2024)
13. Kasmaiee, S.; Tadjfar, M.: Experimental study of the injection angle impact on the column waves: wavelength, frequency and drop size. Experimental Thermal and Fluid Science **148**, 110989 (2023)
14. Mai, L.; Niu, Y.; Liu, F.: Saliency aggregation: A data-driven approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1131–1138 (2013)
15. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2083–2090 (2013)
16. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y.: Learning to detect a salient object. IEEE Transactions on Pattern analysis and machine intelligence **33**(2), 353–367 (2010)
17. Yang, P.; Wang, Q.; Dou, J.; Dou, L.: Learning saliency-awareness siamese network for visual object tracking. Journal of Visual Communication and Image Representation **103**, 104237 (2024)
18. Liang, S.; Liu, R.; Qian, J.: Fast saliency prediction based on multi-channels activation optimization. Journal of Visual Communication and Image Representation **94**, 103831 (2023)
19. Jiang, Y.; Luo, S.; Guo, L.; Zhang, R.: Mct-vhd: Multi-modal contrastive transformer for video highlight detection. Journal of Visual Communication and Image Representation **101**, 104162 (2024)
20. Fu, K.; Gu, I.Y.H.; Yang, J.: Saliency detection by fully learning a continuous conditional random field. IEEE Transactions on Multimedia **19**(7), 1531–1544 (2017)
21. Zhou, T.; Wang, S.; Zhou, Y.; Yao, Y.; Li, J.; Shao, L.: Motion-attentive transition for zero-shot video object segmentation. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 13066–13073 (2020)

22. Zhao, X.; Liang, H.; Li, P.; Sun, G.; Zhao, D.; Liang, R.; He, X.: Motion-aware memory network for fast video salient object detection. arXiv preprint arXiv:2208.00946 (2022)
23. Miao, J.; Wei, Y.; Yang, Y.: Memory aggregation networks for efficient interactive video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10366–10375 (2020)
24. SG, A.; Karibasappa, K.; Reddy, B.E.: Video segmentation for moving object detection using local change & entropy based adaptive window thresholding. Academy & Industry Research Collaboration Center (AIRCC) pp 155–166 (2013)
25. Wen, H.; Zhou, X.; Sun, Y.; Zhang, J.; Yan, C.: Deep fusion based video saliency detection. Journal of Visual Communication and Image Representation **62**, 279–285 (2019)
26. Brostow, G.J.; Shotton, J.; Fauqueur, J.; Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10, Springer, pp 44–57 (2008)
27. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems 29 (2016)
28. Kasmaiee, S.; Tadjfar, M.: Non-circular pressure swirl nozzles injecting into stagnant air. International Journal of Multiphase Flow **175**, 104798 (2024)
29. Ji, G.P.; Fu, K.; Wu, Z.; Fan, D.P.; Shen, J.; Shao, L.: Full-duplex strategy for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4922–4933 (2021)
30. Zhang, M.; Liu, J.; Wang, Y.; Piao, Y.; Yao, S.; Ji, W.; Li, J.; Lu, H.; Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1553–1563 (2021)
31. Tian, X.; Xu, K.; Yang, X.; Du, L.; Yin, B.; Lau, R.W.: Bi-directional object-context prioritization learning for saliency ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5882–5891 (2022)
32. Sharma, G.; Singh, M.; Kumain, S.C.; Kumar, K.: Bmst-net: bidirectional multi-scale spatiotemporal network for salient object detection in videos. Signal, Image and Video Processing **19**(1), 1–9 (2025)
33. Singh, H.; Verma, M.; Cheruku, R.: Vs-net: Multiscale spatiotemporal features for lightweight video salient document detection. In: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, pp 1307–1311 (2022)
34. Huang, K.; Xu, Z.: Lightweight video salient object detection via channel-shuffle enhanced multi-modal fusion network. Multimedia Tools and Applications pp 1–15 (2023)
35. Zhang, X.; Zhou, X.; Lin, M.; Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6848–6856 (2018)
36. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773 (2017)
37. Meng, J.; Jiang, P.; Wang, J.; Wang, K.: A mobilenet-ssd model with fpn for waste detection. Journal of Electrical Engineering & Technology **17**, 1425–1431 (2022)
38. Tokmakov, P.; Alahari, K.; Schmid, C.: Learning video object segmentation with visual memory. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4481–4490 (2017)
39. Gu, Y.; Wang, L.; Wang, Z.; Liu, Y.; Cheng, M.M.; Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 10869–10876 (2020)
40. Oh, S.W.; Lee, J.Y.; Sunkavalli, K.; Kim, S.J.: Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7376–7385 (2018)
41. Cheng, J.; Tsai, Y.H.; Hung, W.C.; Wang, S.; Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7415–7424 (2018)
42. Wang, W.; Shen, J.; Lu, X.; Hoi, S.C.; Ling, H.: Paying attention to video object pattern understanding. IEEE transactions on pattern analysis and machine intelligence (2020)
43. Li, F.; Kim, T.; Humayun, A.; Tsai, D.; Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2192–2199 (2013)
44. Seo, S.; Lee, J.Y.; Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, pp 208–223 (2020)
45. Mao, A.; Yan, J.; Fang, Y.; Liu, H.: Hierarchical boundary feature alignment network for video salient object detection. Journal of Visual Communication and Image Representation p 104435 (2025)
46. Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2663–2672 (2017)
47. Voigtlaender, P.; Chai, Y.; Schroff, F.; Adam, H.; Leibe, B.; Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9481–9490 (2019)
48. Min, D.; Zhang, C.; Lu, Y.; Fu, K.; Zhao, Q.: Mutual-guidance transformer-embedding network for video salient object detection. IEEE Signal Processing Letters (2022)
49. Wang, W.; Shen, J.; Xie, J.; Porikli, F.: Super-trajectory for video segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1671–1679 (2017)
50. Perazzi, F.; Wang, O.; Gross, M.; Sorkine-Hornung, A.: Fully connected object proposals for video segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 3227–3234 (2015)
51. Xu, Y.; Song, D.; Hoogs, A.: An efficient online hierarchical supervoxel segmentation algorithm for time-critical applications. In: BMVC, Citeseer, p 12 (2014)
52. Wang, W.; Song, H.; Zhao, S.; Shen, J.; Zhao, S.; Hoi, S.C.; Ling, H.: Learning unsupervised video object segmentation through visual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3064–3074 (2019)
53. Ballas, N.; Yao, L.; Pal, C.; Courville, A.: Delving deeper into convolutional networks for learning video representations. arXiv preprint arXiv:1511.06432 (2015)
54. Siam, M.; Jiang, C.; Lu, S.; Petrich, L.; Gamal, M.; Elhoseiny, M.; Jagersand, M.: Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: 2019 International Conference on Robotics and Automation (ICRA), IEEE, pp 50–56 (2019)
55. Chen, Z.; Guo, C.; Lai, J.; Xie, X.: Motion-appearance interactive encoding for object segmentation in unconstrained videos. IEEE Transactions on Circuits and Systems for Video Technology **30**(6), 1613–1624 (2019)
56. Fan, D.P.; Wang, W.; Cheng, M.M.; Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8554–8564 (2019)
57. Han, Z.; Hu, S.; Song, H.; Zhang, K.: Open-vocabulary saliency-guided progressive refinement network for unsupervised video

object segmentation. In: ICASSP 2025–2025 IEEE International Conference on Acoustics, pp. 1–5. IEEE, Speech and Signal Processing (ICASSP) (2025)
58. Singh, H.; Verma, M.; Cheruku, R.: Novel dilated separable convolution networks for efficient video salient object detection in the wild. IEEE Transactions on Instrumentation and Measurement (2023)
59. Zhang, Q.; Wang, S.; Wang, X.; Sun, Z.; Kwong, S.; Jiang, J.: Geometry auxiliary salient object detection for light fields via graph neural networks. IEEE Transactions on Image Processing **30**, 7578–7592 (2021)
60. Zhang, Z.; Gao, P.; Peng, S.; Duan, C.; Zhang, P.: Enhanced point feature network for point cloud salient object detection. IEEE Signal Processing Letters (2023)
61. Liu, X.; Wang, L.: Msrmnet: Multi-scale skip residual and multi-mixed features network for salient object detection. Neural Networks **173**, 106144 (2024)
62. Peng, D.; Zhou, W.; Pan, J.; Wang, D.: Msednet: Multi-scale fusion and edge-supervised network for rgb-t salient object detection. Neural Networks **171**, 410–422 (2024)
63. Singh, H.; Verma, M.; Cheruku, R.: Dmfnet: geometric multi-scale pixel-level contrastive learning for video salient object detection. International Journal of Multimedia Information Retrieval **14**(2), 12 (2025)
64. Zhu, X.; Hu, H.; Lin, S.; Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9308–9316 (2019)
65. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al.: 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 (2023)
66. Deng, J.; Dong, S.;, Chen, L.; Hu, J.; Zhuo, C.: Stdf: Spatio-temporal deformable fusion for video quality enhancement on embedded platforms. ACM Transactions on Embedded Computing Systems (2024)
67. Singh, H.; Verma, M.; Cheruku, R.: Dsfnet: video salient object detection using a novel lightweight deformable separable fusion network. IEEE Transactions on Instrumentation and Measurement (2024)
68. Wang, Z.; Zhong, Y.; Miao, Y.; Ma, L.; Specia, L.: Contrastive video-language learning with fine-grained frame sampling. arXiv preprint arXiv:2210.05039 (2022)
69. Gao, S.H.; Han, Q.; Li, Z.Y.; Peng, P.; Wang, L.; Cheng, M.M.: Global2local: Efficient structure search for video action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16805–16814 (2021)
70. Kong, Y.; Wang, Y.; Li, A.; Huang, Q.: Self-sufficient feature enhancing networks for video salient object detection. IEEE Transactions on Multimedia (2021)
71. Cheng, M.M.; Gao, S.H.; Borji, A.; Tan, Y.Q.; Lin, Z.; Wang, M.: A highly efficient model to study the semantics of salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8006–8021 (2021)
72. Su, Y.; Deng, J.; Sun, R.; Lin, G.; Wu, Q.: A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. arXiv preprint arXiv:2203.04708 (2022)
73. Xu, B.; Jiang, Q.; Zhao, X.; Lu, C.; Liang, H.; Liang, R.: Multidimensional exploration of segment anything model for weakly supervised video salient object detection. IEEE Transactions on Circuits and Systems for Video Technology (2024)
74. Zhao, X.; Liang, H.; Li, P.; Sun, G.; Zhao, D.; Liang, R.; He, X.: Motion-aware memory network for fast video salient object detection. IEEE Transactions on Image Processing (2024)
75. Huang, K.; Xu, Z.: Lightweight video salient object detection via channel-shuffle enhanced multi-modal fusion network. Multimedia Tools and Applications **83**(1), 1025–1039 (2024)
76. Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
77. Lu, F.; Tong, Q.; Jiang, X.; Feng, Z.; Xu, J.; Huo, J.: Deep multilayer sparse regularization time-varying transfer learning networks with dynamic kullback–leibler divergence weights for mechanical fault diagnosis. IEEE Transactions on Industrial Informatics (2024)
78. Lei, T.; Wang, R.; Zhang, Y.; Wan, Y.; Liu, C.; Nandi, A.K.: Defednet: Deformable encoder-decoder network for liver and liver tumor segmentation. IEEE Transactions on Radiation and Plasma Medical Sciences **6**(1), 68–78 (2021)
79. Chen, C.; Wang, G.; Peng, C.; Fang, Y.; Zhang, D.; Qin, H.: Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. IEEE Transactions on Image Processing **30**, 3995–4007 (2021)
80. Liu, Y.; Duanmu, M.; Huo, Z.; Qi, H.; Chen, Z.; Li, L.; Zhang, Q.: Exploring multi-scale deformable context and channel-wise attention for salient object detection. Neurocomputing **428**, 92–103 (2021)
81. Chen, Y.W.; Jin, X.; Shen, X.; Yang, M.H.: 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 (2022)
82. Pei, G.; Yao, Y.; Shen, F.; Huang, D.; Huang, X.; Shen, H.T.: Hierarchical co-attention propagation network for zero-shot video object segmentation. IEEE Transactions on Image Processing (2023)
83. Lee, M.; Cho, S.; Lee, S.; Park, C.; Lee, S.: Unsupervised video object segmentation via prototype memory network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 5924–5934 (2023)
84. Ponimatkin, G.; Samet, N.; Xiao, Y.; Du, Y.; Marlet, R.; Lepetit, V.: 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 (2023)
85. Zy, Liu; Jw, Liu: Part-aware attention correctness for video salient object detection. Engineering Applications of Artificial Intelligence **119**, 105733 (2023)
86. Liu, N.; Nan, K.; Zhao, W.; Yao, X.; Han, J.: Learning complementary spatial–temporal transformer for video salient object detection. IEEE Transactions on Neural Networks and Learning Systems (2023)
87. Li, P.; Zhang, Y.; Yuan, L.; Xiao, H.; Lin, B.; Xu, X.: Efficient long-short temporal attention network for unsupervised video object segmentation. Pattern Recognition **146**, 110078 (2024)
88. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7479–7489 (2019)
89. Wei, J.; Wang, S.; Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. Proceedings of the AAAI conference on artificial intelligence **34**, 12321–12328 (2020)
90. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 724–732 (2016)
91. Wang, H.; Gan, W.;, Hu, S.; Lin, J.Y.; Jin, L.; Song, L.; Wang, P.; Katsavounidis, I.; Aaron, A.; Kuo, C.C.J.: Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In: 2016 IEEE international conference on image processing (ICIP), IEEE, pp 1509–1513 (2016)

92. Brox, T.; Malik, J.; Ochs, P.: Freiburg-berkeley motion segmentation dataset (fbms-59). In: European Conference on Computer Vision (ECCV), vol 1, p 9 (2010)
93. Fan, D.P.; Wang, W.; Cheng, M.M.; Shen, J.: Shifting more attention to video salient object detection. In: IEEE CVPR, p 308 (2019)