DSFNet: Video Salient Object Detection Using a Novel Lightweight Deformable Separable Fusion Network

Hemraj Singh, Mridula Verma, *Member, IEEE*, and Ramalingaswamy Cheruku, *Senior Member, IEEE*

***Abstract*— Geometric variations of spatial and temporal features of objects in video streams cause great difficulty in video salient object detection (VSOD) tasks. Most existing deep-learning methods utilize fixed-sized kernels, which limits the receptive field (RF) to extract the local and global features and fails to understand the visual semantics of the deformed objects' foreground and background. Moreover, due to their complex architectures, these methods need more computational resources, which limits their deployment in real-world scenarios. To address the aforementioned challenges and to make a balance between performance and computational complexity, a deformable separable fusion network (DSFNet) is proposed, which extracts the geometric spatiotemporal variations at multiscale features dynamically without compromising the network's complexity. A Swarm-Enhanced Adam (SEAdam) optimizer has been proposed to adaptively balance the exploration and exploitation of gradients locally and globally and improve the convergence speed. This is the first work that extracts the multiscale geometric local and global context-based visual information. With the help of extensive experimentation on six benchmark highly challenging datasets, we show that the proposed model outperforms state-of-the-art models in terms of the number of parameters, floating-point operations (FLOPs), and latency.**

***Index Terms*— Deformable attention fusion, deformable global attention (GA), deformable region context, deformable separable fusion network (DSFNet), spatial and temporal features, video salient object detection (VSOD).**

I. Introduction

VIDEO salient object detection (VSOD) is a fundamental task in computer vision that aims to detect and segment objects capturing human visual attention in multiple applications such as autonomous cars [1], medical image segmentation [2], smart surveillance system [3], smart home [4], smart agriculture [5], and many more. In this work, the most popular unsupervised VSOD [4], [6], [7] studies are explored for automatically detecting and segmenting target objects without giving any annotation mask as visual guidance across video sequences. In the real scenario, the target objects dynamically change their geometric appearance in the video sequences, creating confusion with similar instances from the same category, which poses significant challenges in this task. A few examples of geometric variations include scale changes, pose alterations, rapid motion, truncation, blurry effects, occlusions, and more. Deep VSOD methods analyze the spatial and temporal dependencies among the sequence of frames in a video and extract effective appearance and motion features. The majority of these methods are based on multistream networks [6], [8], [9], multiscale encoder–decoder architecture [10], [11], [12], global-to-local search strategies [13], [14], [15], recurrent neural networks (RNNs) and ConvLSTM-based methods [16], [17], [18], recent approaches based on deformable convolution-based networks [19], [20], [21], [22], [23], or hybrid approaches, for example, multiscale deformable convolution networks [24], [25] or deformable encoder–decoder strategies [19], [26]. To capture temporal information, optical flow is often utilized [4], [6], [27] to convey motion cues. However, the efficacy of these methods is heavily dependent on the accuracy of flow computation and is susceptible to scenarios where foreground objects exhibit minimal motion.

A few other limitations of most of these models are the loss of spatial context because of pooling and strided-convolution operations [26], loss of small objects due to many downsampling operations, the aliasing influence due to many upsampling [28] or unadaptability toward long/short receptive fields (RFs) [29]. Several solution approaches available, however, are applied for SOD on images [24], [25], [28] or suffer from increased computational complexity and decreased inference speed due to stack of convolution blocks or larger spatial filters [15], [22], [23], [24], [26]. Moreover, they come with a significant computational cost associated with modeling temporal information and hence are not feasible to be applied in resource-constrained environments [30], [31].

Recently, many lightweight SOD models [1], [32], [33] are proposed, which mainly focus on employing lightweight backbone networks, such as MobileNets [30] and EfficientNet [8] to capture salient objects. Other works focus on knowledge distillation [18], [33], quantization [1], [32], and model pruning [34], [35] to capture salient objects using lesser parameters. Maintaining a higher detection performance, however, is still an open challenge. To balance the performance and number of network parameters, we design a novel lightweight deformable separable fusion network (DSFNet), which can extract efficient

Received 11 March 2024; revised 15 June 2024; accepted 24 June 2024. Date of publication 30 September 2024; date of current version 14 October 2024. The Associate Editor coordinating the review process was Dr. Damodar Reddy Edla. *(Corresponding author: Ramalingaswamy Cheruku.)*

Hemraj Singh and Ramalingaswamy Cheruku are with the Department of Computer Science and Engineering, National Institute of Technology Warangal, Hanamkonda, Telangana 506004, India (e-mail: 720079@student.nitw.ac.in; rmlswamy@nitw.ac.in).

Mridula Verma is with the Institute for Development and Research in Banking Technology, Hyderabad, Telangana 500057, India (e-mail: vmridula@idrbt.ac.in).

Digital Object Identifier 10.1109/TIM.2024.3470045

multiscale local and global context geometric information without escalating network parameters, making it suitable for IoT applications. This is the first work where multicontext-based, multiscale features are extracted and fused using a multiscale deformable and separable fusion network in videos.

The existing weighted momentum particle swarm optimizers (WMPSO) [36], [37] compute momentum and weighted to explore and exploit the search space while adjusting their positions based on the personal best and global best solution. However, these methods converge prematurely to suboptimal solutions, especially if the momentum factor is too high or the weighting scheme is not appropriately designed due to getting stuck in local optima [38], [39]. To solve the local optima problem, Adam [36], [40], [41] is used to explore the search space adaptively using adaptive learning rate and momentum estimation. However, the adaptive nature of Adam often leads to over-adaption when the learning rates become too small and cause slower convergence or getting stuck in poor local optima. To overcome the above problems, the SEAdam is proposed, which adaptively balances the exploration and exploitation of gradients locally and globally and improves the convergence speed. The main contributions of our proposed model are listed as follows.

1) A novel lightweight DSFNet is proposed, which extracts attention-based multiscale local and global geometric variations from video streams with the help of the following newly designed modules.
   a) Deformation attention fusion module (DAFM) to extract the attention-based multiscale geometric spatiotemporal features.
   b) Deformable global attention module (DGAM) to extract geometric variation of spatial and temporal information and the fused cross-modal context.
   c) Deformable separable receptive field blocks (DSRFB) to extract the multiscale geometric information.
   d) Deformable region context module (DRCM) to fuse geometric variations of spatiotemporal local and global attention (GA) features.
   e) Local Feature Indicator (LFI) to extract and preserve the lightweight local context information.
2) For balancing the vanishing gradient and well exploitation of solution space, a novel Swarm-Enhanced Gradient-based Adam (SEAdam) optimizer is proposed.
3) With the help of extensive experimentation on six VSOD benchmark datasets, we illustrate that the proposed DSFNet model achieves SOTA performance on four benchmark datasets with a majority of metrics.

II. Related Work

The unsupervised VSOD goals are to detect and segment the object without any prior information or annotation mask in the video. Most of the papers [4], [6], [7], [42] tackled the VSOD task in an unsupervised way, which uses the attention-aware features learning techniques from a video clip. These methods have major drawbacks in that they are unable to balance the spatial and temporal feature consistency from previous

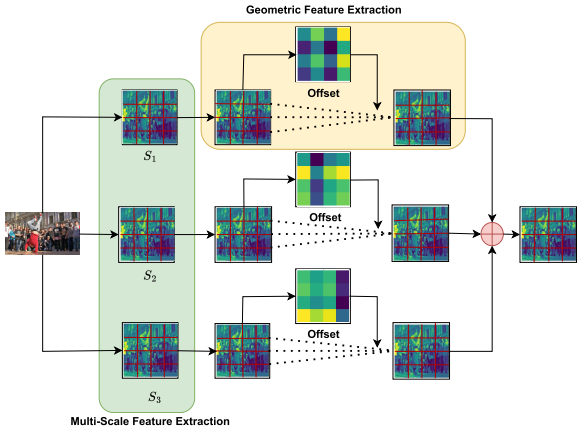


Fig. 1. Visualization of proposed multiscale geometric spatial and temporal features extraction techniques via DSFNet, where S_1, S_2, and S_3 are the three different scales.

and current features and are unreliable for correct detection. Zhang et al. [43] presented a Wasserstein distance function to design a short-distant Sinkhorn layer and a long-distant Sinkhorn layer, which computes the global optical flow using a transport layer from one modality to another. Cho et al. [4] designed a two-stream motion-as-option network that uses the motion causes optionally using a collaborative learning strategy and reduces the motion dependency problems. Furthermore, Lee et al. [7] designed a prototype memory network (PMNet) to extract the effective RGB and motion information based on the superpixel component. For accurate detection, the self-learning method is used to enhance performance. Zhang et al. [44] proposed a dynamic context-sensitive filtering network (DCFNet) to extract spatiotemporal-based location-related similarity features from consecutive frames.

Most of the existing multiscale feature extraction works [3], [11], [28], [45], [46], [47] have been addressed using the multiencoder, ASPP, and different kernels with different dilation rates. Singh et al. [3] used the ASPP module to extract the multiscale information. Zhang et al. [47] introduced a multiscale information enhancement (MIE) module to enrich shared information while integrating supplementary details. This module facilitates extracting RGB features across various scales and their conversion into point features. Zhang et al. [45] used multiscale graph neural networks to effectively investigate spatial and disparity correlations across multiple views, facilitating a deeper comprehension of light field content and the generation of saliency features that are both more representative and discriminative. Ponimatkin et al. [12] designed an object function that minimizes the wrong prediction and scales the right resolution without increasing the complexity of the network. Liu and Wang [28] proposed a multiscale deformation module (MSDM) to capture diverse visual cues across different scales and adapt to the varying shapes of salient objects. Peng et al. [11] employed a multiscale encoder–decoder network to acquire multiscale semantic features for estimating saliency, particularly foreground extraction. However, these SOTA methods require more computational complexity and leakage of information at each step. To overcome these problems, we introduce DSRFB blocks, which require less computation complexity than SOTA

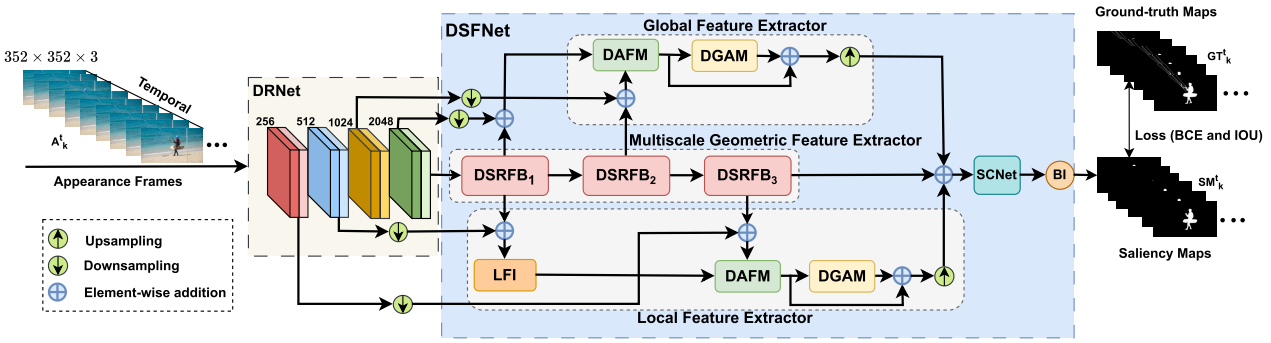


Fig. 2. Architecture of the proposed DSFNet. Where DRNet is the deformable residual network, DSRFB is the deformable separable receptive field blocks, DAFM is the deformable attention fusion module, DGAM is the deformable global attention module, UP is the upsampling, SCNet is the separable convolution network, and BI is a bilinear interpolation.

methods due to the use of deformable and depth-wise separable convolution for fusing the multiscale and geometric information.

The deformable convolution-based approaches aim to extract the geometric spatial structure of objects adaptively. The traditional convolution neural network follows the fixed kernel structure and fails to extract geometric spatial information from the objects. So, to overcome this problem, Zhu et al. [21] came up with Deformable Convnets v2 with an extra modulation mechanism, which enhances the modeling power in the network at region levels. Still, it fails to handle long-range dependency on spatial and temporal information. Next, to overcome this problem, Wang et al. [22] designed the InternImage ViT-based technique, which generates a large effective RF for detection and segmentation tasks. Furthermore, Deng et al. [23] designed spatiotemporal deformable convolution (STDC) to extract the motion information and fuse it effectively. Singh et al. [19] designed a DSNet that extracts attention-based spatial and temporal information without increasing the model parameter.

To remedy the problems raised by the heavy-weighted models, Cheng et al. [48] proposed an extremely lightweight holistic model trained from scratch to extract the spatial and temporal features. This model takes more training time and cannot detect clutter background and deformation scenes. Hu et al. [18] proposed a lightweight model that utilizes multiple heterogeneous decoders using 3-D convolution to improve accuracy, but it does not consider the training time and inference time. It is unable to tackle the correlation features of deformed objects/backgrounds. Singh et al. [10] proposed a VS-Net lightweight model to detect the salient document using multiscale spatiotemporal features. Still, it is not as effective due to the long dependency of the features. Furthermore, to handle this problem, Hu and Zhu [8] proposed a dual-stream network to extract the appearance and motion representations to detect the objects but is unable to overcome the burden of the feature sparse matrix.

III. Proposed Method

A. Motivation

The existing multiscale spatial and temporal-based models [11], [28], [49], [50] require significant computational resources, which lead to slower inference speeds and increased hardware requirements. The deformable convolution-based methods [19], [20], [21], [22] are popular for extracting the geometric spatiotemporal features at low computational costs; however, they are unable to generalize the performance on unseen data. To overcome the challenges of both approaches, we combine multiscale and deformable concepts as shown in Fig. 1. The benefit of combining both the concepts of multiscale and deformable is that the resulting extracted features have the potential to improve accuracy and robustness by dynamically capturing the diverse appearance and motion patterns of objects (such as occlusions, scale variations, and rapid motion) in videos.

B. Overview of Architecture

Consider a dataset having T video clips with k consecutive frames (where $k = 1, 2, \ldots, T$). The proposed DSFNet takes as inputs the appearance frames denoted as $A_k \in \mathbb{R}^{H\times W\times C}$ (where H is the height, W is the width, and C is the channels of the appearance frame) and generates the saliency maps ($\mathrm{SM}_k = \mathrm{SM}_1, \mathrm{SM}_2, \ldots, \mathrm{SM}_k$). As shown in Fig. 2, first, the appearance frames A_k are passed to the backbone Deformable Residual Network (DRNet), which has four deformable residual blocks, $\mathrm{DR}_1, \mathrm{DR}_2, \mathrm{DR}_3$, and DR_4, to adaptively incorporate the spatial and temporal information and generates backbone features from the input video clips at different dimensions 256, 512, 1024, and 2048. This DRNet is designed by utilizing the deformable convolution [21] with the backbone ResNet-50 [51].

The proposed DSFNet extracts the multicontext, multiscale geometric spatial and temporal information in three ways: 1) the multiscale global contextual spatial features are extracted from DRNet blocks outputs, and temporal features are extracted from consecutive feature maps using the DAFM and DGAM, which enhances the backbone geometric features coming from DRNet; 2) the local contextual features are extracted from the LFI module along with geometric features. This local context is fused with the multiscale global context to obtain enhanced local–global contextual features; and 3) the DSRF blocks extract lightweight multiscale geometric spatial and temporal dependency information. A SEAdam optimizer, which minimizes total weighted loss (BCE and IoU loss

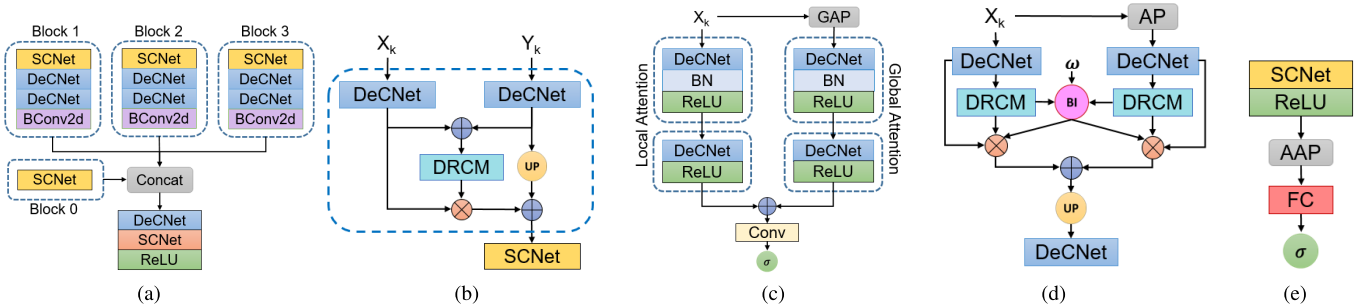


Fig. 3. Configurations of (a) DSRFB, (b) DAFM, (c) DRCM, (d) DGAM, and (e) LFI modules. $\otimes$ is the element-wise multiplication operation and $\oplus$ is the element-wise addition operation. GAP: Global Avg. Pooling, AP: Avg. pooling, AAP: Adaptive Avg. Pooling, BN: batch normalization, σ: a Sigmoid operation, ω: learnable parameters, BI: binary interpolation operation, UP: Upsampling operation, FC: fully connected layer, and Concat: element-wise concatenation operation.

function) based on the optimal feature maps, is proposed to update the network weight parameter. Finally, the separable convolution network (SCNet) [19], [30] followed by Binary Interpolation (BI) is used to convert high-level multicontext multiscale geometric semantic information to low-level local and global information and produce the saliency maps. The detailed explanations of each of these components are provided in the upcoming subsections.

C. Deformable Separable Receptive Field Block

A DSRFB module is proposed, which is an updated version of the RF block [52] to extract the multiscale geometric local and global spatiotemporal features. It is designed using DeCNet, which extracts the geometric spatial and temporal multiscale information, and SCNet to reduce the network parameters and floating-point operation (FLOP) and increase the inference speed. DSRFB enhances the feature discriminability and robustness and establishes the correlation between the eccentricity and size of RFs. As illustrated in Fig. 3(a), the four blocks extract geometric variations of multiscale spatial and temporal features at different dilation rates, which are then fused together using element-wise concatenation. Furthermore, with the help of DeCNet and SCNet, spatial- and temporal-based variations changes in feature maps are accommodated by dynamically adjusting the kernels. At last, the feature quality is normalized using a nonlinear ReLU activation function.

D. Deformable Attention Fusion Module (DAFM)

The DAFM is designed using a DeCNet, SCNet, and DRCM, as shown in Fig. 3(b). It takes fused multiscale spatial–temporal features X_k^{ms} and Y_k as inputs, coming from DSRF blocks and layers of DRNet, and with the help of two DeCNet geometric features are extracted. These fused features are passed to the DRCM, which generates the global and local context-based attention features. The element-wise multiplication operation is performed between geometric- and attention-based global contextual features and fused with upsampled Y_k^{ms} features. At last, SCNet generates compressed generalized spatial and temporal features.

1) Deformable Region Context Module: The DRCM module, shown in Fig. 3(c), extracts geometric multiscale contextual attention features using the deformable convolution layer, which helps to handle the geometric variation of region context information dynamically and detect the salient objects, which are surrounded by background and similar objects with the same color and shape. The geometric multiscale contextual attention information is extracted in two ways: 1) local attention (LA), which is generated by the attention mask from low-level to high-level information by capturing context areas and 2) GA, which is generated by the attention mask from high-level to low-level information with resolution $64 \times 64 \times 3$. Furthermore, LA and GA outputs are fused to persistently preserve the quality of features. The fused attention-based global geometric features are passed to the *Conv* layer with 1×1 filters to generate the attention mask. Finally, the Sigmoid (σ) operation is normalized to the feature maps.

E. Deformable Global Attention Module

The DGAM [shown in Fig. 3(d)] goals to extract more precise representations of the attention-based discriminative local and global information while diminishing background noise and amplifying saliency information. The first DeCNet receives X_k^a directly, whereas the second one receives input after average pooling (AP) operation, which calculates and enhances the temporal information on X_k^a. The outputs of these two DeCNets are passed to the two DRCMs, respectively, which provide input to bilinear interpolation (BI) along with learnable parameters (ω). Furthermore, both the consistent attention-based geometric features are fused after applying upsampling, followed by DeCNet.

F. LFI Modules

In contrast to the existing LFI methods [53], [54], which fail to capture the geometric variation of complex spatial relationships within local regions, an LFI module is designed. With the indicative local features extracted by LFI, the proposed network can handle challenging scenarios such as motion blur, deformation, clutter scene, and illumination. The configuration of the LFI module is shown in Fig. 3(e). The fusion of lower scale features and deformed backbone features is passed to

Algorithm 1 Deformable Separable Fusion Network

Input: A_k: Appearance frames and GT_k: Ground-truth maps

Output: SM_k: Saliency maps.

1 A_k is passed to DRNet to generate backbone spatial and temporal feature map X_k.
2 The local context multiscale spatial and temporal feature X_k^l is extracted by LFI, DAFM, DGAM, and upsampling (UP) in the bottom part using the backbone spatial and temporal feature.
3 The global context multiscale spatial and temporal feature X_k^g is extracted by DAFM, DGAM, and UP in the upper part.
4 A stack of DSRFB modules is used to extract the multiscale geometric spatial and temporal information X_k^{ms} in the middle part.
5 All the extracted features (X_k^l, X_k^g, X_k^{ms}) are fused together using element-wise addition operation and produce multicontext multiscale geometric spatial and temporal features X_k^f.
6 A SEAdam optimizer updates the network weight parameters and minimizes total loss (weighted BCE and IoU loss) using *Eq.* 9.
7 At last, the saliency map is generated using SCNet and Bi-linear Interpolation operation (BI) on the fused multicontext multiscale geometric spatial and temporal features X_k^f.

the AAP to extract the local variation of geometric features and store in X_k with dimension 64. The X_k features are passed to SCNet, which uses a 3×3 filter for depth-wise separable and 1×1 filter for point-wise separable convolution to reduce the multiplication and addition operations. Further nonlinear ReLU activation followed by the FC layer and Sigmoid σ operation generates the generalized local spatial- and temporal-based geometric features.

G. Contextual Feature Extraction

1) Geometric Backbone Spatial and Temporal Feature Extraction: The geometric backbone spatial and temporal features are extracted using DRNet blocks (DR_i, $i = 1, 2, 3, 4$) using appearance frames information and generate the backbone features maps (X_k^{en}, $k = 1, 2, 3, 4$). The process is given as follows:

$$X_k^{en} = \text{DRNet}(A_k). \tag{1}$$

2) Local-Context Spatial and Temporal Feature Extraction: The local context is extracted using the DSRFB, LFI, DAFM, DGAM, and bilinear upsampling operation (UP). LFI locates the local context feature with the help of adaptive average pooling (AAP). The process is given as follows:

$$\begin{aligned} X_{k_1}^l &= \text{LFI}\big(X_{k_2} \oplus \text{DSRFB}_1\big(X_k^{en}\big)\big) \\ X_{k_2}^l &= X_{k_1} \oplus \text{DSRFB}_3\big(\text{DSRFB}_1\big(X_k^{en}\big)\big) \\ X_k^{lc} &= \text{DAFM}\big(X_{k_1}^l, X_{k_2}^l\big) \\ X_k^{lc} &= \text{UP}\big(X_k^{lc} \oplus \text{DGAM}\big(X_k^{lc}\big)\big). \end{aligned} \tag{2}$$

Here, X_k^{lc} is multiscale local-context geometric spatial and temporal features, and $X_{k_1}^{en}$ and $X_{k_2}^{en}$ are the DRNet blocks DR_1 and DR_2 outputs, respectively.

3) Global-Context Spatial and Temporal Feature Extraction: The global contextual features, which help in finding the coarse RF, are extracted using DSRFBs, DAFM, and DGAM with one bilinear UP. SCNet is used to reduce the network parameters. The process is given as follows:

$$\begin{aligned} X_{k_1}^g &= X_{k_4}^{en} \oplus \text{DSRFB}_1\big(X_{k_4}^{en}\big) \\ X_{k_2}^g &= X_{k_3}^{en} \oplus \text{DSRFB}_2\big(\text{DSRFB}_1\big(X_{k_4}^{en}\big)\big) \\ X_k^g &= \text{DAFM}\big(X_{k_1}^g, X_{k_2}^g\big) \\ X_k^g &= \text{UP}\big(X_k^g \oplus \text{DGAM}\big(X_k^g\big)\big). \end{aligned} \tag{3}$$

Here, $X_{k_3}^{en}$ and $X_{k_4}^{en}$ are the DRNet blocks DR_3 and DR_4 outputs.

4) Multiscale Geometric Spatial and Temporal Feature Extraction: The last block output $X_{k_4}^{en}$ of DRNet is used as the input to the DSRFB modules (DSRFB_i, $i = 1, 2, 3$) to extract the geometric multiscale spatial and temporal information X_k^{ms}. The DSRFB module helps to capture different attention responses adaptively at different resolutions. The process is given as follows:

$$X_{k_4}^{ms} = \text{DSRFB}_3\big(\text{DSRFB}_2\big(\text{DSRFB}_1\big(X_{k_4}^{en}\big)\big)\big) \tag{4}$$

where X_k^{ms} is the multiscale geometric spatial and temporal features. These features are further fused with local and global context as given as follows:

$$X_k^f = \big(X_k^{ms} \oplus X_k^l \oplus X_k^g\big). \tag{5}$$

At last, X_k^f is used to generate the saliency maps SM_k using the SCNet with 1×1 filter followed by a bilinear interpolation operation (BI). It extracts the multiscale geometric spatial and temporal information and converts strong high-level multicontext multiscale geometric information to low-level weak multicontext multiscale geometric information. Furthermore, the BI operation is used to normalize the feature quality between 0 and 1 and generate the saliency map (SM_k). The process is given as follows:

$$\text{SM}_k = \text{BI}\Big(\text{SCNet}\Big(X_k^f\Big)\Big). \tag{6}$$

The generated saliency maps SM_k are efficient representations of the foreground.

H. Swarm-Enhanced Adam (SEAdam) Optimizer

The existing WMPSO [36], [37] computed momentum and weighted average contribute more toward exploration and exploitation of the search space while adjusting their positions based on the personal best and global best solutions. However, this exploration and exploitation of search space might converge prematurely to suboptimal solutions, especially if the momentum factor is too high or the weighting scheme is not appropriately designed. This limits its ability to explore the search space effectively and get stuck in local optima. To solve the local optima problem, Adam [36], [40], [41] is used, which explores the search space adaptively using adaptive learning rate and momentum estimation. Sometimes,

Algorithm 2 SEAdam Algorithm

Input: η: Learning rate; $\beta, \beta_1, \beta_2 \in [0, 1)$: momentum factor and decay rates; effective step size ϵ: 10^{-8}
F: fitness function; P_i: ith particles;
w_0: initial weight vector; t = 0 (iteration);
for *each particle* $i \in S$ **do**
  initialize particles and evaluate their fitness using F
while w_t *not converged* **do**
  t = t+1;
  for *each particle* $P_i \in S$ **do**
    Apply Eq. 7 and 8 and evaluate its fitness using F
  $g_t = \nabla_w F_{t+1}(\mathrm{w}_t)$ (Get gradient g_t w.r.t fitness function at time t)
  $m^t = \beta_1 \cdot m^{t-1} + (1 - \beta_1) \cdot g_t$;
  $v^t = \beta_2 \cdot v^{t-1} + (1 - \beta_2) \cdot (g_t^2)$;
  Bias Correction
  $\hat{m}^t = \frac{m^t}{(1-\beta_1^t)}$, $\hat{v}^t = \frac{v^t}{(1-\beta_2^t)}$
  Update
  $\mathrm{w}_t = \mathrm{w}_{t-1} - \eta \cdot \frac{\hat{m}^t}{(\sqrt{\hat{v}^t}+\epsilon)}$
  Find the p_i^{best} and g^{best}
return w_t

the adaptive nature of Adam leads to over-adaption problems when the learning rates become too small. This issue can manifest as slower convergence or getting stuck in poor local optima. To overcome both the Adam and WMPSO optimizers, an SEAdam optimizer is proposed, a combination of the Adam [40] optimizer and WMPSO. In SEAdam, the computed momentum-based weighted average of WMPSO is adaptively updated using Adam's adaptive learning rates and stochastic estimation. The adaptive learning rates allow SEAdam to navigate regions of the optimization landscape with varying gradients, while stochastic estimation injects randomness into the updates, enabling the exploration of new directions. Through these mechanisms, SEAdam increases the likelihood of escaping local minima and converging toward better solutions.

Let v_i^{t+1} denote the velocity of the ith particle at iteration $t + 1$, c_1 and c_2 are the cognitive and social learning parameters, respectively, r_1 and r_2 denote random values between [0, 1), p_i^{best} is the personal best position of the ith particle, β is the momentum factor, S is the swarm size, and g^{best} is the global best position observed in swarm S

$$v_i^{(t+1)} = (1-\beta)\left[v_i^{(t)} + c_1 r_1\left(p_i^{\text{best}} - x_i^{(t)}\right) + c_2 r_2\left(g^{\text{best}} - x_i^{(t)}\right)\right] + \beta v_i^{t-1}. \tag{7}$$

Next, this velocity is used to compute the position of particles using as follows:

$$x_i^{(t+1)} = x_i^t + v_i^{(t+1)} \tag{8}$$

where x_i^t denotes the position of the ith particle at iteration t.

In the SEAdam optimizer, first, various particle velocities and positions (model weights) are initialized randomly, and these are updated using (7) and (8). Next, the gradient of weights is calculated and updated according to the Adam optimizer. Finally, these weights are updated using p_i^{best} and g^{best}. This process is shown in Algorithm 2. During training, the SEAdam optimizer minimizes the structure loss function (weighted BCE and IoU loss). The loss function is shown as follows:

$$F(\mathrm{SM}_k, \mathrm{GT}_k) = W_1 * \mathrm{BCE}(\mathrm{SM}_k, \mathrm{GT}_k) + W_2 * \mathrm{IoU}(\mathrm{SM}_k, \mathrm{GT}_k) \tag{9}$$

where W_1 and W_2 are the learnable weights, SM_k is the predicted saliency maps, and GT_k is ground-truth maps.

IV. Experiments and Result Analysis

The implementation of the proposed DSFNet model is carried out on a 64-bit Ubuntu 18.04 system equipped with 32 GB RAM, 1 TB hard disk storage memory, and a 16 GB P5000 NVIDIA GPU. The system runs Anaconda 3.8 and PyTorch version 1.13.0 with CUDA 11.6 facilitated by NVIDIA driver 470 to leverage the capabilities of the NVIDIA P5000 GPU. For saliency map generation, OpenCV2.0 with Imageio v2.0 is employed. The input frame resolution is resized with 352 × 352. The computational cost of DSFNet is evaluated and compared based on the number of network parameters (#Param) in a million (M), the number of FLOPs in gigabytes (G), and the inference speed in frames per second (FPS). DSFNet is tested on six benchmark VSOD datasets: 1) DAVIS [57] is widely recognized, comprising 50 video clips (30 for training and 20 for testing); 2) MCL [58] consists of nine testing videos; 3) FBMS [9] includes 59 videos, 29 clips utilized for training, and 30 clips utilized for testing; 4) SegTrack-V2 [6] features 13 testing video clips; 5) DAVSOD [59] has 142 video clips; 61 are allocated for training and 81 for testing; and 6) DAVSOD-Difficult [59] comprises 20 testing video clips. These datasets collectively cover diverse scenarios, allowing for a comprehensive evaluation of the DSFNet across different challenges in VSOD.

A. Training and Testing Performance

For training, a mini-batch of videos is randomly selected with corresponding GT frames. The training is divided into two parts: 1) unsupervised pretraining is performed statically using the DUTS [6] dataset, which has 10 567 images and 2) fine-tuning is performed dynamically on DAVIS with 2300 video frames and FBMS with 600 video frames in training datasets. During the pretraining and fine-tuning, the SEAdam optimizer optimizes the total (weighted BCE and IoU) loss functions and updates the network weights. The process is given in Algorithm 1. The learnable parameter ω value is 0.0001. To complete 35 epochs with eight batch sizes, DSFNet takes approximately 10 h. The testing of DSFNet is evaluated on test datasets of DAVIS, FBMS, MCL, SegTrack-V2, DAVSOD, and DAVSOD-Difficult. We use the SOTA models' GitHub codes, pretrained weights, and saliency maps to test the performance of models. We implemented some papers independently and produced the results in the same environments. The testing performance of the DSFNet is

TABLE I

COMPARISON OF THE PERFORMANCE OF DSFNET WITH 13 SOTA VSOD MODELS AND SIX LIGHTWEIGHT VSOD MODELS ON SIX DATASETS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE IN THE ORDER RED > GREEN > BLUE. WHERE L-VSOD IS A LIGHTWEIGHT SOTA VSOD MODEL

		Backbone	# Param	FLOPs	Speed	DAVIS			FBMS			DAVSOD			SegTrack-V2			MCL			DAVSOD-Diff		
	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$_{21}$ [26]	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$_{22}$ [8]	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$_{23}$ [10]	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$_{23}$ [18]	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$_{23}$ [22]	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$_{23}$ [19]	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**
	DSFNet(ours)	DRNet	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	EREST$_{21}$ [42]	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$_{21}$ [6]	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$_{21}$ [24]	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$_{22}$ [15]	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$_{22}$ [2]	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$_{23}$ [55]	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$_{23}$ [4]	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$_{23}$ [7]	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$_{23}$ [12]	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$_{23}$ [56]	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$_{23}$ [9]	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$_{24}$ [23]	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$_{24}$ [29]	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

TABLE II

COMPARISON OF PARAMETER TUNING OF THE SEADAM OPTIMIZER FOR TRAINING THE DSFNET

Different Parameter Tuning									DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
β	**S**	β_1	β_2	c_1	c_2	r_1	r_2	η	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**
0.1	5	0.4	0.44	0.5	0.5	0	0	0.1	0.910	0.025	0.826	0.034	0.897	0.028	0.865	0.036	0.769	0.055	0.533	0.101
0.2	10	0.5	0.55	1.0	1.0	0.3	0.3	0.01	0.907	0.027	0.823	0.036	0.894	0.029	0.869	0.034	0.771	0.059	0.535	0.099
0.3	15	0.6	0.66	1.5	1.5	0.4	0.4	0.001	0.927	0.020	0.836	0.031	0.897	0.027	0.894	0.027	0.809	**0.053**	0.538	0.100
0.4	**20**	**0.9**	**0.999**	**2**	**2**	**0.5**	**0.5**	**0.0001**	**0.940**	**0.015**	**0.867**	0.021	**0.901**	**0.024**	**0.912**	**0.014**	**0.816**	0.054	**0.543**	**0.097**
0.5	25	0.5	0.5	2.5	2.5	0.6	0.6	0.00001	0.932	0.019	0.857	**0.019**	0.899	0.026	0.900	0.018	0.798	0.056	0.540	0.099

evaluated on benchmark metrics [60] such as S-measure (S_α), F-measure (F_β), Mean Absolute Error (MAE), and inference speed as presented in Table I.

B. Parameter Tuning

Parameter tuning is performed with different sets of values to check the performance variation between F-measure and MAE. The parameters of optimization follow the greed search techniques to tune the parameters. The tuning parameters are: momentum factor (β), cognitive/social learning parameters (c_1, c_2), random parameters (r_1, r_2), number of particle swarms S, and exponential decay rate and moment estimation (β_1, β_2), and learning rate (η). The range for parameters is as follows: momentum factor is considered as (0 to 1, 0 to 1), cognitive/social learning parameters (1 to 5, 1 to 5), random values between (0 to 1, 0 to 1), number of particle swarms (10, 30), exponential decay rate and moment estimation (0.1 to 0.5, 0.1 to 0.5), learning rate (0.1 to 0.00001), and maximum iteration is (1000 to 5000). As the learning rate, the number of particle swarms, exponential decay rate, and moment estimation with cognitive/social learning parameters increase, the performance of the proposed DSFNet model increases. Still, part of the time, the learning rate and cognitive/social learning parameters with random values decrease. So, the best combination of parameters and corresponding comparison results are presented in Table II.

C. Comparative Analysis

From the results presented in Table I, we observe that DSFNet outperforms SOTA models on datasets such as DAVIS, DAVSOD, MCL, and DAVSOD-Difficult, demonstrating superior performance in terms of complexity and computations across the majority of cases. The proposed model takes less time to generate saliency maps accurately. Furthermore, Table III provides a performance comparison between the proposed optimization algorithm (SEAdam) and other optimizers. In addition, the model complexity is compared regarding storage space (MB) and computation time (minutes, m). The detailed comparative results are outlined in Table IV.

In Fig. 4, the proposed model (DSFNet) is compared with the latest SOTA methods under various difficult scenes. The DSFNet can distribute salient objects and nonsalient objects from difficult situations, including confusing scenarios in crowd scenes (first rows), cluttered and thin object backgrounds with low light (third rows), small objects with occlusion (first and fourth rows), similar background and foreground (second rows), and dynamic scene changing with deformation (first, second, third, and fourth rows). Due to the fusion of deformability and separability, the proposed model implicitly detects the object properly with less error. However, as shown in Fig. 5 rows 1 and 2, the proposed model is getting confused in crowd scenes and cluttered backgrounds.

D. Limitations and Future Directions

DSFNet has faced several challenges: 1) obtaining accurate detection results for crowd scenes, partial occlusion, and fast-moving objects remains challenging for DSFNet and SOTA models from Fig. 5 (rows 1 and 2). When the object has low-light and shadow scenes as given in Fig. 5 (rows 3 and 4), the DSFNet and SOTA models failed to distinguish the object's

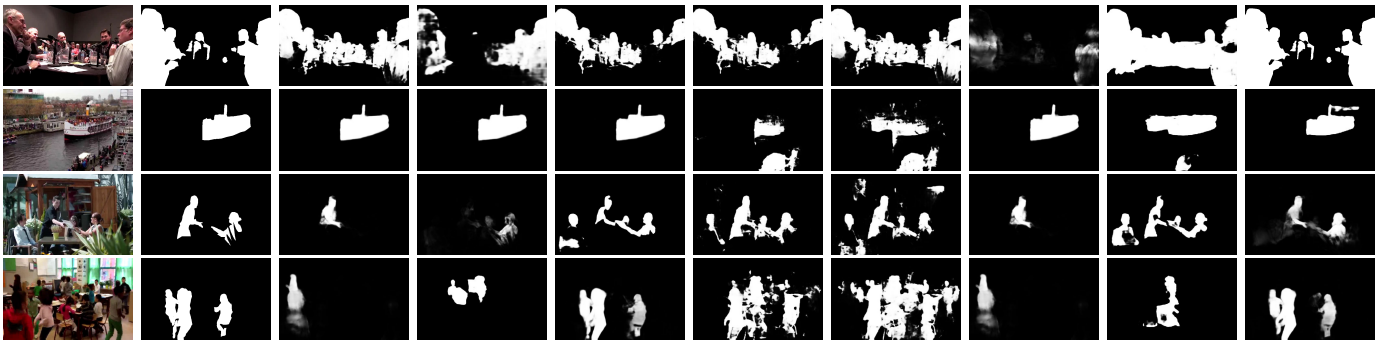


Fig. 4. Performance comparison of the proposed DSFNet model and SOTA models on more difficult scenarios of the DAVSOD-Difficult dataset. IF is the input frame and GT is the ground truth.

TABLE III
COMPARISON OF THE DIFFERENT OPTIMIZERS TO THE SEADAM OPTIMIZER OF DSFNET

	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
Different Optimizer	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
SGD	0.901	0.023	0.824	0.033	0.895	0.029	0.861	0.035	0.766	0.059	0.535	0.102
Adam [40]	0.911	0.021	0.822	0.039	0.889	0.027	0.893	0.028	0.773	0.058	0.536	0.100
AdaNorm [41]	0.927	0.023	0.837	0.034	0.863	0.031	0.870	0.033	0.769	0.060	0.529	0.101
AdaSwarm [36]	0.932	0.019	0.854	**0.018**	0.900	**0.020**	0.885	0.030	0.787	0.055	0.539	**0.096**
SEAdam	**0.940**	**0.015**	**0.867**	0.021	**0.901**	0.024	**0.912**	**0.014**	**0.816**	**0.054**	**0.543**	0.097

TABLE IV
COMPARISON BETWEEN SOTA MODELS AND THE PROPOSED MODEL (DSFNET) REGARDING THE MODEL SIZE AND COMPUTATIONAL TIME

Computation Measure	EUVSOD	VS-Net	TinyHD	DSNet	EREST	DefED-Net	FSNet	MSDM	CFAM	STDF	HCPN	TMO	PMN	SPGO	PACNet	LSTA	CoSTFormer	SKD	InternImage	DSFNet
Model Size (MB)	588.0	70.0	132.0	587.0	499.0	67.9	487.0	83.0	143.2	189.0	321.3	298.0	338.0	458.2	256.3	340.9	180.1	280.2	320.4	**60.0**
Computation Time (m)	5.6	5.2	4.8	4.9	4.4	4.0	4.5	3.1	2.9	2.8	4.6	4.1	3.2	4.5	3.8	3.4	3.0	3.0	3.6	**2.6**

foreground and background effectively and 2) DSFNet still requires high computational systems and large memory space for extracting the multiscale geometric spatial and temporal information and storing the network trainable weight during training due to embedding multiple attention modules. These difficulties arise from the fact that the spatial and temporal information of these objects are inherently weak and can be easily disrupted or obscured by the spatial and temporal information of nearby objects. Furthermore, the crowd scene, partial occlusion, and fast-moving objects are well-known problems for VSOD tasks. In the future, knowledge distillation, quantization, and compression techniques will be adopted to overcome these problems without compromising accuracy. Temporal dependencies will be handled by proposing some temporal-based modules and making it feasible to adopt multiple domains.

E. Effectiveness of DRNet

As explained earlier, the deformable convolution with spatial adaptive pooling mechanisms enables the network to dynamically adjust its pooling regions as per the content of the input feature maps, which enhances its ability to handle variations in object scales and positions. The residual connections help mitigate the vanishing gradient problem and enable the model to capture richer and more discriminative feature representations, particularly in tasks where objects exhibit nonrigid deformations or variations in shape and appearance. From Table V, in comparison to other lightweight backbone networks, DRNet is giving a better result and requires less storage space. Its energy efficiency is lower than that of other SOTA backbone networks. Furthermore, to show the effectiveness of the DRNet, we replace convolution layers of the other simplified versions of ResNet (18, 34, 50, and 101) with deformable convolution and find that as the layers are reduced, the performance of the model decreases and model complexity increases, while as the layer is increased, the performance is saturated and model complexity is increased as shown in Table VI.

F. Effectiveness of SEAdam

The SEAdam is leveraging the adaptive learning rates and stochastic estimation capabilities. The adaptive learning rates allow SEAdam to adjust the learning rates of individual parameters based on their gradients, ensuring that small gradients are amplified and large gradients are attenuated to prevent oscillations or divergence. It navigates regions of the optimization landscape with varying gradients. At the same time, stochastic estimation injects randomness into the updates to adjust the learning rate across all parameters to ensure smooth convergence toward the global minimum of the loss function and enable the exploration of new directions. Through these mechanisms, SEAdam increases the likelihood of escaping local minima and converging toward better solutions. From Table III, in comparison to other optimizers, SEAdam is giving

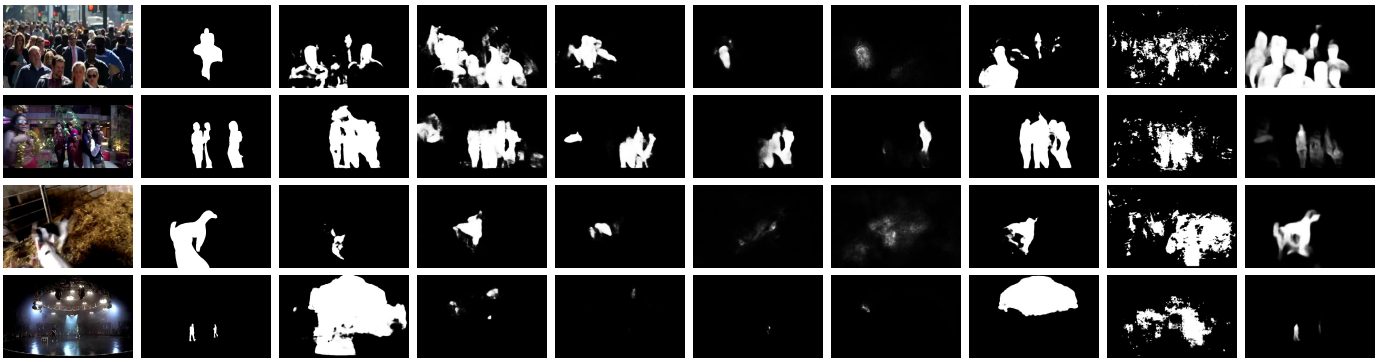


Fig. 5. Failure case of DSFNet and SOTA models on the DAVSOD-Difficult dataset. IF is the input frame and GT is the annotation map.

TABLE V
COMPARISON OF THE DIFFERENT TYPES OF LIGHTWEIGHT BACKBONE NETWORK WITH DSFNET

	# Param	FLOPs	Speed	Memory Footprint	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
Backbone	**(M)**	**(G)**	**(FPS)**	**(MB)**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**
LightViT [61]	9.4	0.73	**100.2**	25	0.899	0.025	0.826	0.043	0.873	0.030	0.862	0.041	0.763	0.054	0.437	0.111
MobileNet2.0 [30]	**3.2**	**1.2**	59.3	32	0.907	0.026	0.820	0.047	0.866	0.044	0.869	0.037	0.775	0.046	0.458	0.109
VGG-16 [62]	6.4	5.9	43.2	29	0.899	0.032	0.832	0.039	0.885	0.025	0.889	0.032	0.788	0.043	0.484	0.099
EfficientNet [8]	12.9	10.5	48.9	32	0.912	0.019	0.869	0.033	0.890	0.033	0.895	0.023	0.784	**0.041**	0.479	0.103
PeleeNet [63]	5.9	**1.2**	46.0	38	0.902	0.025	0.850	0.037	0.863	0.047	0.879	0.041	0.781	0.046	0.475	0.101
SqueezeNet [64]	5.3	2.2	100.0	48	0.909	0.028	0.860	0.035	0.899	0.033	0.883	0.039	0.769	0.053	0.460	0.105
ResNet-50 [51]	26.2	22.3	34.3	46	0.913	0.029	0.854	0.040	0.894	0.037	0.876	0.041	0.760	0.059	0.466	0.100
ResNeXt-50 [65]	25.0	24.3	44.3	43	0.910	0.030	0.819	0.050	0.891	0.030	0.871	0.043	0.771	0.056	0.456	0.103
DRNet	23.4	20.0	84.0	60	**0.940**	**0.015**	**0.867**	**0.021**	**0.901**	**0.024**	**0.907**	**0.014**	**0.816**	0.054	**0.543**	**0.097**

TABLE VI
COMPARISON OF THE DIFFERENT TYPES OF RESNET BACKBONE NETWORKS WITH DEFORMABLE CONVOLUTION USED IN DSFNET

	# Param	FLOPs	Speed	Memory Footprint	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
Backbone	**(M)**	**(G)**	**(FPS)**	**(MB)**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**
DRNet-18	13.5	12.6	90.3	45	0.904	0.019	0.823	0.054	0.857	0.042	0.868	0.021	0.745	0.061	0.510	0.105
DRNet-34	18.4	16.7	86.9	52	0.919	0.017	0.845	0.044	0.877	0.032	0.882	0.017	0.777	0.058	0.523	0.100
DRNet-50	23.4	20.0	84.0	60	**0.940**	**0.015**	**0.867**	**0.021**	**0.901**	**0.024**	**0.907**	**0.014**	**0.816**	0.054	**0.543**	**0.097**
DRNet-101	45.8	41.5	65.0	75	0.933	0.016	0.860	0.026	0.890	0.027	0.898	0.019	0.800	**0.050**	0.536	0.099

TABLE VII
ABLATION STUDIES FOR THE COMPONENTS SETTING OF DSFNET

	Component Setting					DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
No.	**DSRFB**	**DAFM**	**DGAM**	**DRCM**	**LFI**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**	S_α	**MAE**
1	✓	✓		✓		0.920	0.024	0.833	0.038	0.839	0.040	0.872	0.029	0.769	0.067	0.519	0.105
2		✓	✓		✓	0.923	0.020	0.848	0.027	0.843	0.037	0.877	0.027	0.775	0.065	0.521	0.104
3	✓		✓	✓		0.918	0.027	0.840	0.031	0.840	0.039	0.870	0.030	0.768	0.069	0.518	0.107
4		✓		✓	✓	0.930	0.019	0.852	0.030	0.852	0.031	0.883	0.023	0.782	0.061	0.526	0.102
5	✓	✓	✓			0.926	0.024	0.855	0.027	0.848	0.035	0.880	0.025	0.780	0.063	0.529	0.099
6		✓	✓	✓		0.933	0.018	0.839	0.034	0.889	0.029	0.889	0.020	0.779	0.066	0.538	0.101
7	✓		✓	✓	✓	0.934	0.016	0.866	0.022	0.892	0.027	0.893	0.016	0.792	0.057	0.530	0.103
8	✓	✓	✓	✓		0.937	**0.014**	0.845	0.025	0.896	0.026	0.897	0.017	0.799	**0.054**	0.542	0.099
9		✓	✓	✓	✓	0.932	0.017	**0.869**	0.024	0.893	0.028	0.898	0.018	0.795	0.056	0.537	**0.096**
10	✓	✓	✓	✓	✓	**0.940**	0.015	0.867	**0.021**	**0.901**	**0.024**	**0.912**	**0.014**	**0.816**	**0.054**	**0.543**	0.097

a better result and said that it explores and exploits the gradient locally and globally efficiently.

G. Evaluation on the DAVSOD-Difficult Dataset

The DAVSOD-Difficult dataset [59] (DAVSOD-Diff) is notably one of the most challenging datasets, encompassing various instances of wild and unconstrained scenarios. The results obtained from the DAVSOD-Diff dataset underline the significance of our proposed DSFNet model, particularly considering its fewer parameters. As illustrated in Table I, the proposed DSFNet demonstrates a performance improvement, whereas other models experience a drastic decrease in performance. In comparison to the recent and top-performing FSNet model [6], DSFNet exhibits a 17.9% decrease in S_α, a 4.1% increase in F_β, and a 2.0% decrease in MAE. Against the EUVSOD model [8], DSFNet showcases substantial improvements with a 37.5% increase in S_α, a 38.5% increase in F_β, and a 12.2% decrease in MAE. Similarly, in comparison to DSNet [19], DSFNet demonstrates an increase of 4.4% in S_α, 1.9% in F_β, and 1.0% in MAE. Notably, the proposed DSFNet model achieves these results without utilizing any validation set from the DAVSOD-Diff dataset during training, underscoring its robust generalization capabilities.

H. Ablation Study

The first part of the ablation study, as shown in Table VII, explains the requirements of the proposed components. Various combinations of these components are explored to

TABLE VIII
ABLATION STUDIES OF MODULES CONFIGURATION WITH DIFFERENT # FILTERS AND DILATION RATES OF DSFNET

	Parameters		DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
No.	# filters	Dilation rate	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
1	16	1, 4, 8, 9	0.912	0.030	0.798	0.039	0.889	0.040	0.860	0.045	0.749	0.057	0.498	0.109
2	32	1, 5, 8, 12	0.923	0.023	0.844	0.032	**0.903**	**0.023**	0.889	0.035	0.799	**0.048**	0.527	0.103
3	**64**	**1, 6, 12, 18**	**0.940**	**0.015**	**0.867**	**0.021**	0.901	0.024	**0.912**	**0.014**	**0.816**	0.054	**0.543**	**0.097**
4	128	1, 6, 10, 14	0.938	0.017	0.860	0.025	0.888	0.040	0.887	0.038	0.789	0.053	0.533	0.098
5	256	1, 7, 9, 12	0.925	0.020	0.833	0.032	0.873	0.045	0.878	0.037	0.769	0.056	0.458	0.115
6	512	1, 5, 9, 13	0.919	0.021	0.828	0.029	0.866	0.049	0.869	0.040	0.786	0.055	0.455	0.118

TABLE IX
EFFECTIVENESS OF DIFFERENT DSRFB MODULES OF DSFNET

	DAVIS		MCL		FBMS		SegTrack-V2		DAVSOD		DAVSOD-Diff	
DSRFB	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE	S_α	MAE
N=0	0.923	0.020	0.860	0.026	0.885	0.029	0.871	0.029	0.776	0.064	0.435	0.122
N=1	0.930	0.018	0.857	0.025	0.889	0.030	0.890	0.025	0.793	0.060	0.542	**0.096**
N=2	0.934	0.016	0.866	0.024	0.894	0.027	0.893	0.020	0.795	0.059	0.538	0.099
N=3	**0.940**	**0.015**	0.867	**0.021**	**0.901**	**0.024**	**0.912**	**0.014**	0.816	**0.054**	**0.543**	0.097
N=4	0.936	0.015	**0.869**	0.022	0.897	0.026	0.899	0.016	**0.818**	0.057	0.540	0.098

generate the most effective multicontext, multiscale geometric spatiotemporal features. It can be observed from the table that as the components are included in the proposed DSFNet model, the performance is progressively improved. In addition, the proposed solution shows superiority when compared to No. 1 and No. 10 in the baseline, and it shows that the proposed DSFNet model completely depends on component settings. The second part of the ablation study (shown in Table VIII) describes the configurations of various filters and dilation rates: 1) the kernel size and dilation are used in ascending order in each DSFNet module to perform consistently and 2) to improve the proposed model (DSFNet) capacity and performance, the filters and dilation rates are used in different combinations. As the filters are increased from 16 to 64, the performance is increased continuously, as shown in rows No. 1 to No. 3 and from 128 to 512, downgrading the performance of the DSFNet model as shown in rows No. 4 to No. 6. The default settings are considered to balance between the accuracy and efficiency of the DSFNet model as shown in Table VIII. It is interesting to learn from Table IX that the best configuration of the DSFNet model is DSRFB $= 3$ with all components (DAFM, DRCM, DGAM, and LFI).

V. CONCLUSION

In this article, a novel lightweight DSFNet is designed to handle the problem of VSOD in unconstrained scenarios and maintain the accuracy and speed tradeoff, even with the most challenging DAVSOD-Diff dataset. Through extensive experimentation and a comprehensive ablation study on six publicly available benchmark datasets, we show that the DSFNet outperforms all large VSOD models and lightweight VSOD models in terms of both F_β and MAE across most datasets. The performance of DSFNet on the most difficult "DAVSOD-Difficult" dataset is the highlight of the article. With the 53% increase in the number of parameters and 56% increase in FLOPS compared to the existing best-performing lightweight model VSNet [10], DSFNet can achieve 8% increase in S_α, 13.7% increase in F_β, and 28% decrease in MAE on this dataset. One considerable limitation of the work is the reduced detection performance in cases of occluded objects with cluttered backgrounds and deformed moving objects. In the future, we plan to work on this limitation and evaluate the performance of the proposed model on edge devices for real-time federated processing of surveillance videos.

DATA AVAILABILITY STATEMENT

It confirms that the dataset we used in the experiment is publicly available, and their original origin has been cited in the article.

REFERENCES

[1] Y. Dai, C. Xue, and L. Zhou, "Visual saliency guided perceptual adaptive quantization based on HEVC intra-coding for planetary images," *PLoS ONE*, vol. 17, no. 2, Feb. 2022, Art. no. e0263729.

[2] Y. Tang, Y. Li, and W. Zou, "Fast video salient object detection via spatiotemporal knowledge distillation," 2020, *arXiv:2010.10027*.

[3] H. Singh, M. Verma, and R. Cheruku, "Novel dilated separable convolution networks for efficient video salient object detection in the wild," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 5023213.

[4] S. Cho, M. Lee, S. Lee, C. Park, D. Kim, and S. Lee, "Treating motion as option to reduce motion dependency in unsupervised video object segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5140–5149.

[5] X. Jiang et al., "Camouflaged object segmentation based on joint salient object for contrastive learning," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 5023816.

[6] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4922–4933.

[7] M. Lee, S. Cho, S. Lee, C. Park, and S. Lee, "Unsupervised video object segmentation via prototype memory network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5924–5934.

[8] C. Hu and L. Zhu, "Efficient unsupervised video object segmentation network based on motion guidance," 2022, *arXiv:2211.05364*.

[9] N. Liu, K. Nan, W. Zhao, X. Yao, and J. Han, "Learning complementary spatial–temporal transformer for video salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 10663–10673, Aug. 2024.

[10] H. Singh, M. Verma, and R. Cheruku, "VS-net: Multiscale spatiotemporal features for lightweight video salient document detection," in *Proc. IEEE 34th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Oct. 2022, pp. 1307–1311.

[11] D. Peng, W. Zhou, J. Pan, and D. Wang, "MSEDNet: Multi-scale fusion and edge-supervised network for RGB-T salient object detection," *Neural Netw.*, vol. 171, pp. 410–422, Mar. 2024.

[12] G. Ponimatkin, N. Samet, Y. Xiao, Y. Du, R. Marlet, and V. Lepetit, "A simple and powerful global optimization for unsupervised video object segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5892–5903.

[13] T.-N. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5002–5015, Oct. 2018.

[14] S.-H. Gao, Q. Han, Z.-Y. Li, P. Peng, L. Wang, and M.-M. Cheng, "Global2Local: Efficient structure search for video action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16805–16814.

[15] Y.-W. Chen, X. Jin, X. Shen, and M.-H. Yang, "Video salient object detection via contrastive features and attention modules," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1320–1329.

[16] A. Kompella and R. V. Kulkarni, "A semi-supervised recurrent neural network for video salient object detection," *Neural Comput. Appl.*, vol. 33, no. 6, pp. 2065–2083, Mar. 2021.

[17] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial–temporal information," *IEEE Trans. Image Process.*, vol. 29, pp. 1090–1100, 2020.

[18] F. Hu et al., "TinyHD: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2051–2060.

[19] H. Singh, M. Verma, and R. Cheruku, "DSNet: Efficient lightweight model for video salient object detection for IoT and WoT applications," in *Proc. Companion ACM Web Conf.*, 2023, pp. 1286–1295.

[20] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[22] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419.

[23] J. Deng, S. Dong, L. Chen, J. Hu, and C. Zhuo, "STDF: Spatio-temporal deformable fusion for video quality enhancement on embedded platforms," *ACM Trans. Embedded Comput. Syst.*, vol. 23, no. 2, pp. 1–25, 2024.

[24] Y. Liu et al., "Exploring multi-scale deformable context and channel-wise attention for salient object detection," *Neurocomputing*, vol. 428, pp. 92–103, Mar. 2021.

[25] S. Chen, L. Zhang, and L. Zhang, "MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5525614.

[26] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "DefED-net: Deformable encoder–decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 1, pp. 68–78, Jan. 2022.

[27] G. Pei, F. Shen, Y. Yao, G.-S. Xie, Z. Tang, and J. Tang, "Hierarchical feature alignment network for unsupervised video object segmentation," 2022, *arXiv:2207.08485*.

[28] X. Liu and L. Wang, "MSRMNet: Multi-scale skip residual and multi-mixed features network for salient object detection," *Neural Netw.*, vol. 173, May 2024, Art. no. 106144.

[29] P. Li, Y. Zhang, L. Yuan, H. Xiao, B. Lin, and X. Xu, "Efficient long-short temporal attention network for unsupervised video object segmentation," *Pattern Recognit.*, vol. 146, Feb. 2024, Art. no. 110078.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.

[31] T. Hussain, K. Muhammad, J. D. Ser, S. W. Baik, and V. H. C. de Albuquerque, "Intelligent embedded vision for summarization of multiview videos in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2592–2602, Apr. 2020.

[32] A. Khan, M. Kuribayashi, K. Wong, and V. Monn Baskaran, "HDR image watermarking using saliency detection and quantization index modulation," 2023, *arXiv:2302.11361*.

[33] J. Zhang, Q. Liang, and Y. Shi, "KD-SCFNet: Towards more accurate and efficient salient object detection via knowledge distillation," 2022, *arXiv:2208.02178*.

[34] Y. Huang et al., "CP3: Channel pruning plug-in for point-based networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5302–5312.

[35] F. Jia, X. Wang, J. Guan, H. Li, C. Qiu, and S. Qi, "WRGPruner: A new model pruning solution for tiny salient object detection," *Image Vis. Comput.*, vol. 109, May 2021, Art. no. 104143.

[36] R. Mohapatra, S. Saha, C. A. C. Coello, A. Bhattacharya, S. S. Dhavala, and S. Saha, "AdaSwarm: Augmenting gradient-based optimizers in deep learning with swarm intelligence," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 329–340, Apr. 2022.

[37] S. M. Mikki and A. A. Kishk, *Particle Swarm Optimization: A Physics-based Approach*, vol. 20. San Rafael, CA, USA: Morgan & Claypool, 2008.

[38] S. Chen and J. Montgomery, "Particle swarm optimization with thresheld convergence," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2013, pp. 510–516.

[39] R. Mohapatra, R. R. Talesara, S. Govil, S. Saha, S. S. Dhavala, and T. Sudarshan, "A new approach for momentum particle swarm optimization," in *Proc. Adv. Mach. Learn. Comput. Intell. (ICMLCI)*. Singapore: Springer, 2020, pp. 47–63.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[41] S. R. Dubey, S. Kumar Singh, and B. B. Chaudhuri, "AdaNorm: Adaptive gradient norm correction based optimizer for CNNs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5284–5293.

[42] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, and H. Qin, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3995–4007, 2021.

[43] K. Zhang, Z. Zhao, D. Liu, Q. Liu, and B. Liu, "Deep transport network for unsupervised video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8781–8790.

[44] M. Zhang et al., "Dynamic context-sensitive filtering network for video salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1553–1563.

[45] Z. Zhang, P. Gao, S. Peng, C. Duan, and P. Zhang, "Enhanced point feature network for point cloud salient object detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 1617–1621, 2023.

[46] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[47] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "Geometry auxiliary salient object detection for light fields via graph neural networks," *IEEE Trans. Image Process.*, vol. 30, pp. 7578–7592, 2021.

[48] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, Nov. 2022.

[49] Y. Tang, W. Zou, Z. Jin, and X. Li, "Multi-scale spatiotemporal Conv-LSTM network for video saliency detection," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2018, pp. 362–369.

[50] M. Xu, P. Fu, B. Liu, and J. Li, "Multi-stream attention-aware graph convolution network for video salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4183–4197, 2021.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.

[53] M. Bansal, M. Kumar, and M. Kumar, "2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 18839–18857, May 2021.

[54] H. Seong, J. Hyun, and E. Kim, "Kernelized memory network for video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, vol. 33. Springer, 2020, pp. 629–645.

[55] G. Pei, Y. Yao, F. Shen, D. Huang, X. Huang, and H.-T. Shen, "Hierarchical co-attention propagation network for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 2348–2359, 2023.

[56] Z.-Y. Liu and J.-W. Liu, "Part-aware attention correctness for video salient object detection," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105733.

[57] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 724–732.

[58] H. Wang et al., "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1509–1513.

[59] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8546–8556.

[60] H. Zhou, Y. Lin, L. Yang, J. Lai, and X. Xie, "Benchmarking deep models on salient object detection," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109951.

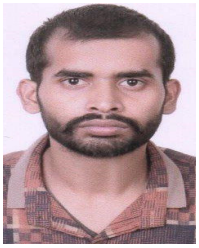
[61] T. Huang, L. Huang, S. You, F. Wang, C. Qian, and C. Xu, "LightViT: Towards light-weight convolution-free vision transformers," 2022, *arXiv:2207.05557*.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[63] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[64] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.

[65] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1492–1500.



Hemraj Singh received the Diploma degree in computer science and engineering from BTEUP, Lucknow, Uttar Pradesh, India, in 2013, the B.Tech. degree in computer science and engineering from AKTU University, Lucknow, in 2017, and the M.Tech. degree in artificial intelligence from NIT Uttarakhand, Srinagar, India, in 2020. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, the National Institute of Technology Warangal, Hanamkonda, Telangana, India.

His research interests include computer vision and image processing, video processing, artificial intelligence, machine learning, and deep learning.

Mridula Verma (Member, IEEE) received the M.Tech. degree in computer science engineering from the Indian Institute of Technology at Roorkee, Roorkee, India, in 2009, and the Ph.D. degree in computer science engineering from the Indian Institute of Technology at Varanasi, Varanasi, India, in 2017.

She is an Assistant Professor and the Head of the Artificial Intelligence and Machine Learning Laboratory, Institute for Development and Research in Banking Technology (IDRBT), Hyderabad, India. Her research interests include practical machine learning, federated learning, privacy-preserving machine learning, and financial NLP.

Ramalingaswamy Cheruku (Senior Member, IEEE) received the B.Tech. degree in CSE from JNT University, Kakinada Campus, Kakinada, India, in 2008, the M.Tech. degree in CSE from ABV-IIIT, Gwalior, India, in 2011, and the Ph.D. degree in CSE from NIT Goa, Cuncolim, Goa, India, in 2018.

He is currently an Assistant Professor with the Department of CSE, National Institute of Technology Warangal, Hanamkonda, Telangana, India. He has published more than 30 journal articles and 20 conference papers in reputed venues.