Several-Attention Network for Unsupervised Video Object Division

Prasant Sethi,
College of Engineering Bhubaneswar

***Abstract*—Some helpful unsupervised video object segmentation techniques that highlight the common information in videos have been put out in recent years. Although these techniques perform well, they are unable to separate the specifics of the objects since they do not take into account the data from the lower layers of the network. We suggest a multi-attention network for unsupervised video object segmentation (MANet) as a solution to this issue. According to recent research, deep layers of networks are more responsive to high-level semantic information than messy details, but shallow layers are the opposite. Based on this insight, a multi-attention module is built by considering not only the deep layer information but also the shallow layer information. By optimizing the shared information across video frames and combining features from both the shallow and deep layers, this module is able to successfully discern the main object and segment its details. Based on experimental results on the SegTrack v2 and DAVIS-2016 datasets, our network surpasses the state-of-the-art techniques.**

***Index Terms*—Attention, deep networks, unsupervised video object segmentation.**

I. Introduction

VIDEO object segmentation (VOS) is a fundamental task of associating each pixel of each frame in a video with a label. VOS has a wide range of applications, such as activity recognition, content-based retrieval, and object tracking [20]. VOS methods can be roughly classified into unsupervised VOS methods (UVOS) and semi-supervised VOS methods. In this letter, we focus on the UVOS, which aims to segment the primary object in each frame without any annotation. UVOS is a challenging task because it suffers from the typical challenges for video object segmentation, such as object deformation and occlusion. Besides, due to the lack of annotations, UVOS methods need to automatically distinguish the primary object.

In the early stages, motion information and saliency information were utilized to segment the primary object. For example, Hu *et al.* [6] developed a novel saliency estimation method and a graph neighborhood for UVOS. Later, with the development of deep learning [7], many deep learning-based UVOS methods have been proposed. Li *et al.* [10] leveraged the instance embeddings and optical flow features to achieve UVOS. Wang *et al.* [17] assumed that the visual attention mechanism is related to the primary object and adopted the visual attention mechanism to UVOS tasks. Although these deep learning-based UVOS methods can obtain satisfactory results in most cases, they only take the advantage of short-term information, the related information between video frames is not used.

Recently, many methods segment the primary object by introducing the attention mechanism to emphasize the common information between video frames. The attention mechanism works to strengthen the context information of the features by making the network focus on the critical part [15]. In [11] and [21], the attention mechanism works to strengthen the foreground information and weaken the background information. In [11], Lu *et al.* adopted the co-attention mechanism to UVOS and utilized co-attention layers to capture the global information. In [21], Yang *et al.* combined the intra-frame information and inter-frame information for better segmenting the primary object. Both methods emphasize the common information in videos. However, they only leverage the information from the deep layers of the network, ignoring the information from the shallow layers. Recent studies show that the deep layers of networks capture high-level semantic information but messy details, while it is opposite for shallow layers [3]. From this insight, we consider that the information from the shallow layers should be equally treated. Therefore, we propose a multi-attention network for unsupervised video object segmentation (MANet) to take into account the information from the shallow layers for segmenting the details of the objects.

The contributions of this letter are as follows. 1) We propose a multi-attention network for unsupervised video object segmentation method that can effectively segment the details of the objects to obtain satisfactory VOS results. 2) We design a multi-attention module that computes the attention information not only based on information from the deep layers, but also based on information from the shallow layers. This module helps to distinguish the primary object and process the details of the object. Experiments on the DAVIS-2016 [13] and SegTrack v2 [8] datasets demonstrate that our proposed MANet can achieve better performance than those of the state-of-the-art methods.

II. Method

Our network consists of two cascaded parts: the encoder and decoder. Fig. 1 shows the architecture of our network. In the following, we describe each of the components in detail.

Manuscript received December 10, 2020; accepted December 14, 2020. Date of publication December 17, 2020; date of current version January 15, 2021. This work was supported by the Science and Technology Development Fund of Macao SAR under Grant 0016/2019/A1 and Grant 0027/2018/A1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianxin Li. *(Corresponding author: Hon-Cheng Wong.)*

The authors are with the Faculty of Information Technology, Macau University of Science and Technology, Macao, China (e-mail: 1439402142@qq.com; hcwong@ieee.org; sllo@must.edu.mo).

Digital Object Identifier 10.1109/LSP.2020.3045641

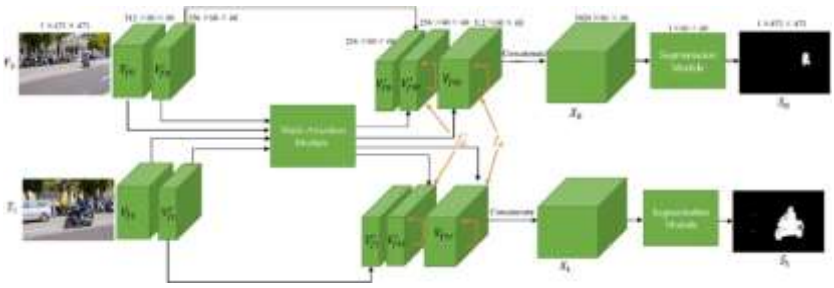

Fig. 1. Overview of MANet. The first frame F_0 and the current frame F_t are fed into the Deeplabv3 to obtain their feature representations V_{ft}, $V_{ft}^{'}$, V_{f0}, and $V_{f0}^{'}$. Then, V_{ft}, $V_{ft}^{'}$, V_{f0}, and $V_{f0}^{'}$ are fed into the multi-attention module to obtain their refined feature representations V_{f0r}, V_{ftr}, $V_{f0r}^{'}$ and $V_{ftr}^{'}$. After that, V_{f0r} is concatenated with $V_{f0r}^{'}$ and V_{f0} to generate X_0; V_{ftr} is concatenated with $V_{ftr}^{'}$ and V_{ft} to generate X_t. Finally, X_t and X_0 are fed into a segmentation module to produce VOS segmentation results S_0 and S_t.

A. Encoder Network

We take the DeepLabv3 [2] which is mainly composed of ResNet101 [22] and ASPP (Atrous Spatial Pyramid Pooling) as our encoder network. Our encoder is a shared network with two branches. These two branches respectively take the first frame F_0 and the current frame F_t as inputs to extract their feature representations from the deep layers and shallow layers. In this work, we take the features from Conv3_x as our shallow layer feature representations, and the features from ASPP as our deep layer feature representations: $V_{f0} \in \mathrm{R}^{W\times H\times 2C}$, $V_{f0}^{'} \in \mathrm{R}^{W\times H\times C}$, $V_{ft} \in \mathrm{R}^{W\times H\times 2C}$, $V_{ft}^{'} \in \mathrm{R}^{W\times H\times C}$, where V_{f0} and V_{ft} denote the shallow layer feature representations, $V_{f0}^{'}$ and $V_{ft}^{'}$ denote the deep layer feature representations, H and W denote the height and width of the frame, and C and $2C$ represent the number of channels in the deep layers and the shallow layers, respectively.

B. Decoder Network

The decoder network takes the original feature representations ($V_{f0} \in \mathrm{R}^{W\times H\times 2C}$, $V_{f0}^{'} \in \mathrm{R}^{W\times H\times C}$, $V_{ft} \in \mathrm{R}^{W\times H\times 2C}$, $V_{ft}^{'} \in \mathrm{R}^{W\times H\times C}$) as inputs to obtain segmentation results (S_0 and S_t). First, feature representations are refined by the multi-attention module to obtain refined feature representations (V_{ftr}, V_{f0r}, $V_{ftr}^{'}$ and $V_{f0r}^{'}$). And then, the refined feature representations are concatenated together along the original feature representations to generate the concatenated feature representations (X_t and X_0). Finally, the concatenated feature representations are processed by the segmentation module to carry out VOS segmentation results.

Recent studies show that the deep layers of the network capture high-level semantic information but messy details, while it is opposite for shallow layers [3]. From this insight, our multi-attention module is designed to enhance the common information between video frames and combine the features from the shallow layers and deep layers. The purpose of enhancing common information is to distinguish the primary object. Combing the features from the shallow layers and deep layers helps to process the details of the object. Fig. 2 illustrates the multi-attention module. In this module, we first respectively compute the distance maps D and $D^{'}$ of the shallow layer feature representations and the deep layer feature representations

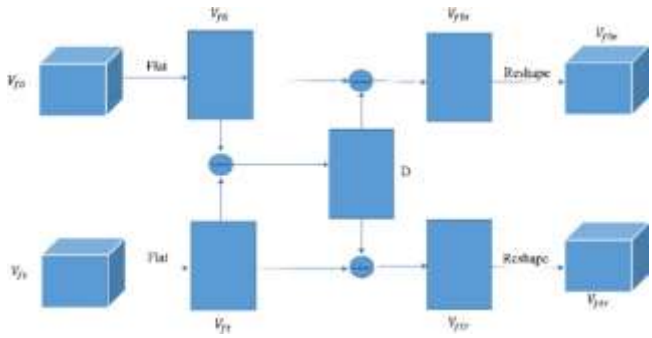

Fig. 2. Illustration of the multi-attention module. $\ominus$ denotes the element-wise subtract. We design the multi-attention module based on the features from one deep layer and one shallow layer. In this illustration, we just give the attention computation for the features (V_{f0} and V_{ft}) from the deep layer. First, we flat V_{f0}, and V_{ft} to obtain two flatted feature maps. And then, we generate the distance map D according to the two flatted feature maps. Finally, we use the original V_{f0}, and V_{ft} to respectively subtract the distance map D to obtain two refined feature maps V_{ftr} and V_{f0r}. In order to get the same shape as V_{f0}, and V_{ft}, a reshape operation is applied to V_{ftr} and V_{f0r}.

between the current frame and the first frame. D represents the pixel-level difference between V_{ft} and V_{f0}, $D^{'}$ represents the pixel-level difference between $V_{ft}^{'}$ and $V_{f0}^{'}$. In D, the feature values denote the dissimilarity (distance) between V_{ft} and V_{f0}. The dissimilar features in V_{ft} and V_{f0} have higher feature values than those of the similar features. The following formulas are used to compute D and $D^{'}$:

$$D = |V_{ft} - V_{f0}|, D^{'} = |V_{ft}^{'} - V_{f0}^{'}| \quad (1)$$

where $D \in \mathrm{R}^{WH\times 2C}$, $D^{'} \in \mathrm{R}^{WH\times C}$, here $V_{ft} \in \mathrm{R}^{WH\times 2C}$, $V_{f0} \in \mathrm{R}^{WH\times C}$, $V_{ft}^{'} \in \mathrm{R}^{WH\times C}$, and $V_{f0}^{'} \in \mathrm{R}^{WH\times C}$.

Then we use the distance map D to refine V_{ft} and V_{f0}, and use the distance map $D^{'}$ to refine $V_{ft}^{'}$ and $V_{f0}^{'}$:

$$V_{ftr} = V_{ft} - D, V_{f0r} = V_{f0} - D \quad (2)$$

V_{ftr} and V_{f0r} denote the refined shallow layer feature representations. The refined deep layer feature representations $V_{ftr}^{'}$ and $V_{f0r}^{'}$ can also be obtained after the same operations.

For the sake of better utilizing multi-attention to segment the primary object, inspired by COSNet [11], we use two gates (f_g and $f_g^{'}$) to allocate the multi-attention confidence to multi-attention. f_g is the gate of V_{ftr} and V_{f0r}, $f_g^{'}$ is the gate of $V_{ftr}^{'}$

and $V_{f0r}^{\downarrow}$. The gate f_g is formulated as follows:

$$f_g(V_{ftr}) = \sigma(w_f V_{ftr} + b_f) \in [0, 1]^{WH}$$
$$f_g(V_{f0r}) = \sigma(w_f V_{f0r} + b_f) \in [0, 1]^{WH} \quad (3)$$

where σ is the logistic sigmoid activation function, w_f and b_f are the convolution kernels and bias, respectively. We can compute the gate $f_g^{\downarrow}$ based on the same operations. After calculating the gate confidences, the refined feature representations V_{ftr} and V_{f0r} can be updated by:

$$V_{ftr} = V_{ft} * f_g(V_{ftr}),\ V_{f0r} = V_{f0r} * f_g(V_{f0r}) \quad (4)$$

where $*$ denotes the channel-wise Hadamard product. Gated $V_{ftr}^{\downarrow}$ and $V_{f0r}^{\downarrow}$ can also be obtained according to the same operations.

After obtaining the gated features, we concatenate the gated features with the original features together:

$$X_0 = [V_{f0r}, V_{f0r}^{\downarrow}, V_{f0}] \in R^{W \times H \times 4C}$$
$$X_t = [V_{ftr}, V_{ftr}^{\downarrow}, V_{ft}] \in R^{W \times H \times 4C} \quad (5)$$

where $[\,]$ denotes the concatenate operation. Finally, X_0 and X_t are fed into a segmentation module composed of a convolution operation to change the number of channel, a 1×1 convolution and a up-sample operation to obtain segmentation results S_0 and S_t.

The weighted binary cross entropy loss and L1 loss are used to train our entire network.

$$L_C(S, O) = -\frac{1}{n}\sum_{x=1}^{n}(1-\eta)o_x \log(s_x) + \eta(1-o_x)\log(1-s_x)$$
$$L_1(S, O) = \frac{1}{n}\sum_{x=1}^{n}|o_x - s_x|$$
$$L = \lambda L_C + L_1 \quad (6)$$

where $O \in \{0, 1\}^{W \times H}$ is the ground-truth, $S \in \{0, 1\}^{W \times H}$ is the segmentation result, s_x is the prediction at pixel x, η is the foreground-background pixel number ratio, and λ is set to 0.8.

III. Experiments

A. Training and Testing

We utilized the DeepLabv3 [2] provided by the COSNet [11] as our initial model. We have two training strategies, one is to train our network only with DAVIS-2016 [13] dataset, and the other is first to use MSRA10K [2] and DUT [2] to fine-tune our feature encoder network (DeepLabv3), and then take the advantage of DAVIS-2016 [13] to train the entire model. The first training strategy only obtained the performance of 79.7 in J and 77.9 in F. Therefore, we took the second training strategy. In the training phase, the current frame F_t and the first frame F_0 are fed into MANet. The size of our input frames was set to $473 \times 473 \times 3$, and the batch size was set to 5. Our network was optimized by the SGD optimizer with an initial learning rate of 2.5×10^{-4}. We utilized PyTorch to implement our network. All experiments were conducted on a 32 G NVIDIA Tesla V100-SXM2 GPU. In the testing phase, we used the DAVIS-2016 and SegTrack v2 [8] datasets to test our network. First, the current frame and the first frame were fed into our MANet to generate segmentation results. And then, CRF was utilized as a post-processing step to obtain our final results. To evaluate our results, we took the region similarity J and contour accuracy F as our evaluation metrics.

TABLE I
Ablation Study

Comparison of different attention modules.			
	Deep layers	Shallow layers	Deep layers + Shallow layers
J mean	81.4	80.9	**82.1**
F mean	80.3	79.2	**80.5**
Comparison of different reference frames			
	First frame	Previous frame	Random frame
J mean	**82.1**	80.2	80.2
F mean	**80.5**	78.8	79.0

B. Ablation Study

Comparison of different attention modules. As mentioned, we design a multi-attention module based on the features from the shallow layers (Conv3_x) and the deep layers (ASPP). To demonstrate the effectiveness of the proposed multi-attention module, we design two other attention modules. One is based on the deep layers, the other is based on the shallow layers. Fig. 3 shows the VOS results of different attention modules. It can be found that the deep layers can basically capture the primary object, but the details of the object cannot be handled well. And we can observe that the shallow layers can preserve the object details, but the background was wrongly classified as the primary object. However, combining the deep layers with the shallow layers can perform well between background removal and preserving the object details. Table I shows that the proposed multi-attention module can obtain the highest mean J and mean F, proving the effectiveness of our multi-attention module.

Comparison of selecting different reference frames. The inputs of our network consist of the current frame and reference frame. And our multi-attention is computed according to the features of the current frame and reference frame. The reference frame can be any frame in the same video of the current frame. In our network, we take the first frame as the reference frame. To investigate the effectiveness of the frame selection strategy, we took the first frame, the previous frame of the current frame, and a random frame as the reference frame respectively to train our whole network, and compared the VOS results of selecting different reference frames. Fig. 4 shows the comparison results. It can be observed that taking the first frame as the reference frame can achieve better VOS results than that of the previous frame and of the random frame. The previous frame and the random frame are inclined to capture short-term temporal dependencies and are prone to accumulate errors. Taking the first frame as the reference frame helps to obtain the global information and obtain satisfactory results. Table I shows that our network with the first frame as the reference frame can achieve higher mean J and mean F than those of the other two frames.

C. Comparison With the State-of-The-Art Methods

To prove the effectiveness of the proposed method, we compared our methods with several deep learning based VOS methods including three classical semi-supervised VOS methods

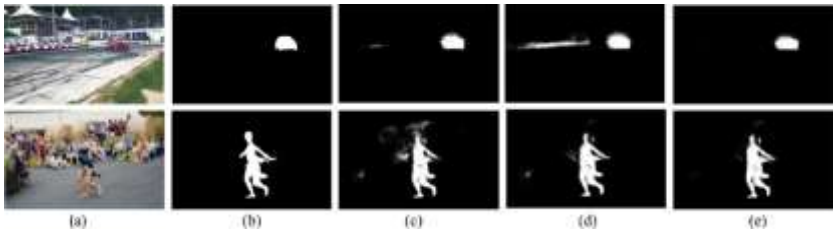

Fig. 3. Segmentation results of different attention modules. (a) Original frame; (b) Ground truth; (c) Segmentation results with feature representations from the deep layers to design the attention module; (d) Segmentation results with feature representations from the shallow layers to design the attention module; (e) Segmentation results with feature representations from the deep layers and the shallow layers to design the attention module.

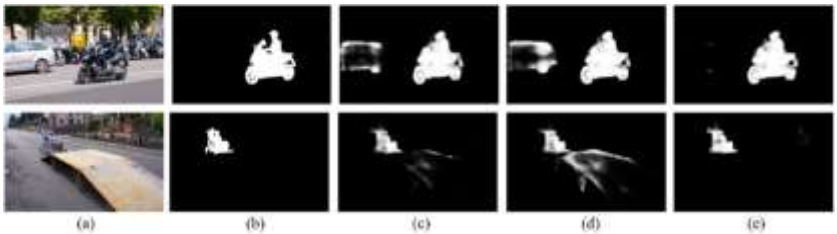

Fig. 4. Segmentation results of selecting different reference frames. (a) Original frame; (b) Ground truth; (c) Segmentation results taking the random frame as the reference frame; (d) Segmentation results taking the previous frame as the reference frame; (e) Segmentation results taking the first frame as the reference frame.

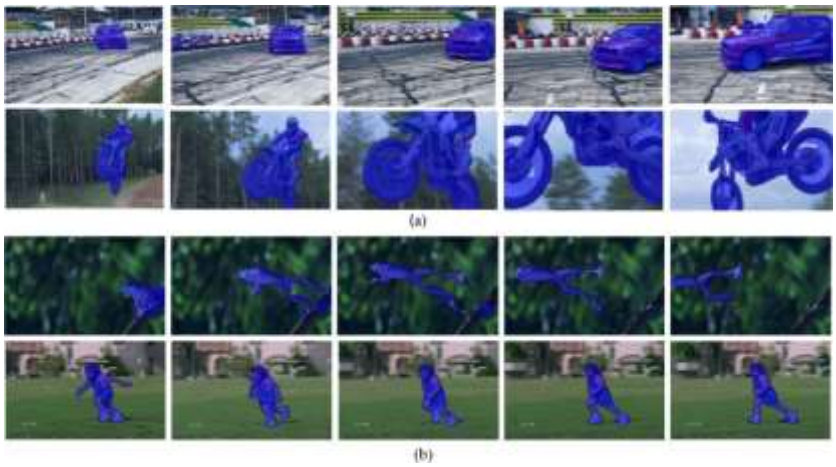

Fig. 5. Segmentation results obtained with our MANet on different datasets. (a) Segmentation results on DAVIS-2016 validation set; (b) Segmentation results the SegTrack v2 dataset.

TABLE II
COMPARISON OF DIFFERENT METHODS

	Method	$\mathcal{J}$ mean	$\mathcal{F}$ mean
Semi-supervised	OSVOS [1]	79.8	80.6
	MSK [12]	79.7	75.4
	PLM [14]	75.5	79.3
Unsupervised	AGNN [17]	80.7	79.1
	COSNet [11]	80.5	79.4
	AD-Net [21]	81.7	80.5
	Our network (w/o CRF)	80.3	**81.3**
	Our network (w/ CRF)	**82.1**	80.5

(OSVOS [1], MSK [12], PLM [14]) and three state-of-the-art UVOS methods (AGNN [17], COSNet [11], AD-Net [21]) on the DAVIS-2016 dataset. Table II shows the experimental results of different VOS methods. We can see that the proposed MANet can obtain higher performance than those of the state-of-the-art UVOS methods, while also performing better than those of the classical semi-supervised VOS methods. The effectiveness of MANet can also be observed in Fig. 5 that visualizes the VOS results of MANet on DAVIS-2016 and SegTrack v2 datasets.

IV. CONCLUSION

In this correspondence, we put forth a multi-attention network (MANet) for unsupervised video object segmentation that mines related information from both the deep and shallow layers. Additionally, a multi-attention module should be designed based on the relationship between the information in the shallow and deep levels. This module aids in identifying the main object and handling its details. We evaluated the performance of our network using the DAVIS-2016 dataset in order to confirm the efficacy of the suggested network. According to experimental results, our network can outperform state-of-the-art techniques in terms of results.

REFERENCES

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 221–230.

[2] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[3] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[4] M. M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[5] Y. T. Hu, J. B. Huang, and A. G. Schwing, "Videomatch: Matching based video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 54–70.

[6] Y. T. Hu, J. B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 786–802.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–455, 2015.

[8] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, " Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2192–2199.

[9] X. Li and C. Change Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 90–105.

[10] S. Li *et al.*, "Instance embedding transfer to unsupervised video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6526–6535.

[11] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3623–3632.

[12] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2663–2672.

[13] 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.*, 2016, pp. 724–732.

[14] J. Shin Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. So Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2167–2176.

[15] G. Song and K. M. Lee, "Bi-directional seed attention network for interactive image segmentation," *IEEE Signal Process. Lett.*, vol. 27, pp. 1540–1544, 2020.

[16] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L. C. Chen, "Feelvos: Fast end-to-end embedding learning for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9481–9490.

[17] W. Wang, H. Song, S. Zhao, J. Shen, S. C. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3064–3074.

[18] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 701–716.

[19] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, 2020, Art. no 36.

[20] J. Zhang, K. Li, and W. Tao, "Multivideo object cosegmentation for irrelevant frames involved videos," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 785–789, Jun. 2016.

[21] Y. Zhao, Q. Wang, L. Bertinett, W. Hu, S. Bai, and P. H. Torr, "Anchor diffusion for unsupervised video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 931–940.

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

[23] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "Motion-attentive transition for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 8326–8338, 2020.