Dual-stream Co-enhanced Network for Unsupervised Video Object Segmentation

Hongliang Zhu[1,3], Hui Yin[1,3,4*], Yanting Liu[1,3,4], and Ning Chen[2,3]
[1]Beijing key lab of traffic data analysis and mining, Beijing Jiaotong University, Beijing 100044, China
[2]Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing 100044, China
[3]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[4]Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University, Beijing 100044, China
[e-mail: hyin@bjtu.edu.cn]
[*]Corresponding author: Hui Yin

Received March 17, 2021; revised November 18, 2022; accepted March 11, 2024; published April 30, 2024

Abstract

Unsupervised Video Object Segmentation (UVOS) is a highly challenging problem in computer vision as the annotation of the target object in the testing video is unknown at all. The main difficulty is to effectively handle the complicated and changeable motion state of the target object and the confusion of similar background objects in video sequence. In this paper, we propose a novel deep Dual-stream Co-enhanced Network (DC-Net) for UVOS via bidirectional motion cues refinement and multi-level feature aggregation, which can fully take advantage of motion cues and effectively integrate different level features to produce high-quality segmentation mask. DC-Net is a dual-stream architecture where the two streams are co-enhanced by each other. One is a motion stream with a Motion-cues Refine Module (MRM), which learns from bidirectional optical flow images and produces fine-grained and complete distinctive motion saliency map, and the other is an appearance stream with a Multi-level Feature Aggregation Module (MFAM) and a Context Attention Module (CAM) which are designed to integrate the different level features effectively. Specifically, the motion saliency map obtained by the motion stream is fused with each stage of the decoder in the appearance stream to improve the segmentation, and in turn the segmentation loss in the appearance stream feeds back into the motion stream to enhance the motion refinement. Experimental results on three datasets (Davis2016, VideoSD, SegTrack-v2) demonstrate that DC-Net has achieved comparable results with some state-of-the-art methods.

Keywords: Unsupervised video object segmentation, Dual-stream Co-enhanced, Motion refinement, Feature aggregation, Dual-stream neural network.

This work is supported by the Fundamental Research Funds for the Central Universities (Science and technology leading talent team project) (2022JBQY009), National Natural Science Foundation of China (51827813), National Key R&D Program “Transportation Infrastructure”“Reveal the list and take command” project (2022YFB2603302) and R&D Program of Beijing Municipal Education Commission (KJZD20191000402).

http://doi.org/10.3837/tiis.2024.04.007 **ISSN : 1976-7277**

1. Introduction

Video object segmentation is a basic task in the field of computer vision, which has played an important role in video editing [1-2], autonomous driving [3-4] and video surveillance [5-6]. The segmented object is generally the most distinct and primary object in the whole video sequence. The fundamental challenges of video object segmentation lie in the complexity of the video scene and the variety of the objects. Meanwhile, the video usually contains rich motion states and complex background interference, resulting in occlusion, fast moving and appearance variations, which brings serious challenges for accurate and stable object segmentation task.

In this paper, we focus on the Unsupervised Video Object Segmentation (UVOS) task， where no manual annotation about the target object to be segmented is provided. In addition to the common challenges of the video data itself mentioned above, UVOS does not require any human involvement in the test phase. Therefore, accurately locating the most prominent objects in the entire video sequence will be even more challenging.

In UVOS, motion cues are very important as the saliency of the target object in a video depends not only on the appearance but also on its continuous movement in successive frames. And motion cues also provide high discriminable features for the confusion caused by other surroundings which have similar colors, textures with the target object. However, most previous approaches [7-13] using motion cues as auxiliary information to help improve segmentation performance are often unsatisfactory when dealing with some complex motion state video sequences. The main reason is that only considering single-direction motion in the video sequences containing complex motion states often leads to inaccurate motion estimation. As shown in **Fig. 1**, we found that in some video sequences with complex environments and changeable motion states, the forward optical flow and backward optical flow of the target have some complementary parts (the dancer's arm in the second row, the woman's right leg in the third row, and the bicycle wheel in the last row).

Furthermore, for current feature extraction framework based on CNNs architecture, the lower level reflects more fine-grained details information (edge and color), while the top layer of the network reflects high-level semantic information which drops out some meaningless or irrelevant detail information. The low-level features and high-level semantic features have different effects on the details and locations of the foreground object, but existing methods do not effectively integrate different level features together, and the high-level semantic features are not fully utilized in the top-down transfer process.

Motivated by the above observations, a novel deep Dual-stream Co-enhanced Network (DC-Net) is proposed in this paper for UVOS. As shown in **Fig. 2**, DC-Net is a dual-stream architecture where the two streams are co-enhanced each other, where the motion stream learns from bidirectional optical flow and produces fine-grained and complete distinctive motion saliency map, and the appearance stream is specifically designed to integrate the different level features effectively. Specifically, the motion saliency map obtained by the motion stream is fused with each stage of the decoder in the appearance stream to improve the segmentation and the segmentation loss in the appearance stream feeds back into the motion stream for the motion cues refinement. Through the co-enhance process, DC-Net can fully take advantage of motion cues and effectively integrate different level features to produce high-quality segmentation mask.

In summary, our contribution can be summarized as follows:

1. A novel deep dual-stream co-enhanced network (DC-Net) is proposed for UVOS task, which can effectively integrate appearance and motion features, and generate high-quality

segmentation masks in a co-enhance process.
2. In order to make up for the insufficient amount of appearance information and improve the problem of inaccurate motion estimation in single-direction optical flow, we propose a Motion Refine Module (MRM), which makes full use of forward and backward optical flow information to learn more complete motion cues of the target object, thereby making the segmentation results more complete.
3. In order to enhance the attention to semantic information in the segmentation process and avoid the loss of the information of different scales in the process of propagation, we propose a Context Attention Module (CAM) to enhance the representation of high-level semantic features and model the relationships among multiple salient objects. Then through the designed Multi-level Feature Aggregation Module (MFAM), the low-level features, high-level features and enhanced high-level features can be effectively integrated to improve the segmentation effect.
4. Extensive experiments are conducted on the Davis2016 [14] dataset to verify the effectiveness of the key modules of our proposed network DC-Net as well as its own. At the same time, we also conducted experiments on these two datasets, VideoSD [15] and Segtrack-v2 [16]. The results on them again demonstrate the effectiveness of DC-Net.

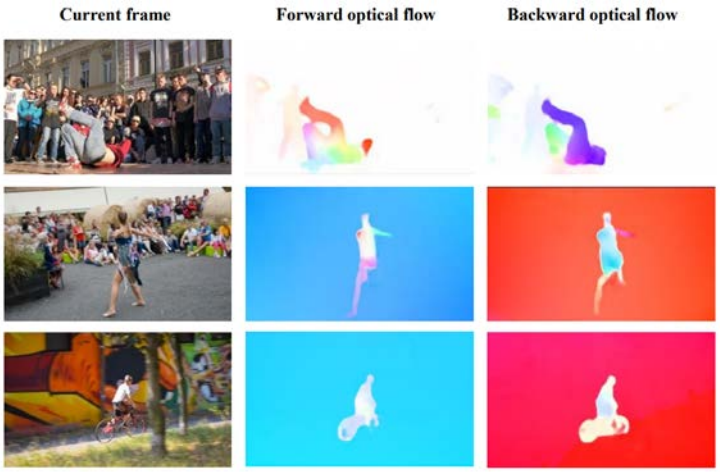

Fig. 1. Forward and backward optical flow images of several video frames on DAVIS-2016 dataset. The forward optical flow is calculated from the displacement of pixels from the previous frame to the current frame, and the backward optical flow is calculated from the displacement of pixels from the next frame to the current frame.

2. Related Work

According to how much information about the target object is given, the problem of video object segmentation is usually divided into unsupervised and supervised (including fully-supervised, semi-supervised and weak supervised) video object segmentation. In this paper, we mainly solve the Unsupervised Video Object Segmentation (UVOS), which extracts the most salient object mask without any manual annotation.

Unsupervised video object segmentation (UVOS). Unsupervised video object segmentation is a challenging task, which doesn't need any annotation of the video sequence but assumes that the object to be segmented has some salient features, for example, moving objects. Before the well-known CNNs structure appeared, the traditional methods mainly relied on hand-

crafted features to segment the primary object in video sequences, such as long-term point trajectory [17]-[21], motion boundary [22], objectness [23]-[27], and saliency [7, 28-30].

In recent years, there have been many methods of constructing CNNs for UVOS [7-12, 31-34]. In [9-12, 33], a CNN architecture is built with appearance model or motion information to segment the primary object in videos. For example, [9] used the artificially synthesized dataset and input the optical flow images obtained from two consecutive frames into the designed CNN architecture to obtain the motion label of each pixel. [11] introduced a dual-stream network with appearance and motion cues for object segmentation. A visual memory module composed of convolutional Gated Recurrent Units [35] is designed in [11] to process those foreground objects whose initial state of the video sequence is still. FSEG [10] also designed a dual-stream fully convolutional neural network and fused motion and appearance features at the end of the network to generate segmentation mask. [31] used pyramid dilated bidirectional ConvLSTM architecture to solve video salient object detection task, and apply it to the UVOS. EPO [33] combined geometric constraints with CNN to convert optical flow into long-term point trajectories to segment the main objects in the video. [34] used the teacher-student learning paradigm for UVOS. Most of the methods mentioned above take into account the appearance and temporal motion of prominent targets in video sequences, and design useful structures to exploit these features fully to perform accurate segmentation, which shows that the effective use of appearance and temporal motion information can improve the segmentation accuracy. Different from the methods using appearance and motion cues individually or separately, we try to effectively integrate appearance and motion feature streams, and generate high-quality segmentation masks by a co-enhance strategy.

Dual-stream network for UVOS with motion and appearance cues. The dual-stream network architecture for UVOS is to allow the two streams to play their respective advantages for different inputs, and then to effectively integrate the features of the two streams. FSEG [10] fused an appearance stream and a motion stream for object inference and then performs the fusion operation of the two streams at the end of the network to obtain the final segmentation mask. It simply treats motion cues as equal to appearance information, without considering the characteristics of different input data. To learn the spatial-temporal feature, [11] utilized a dual-stream architecture and attempt to joint two streams via concatenation and a convolutional visual memory module. However, RNN-based methods are not conducive to modeling long-term video sequences and require a lot of computation and memory usage. [13] employed salient motion detection and object proposals techniques for unsupervised video object segmentation, but its appearance features are extracted using a well pretrained Mask R-CNN [36] model, and the segmentation performance is very dependent on the reliability of the appearance model. EPO [33] exploited multiview geometric constraints combining epipolar distances with optical flow to define motion saliency and used a common appearance model to extract appearance features. These existing methods only consider single-direction optical flow to model motion saliency, which leads to insufficient learned motion information. In this paper, we take bidirectional optical flow into account to fully learn the motion cues of salient objects in a video sequence. In addition, different from the above dual-stream methods for UVOS, our proposed network designs special models to improve the feature integration capability, and then further optimizes the network through co-enhanced manner to improve the segmentation performance of the whole network.

Multi-level features aggregation. In general, all level features have different contributions to video object segmentation. In CNNs structure, feature maps from low-layers encode low-level details information such as color and edge information, which is very beneficial for improving the segmentation accuracy. However, such features typically contain more noises and require

further processing. While top layers encode high-level context and semantic features help integrate global context into the network. Therefore, fusing different levels of features are commonly used in object saliency segmentation models to obtain more accurate mask [37-41]. Recently, many works [42-45] show that the importance of multi-level feature aggregation effectively and sufficiently for their task. For example, [42] combined several prediction outputs obtained by fusing some specific side outputs with short connections. [45] first integrated multi-level feature maps into multiple resolutions and then adaptively learned to combine these feature maps at each resolution. There are also many methods that use the skip-layer architecture like U-shape network [46] due to its simplicity and effectiveness in the task of saliency target segmentation. However, most of methods based on U-shape architecture usually fuse low-level and high-level features with concatenation and addition directly, resulting in limited feature representation and robustness. Furthermore, although CNNs can restore the original feature map resolution by using up-sampling operations in the decoder, the detailed spatial context lost during the down-sampling process cannot be fully recovered, and the high-level semantic features are not fully utilized in the top-down transmission process.

To remedy these problems mentioned above, in this paper, we use a channel attention mechanism to enhance the valuable channels in high-level semantic features and to catch the global context features while suppressing some useless information. Then pass it to each stage of the decoder for effective aggregation with different level features, leading to improved learning capability.

3. Proposed method

In this part, we first outline our proposed DC-Net, and then clarify details of each component we designed. Finally, we will introduce our training schema and loss function.

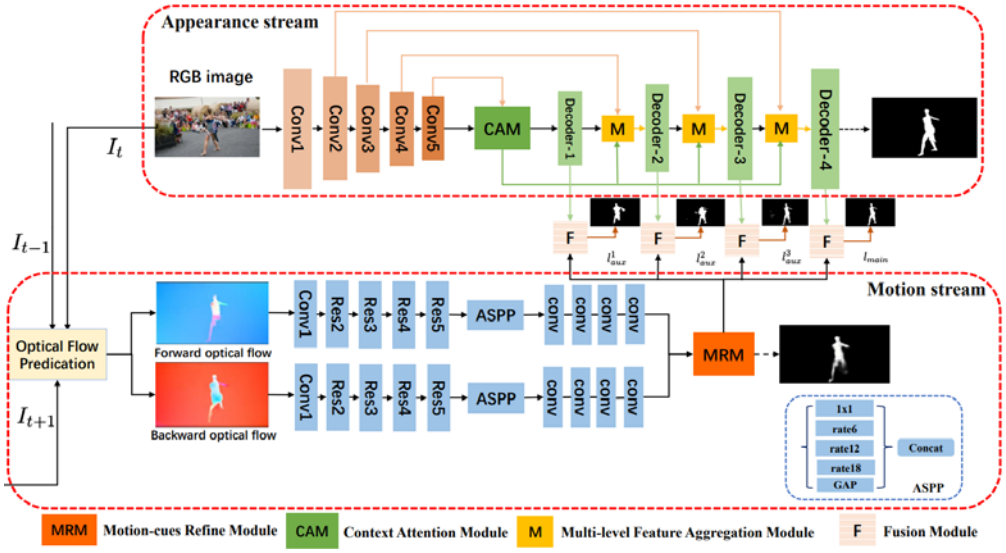

Fig. 2. Pipeline of DC-Net. The frame I_t is input into the appearance stream, and the bidirectional optical flow predicted by the former frame I_{t-1}, the next frame I_{t+1} and I_t are input into the motion stream. The two streams are combined to generate the final segmentation result through fusion modules.

3.1 Proposed network framework

As illustrated in **Fig. 2**, DC-Net is a deep two-stream neural network for UVOS, consisting of an appearance stream, a motion stream and some fusion modules. The input of the appearance stream is a static image, and the appearance stream consists of two modules, Context Attention Module (CAM) and Multi-level Feature Aggregation Module (MFAM). The input of the motion stream is the forward and backward optical flow images [47]. After extracting the respective features, the refined motion saliency map is obtained by the Motion-cues Refine Module (MRM). Fusion Module (FM) is designed to obtain the fine-grained segmentation result by combining the information of these two streams via a co-enhance strategy.

3.2 Appearance stream

The appearance stream is an encoder-decoder structure similar to U-Net [46], using ResNet-50 [58] as backbone of the encoder. The feature map is continuously down-sampled through the convolutional layer and pooling layer, and high-level semantic features are gradually obtained at the end of the encoder.

First of all, in order to infer the relationship between the semantics of different salient objects or regions from a global perspective, and because receptive field of the high-level feature is relatively large and can better reflect the semantic information of the image, we first design the Context Attention Module (CAM) and applied it to the top-level features (the output of the last layer of the encoder). In this way, it is used to further enhance the high-level semantic features and model the relationship between the salient objects, ultimately helping to generate a more complete saliency map.

Features at different scales contain different information of the image. Generally speaking, the low-level features mainly contain details such as edges and colors of the image, and the high-level features are more concerned with the semantic information of the image. Moreover, the U-Shape networks make low-level features and high-level features to directly concatenate to fuse features of different scales. Unlike them, in order not to lose any information during the decoding process, we propose the Multi-level Feature Aggregation Module (MFAM) to effectively integrate low-level features, high-level features and enhanced high-level features (the output of CAM) to improve feature representation performance and the segmentation effect. As follows, we will introduce these proposed modules CAM and MFAM in detail.

3.2.1 Context Attention Module

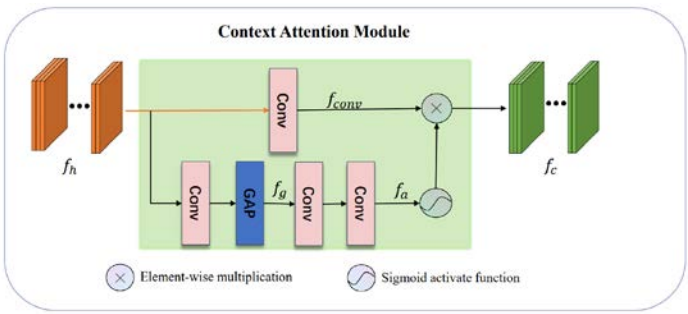

Fig. 3. Context Attention Module

Our proposed CAM is shown in **Fig. 3**. We firstly use a convolution operator and employ global average pooling [50] to obtain the global contextual information, and then enhance the

response of the corresponding channel according to the contribution of different channels to the salient features. Specifically, the process can be described as

$$f_g = GAP(\theta_1(f_h)) \tag{1}$$

$$f_{conv} = \theta_2(f_h) \tag{2}$$

$$f_a = \sigma\left(\theta_4\left(\theta_3(f_g)\right)\right) \tag{3}$$

$$f_c = f_{conv} \times f_a \tag{4}$$

where f_h refers to the high-level features from top layer, and f_g refers to the feature that the high-level features after global average pooling GAP processing. f_g includes global context samantic information, σ is a sigmoid activation function, θ_i $(i = 1,2,3,4)$ refer to 1×1convolution operation, × denotes element-wise multiplication. To enhance the saliency response of target objects and to alleviate the problem of insufficient exploit of high-level semantic features. The output feature f_c after CAM will be passed to the various MFAMs of the decoder, which will be elaborated in the next part.

3.2.2 Multi-level Feature Aggregation Module

We introduce a multi-level feature aggregation module (MFAM) to fully integrate low-level features, high-level features, and global context semantic features to improve segmentation performance, which is illustrated in **Fig. 4**. We take the output of the previous layer in the decoder as the high-level feature f_h^i $(i = 1, 2, 3)$, the low-level f_l^i $(i = 1, 2, 3)$ feature is the shallow layer feature of the corresponding encoder, and the global context semantic feature f_c is the output of the CAM.

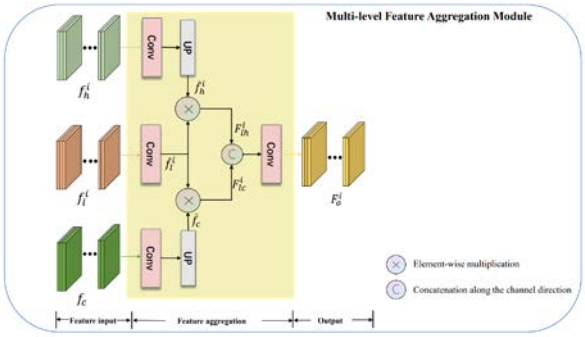

Fig. 4. Multi-level Feature Aggregation Module

Specifically, we first compress the low-level features through a 1×1 convolution layer θ_5 so that it has the same number of channels as the corresponding high-level features, and then the high-level features are fed into a 3×3 convolution layer θ_6. After up-sampling, obtain an attention saliency map $\hat{f}_h^i$ with semantics. Further, we apply element-wise multiplication on $\hat{f}_h^i$ and the compressed low-level feature $\hat{f}_l^i$. Similarly, we perform the mirror operation between low-level features and the output f_c from CAM. Finally, concatenate the two output feature maps, and through a 3×3 convolution layer θ_8 to produce the final aggregation features. The above process can be described as

$$\hat{f}_l^i = \theta_5(f_l^i) \tag{5}$$

$$\hat{f}_h^i = upsample\left(\theta_6(f_h^i)\right) \tag{6}$$

$$F_{lh}^i = Relu(\hat{f}_l^i \times \hat{f}_h^i) \tag{7}$$

$$\hat{f}_c = upsample\left(\theta_7(f_c)\right) \tag{8}$$

$$F_{lc}^i = Relu\left(\hat{f}_l^i \times \hat{f}_c\right) \tag{9}$$

$$F_o^i = \theta_8\left(concat\left(F_{lh}^i, F_{lc}^i\right)\right) \tag{10}$$

where $\theta_t(t = 5,6,7,8)$ refers to convolution layer, F_{lh}^i indicates the i-th stage comprehensive features that combines low-level and high-level features, F_{lc}^i indicates the i-th stage comprehensive features that combines low-level and global context features, $upsample$ is up-sampling operation via bilinear interpolation, $Relu$ represents the ReLU activation function, $concat$ is concatenation operation, and i is the stage index. F_o^i indicates the aggregated features.

3.3 Motion stream

As is shown in **Fig. 2**, the motion stream is siamese neural network architecture, which is composed of an encoder and a decoder. Considering that the optical flow images do not have rich and detailed appearance feature compares with RGB images, ResNet-34 [48] is used as the encoder in order to accelerate the network inference speed and reduce the number of network parameters. The forward optical flow and the backward optical flow are simultaneously fed into the motion stream, and then the two types of features of the decoder are sent to the Motion-cues Refine Module together to obtain the final single-channel motion saliency map. The ASPP [51] module is introduced to model the long-range dependencies of the feature map and integrates local and global feature representations through a series of dilated convolution operations.

In order to effectively integrate the forward optical flow features and the backward optical flow features, we designed a motion-cues refine module (MRM), as shown in **Fig. 5**, which can make full use of the motion cues of salient objects to produce more complete motion saliency map.

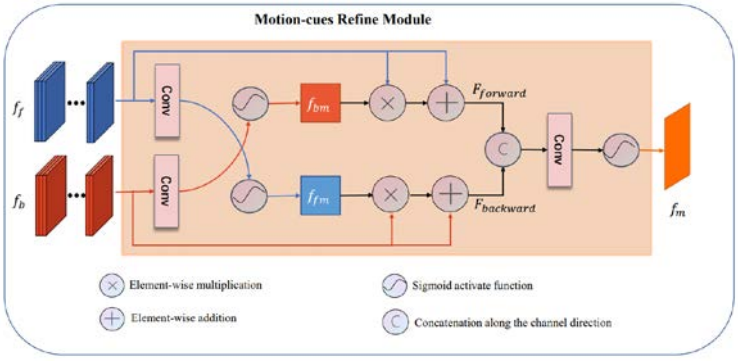

Fig. 5. Motion-cues Refine Module

Fig. 5 shows the Motion-cue Refine Module we proposed. f_f and f_b are the forward optical flow feature and backward optical flow feature through the decoder respectively. Further, each of them is applied a 3×3 convolutional layer and a sigmoid function to get their respective saliency attention maps f_{bm} and f_{fm}. Then, we apply element-wise multiplication between f_{bm} and each channel slice of f_f. This element-wise multiplication is essentially an attention mechanism, which is sample and efficient. However, due to the complexity of the motion and the frame rate of the video data, the corresponding part of the motion part represented in the optical flow may be predicted to be 0. A simple multiplication operation will suppress this part

of the target. In this case, the complete segmentation of moving objects cannot be maintained. Based on the above considerations, we use a skip connection that adds the original feature. With this element-wise addition operation, the motion parts predicted by the optical flow in other directions can be retained without affecting the common prominent targets or parts. Asymmetric operation is also applied to the backward optical flow feature and the saliency attention map generated by the forward optical flow. Finally, concatenate these two features, and the final refined motion saliency map f_m is obtained through a 3×3 convolution and sigmoid activation function. The sigmoid activation function is to compress the pixel value between [0, 1].

The main process can be described as

$$f_{bm} = \sigma\left(\theta_9(f_f)\right) \tag{11}$$

$$f_{fm} = \sigma(\theta_{10}(f_b)) \tag{12}$$

$$F_{forward} = f_f \times f_{bm} + f_f \tag{13}$$

$$F_{backward} = f_b \times f_{fm} + f_b \tag{14}$$

$$f_m = \sigma\left(\theta_{11}\left(concat(F_{forward}, F_{backward})\right)\right) \tag{15}$$

where $\theta_l(l = 9, 10, 11)$ refers to 3×3 convolutional operation, σ is the sigmoid activation function, $+$ denotes element-wise additions. f_m is the final output of the motion stream, which will be fed to our proposed fusion module and effectively fused with the output of each stage of the decoder in the appearance stream. Through our fusion strategy, the final segmentation result of the entire DC-Net is obtained.

3.4 Dual-stream fusion and Co-enhanced strategy

In general, the appearance stream focuses on the appearance features of salient objects in a single frame, such as edges, textures and other details, but it lacks the temporal continuity of foreground objects between frames. The results segmented by the appearance stream may contain non-motion salient parts. While the motion stream mainly concentrates on the motion cues of the object, which can fully extract the motion pattern of the foreground target between frames. But the motion feature does not have the rich detail information as the appearance stream, especially the fine contour details. Both appearance and motion stream can only achieve limited segmentation capabilities. If the dual-stream feature can effectively make up for the shortcomings of the other side, the segmentation performance will have a greater improvement.

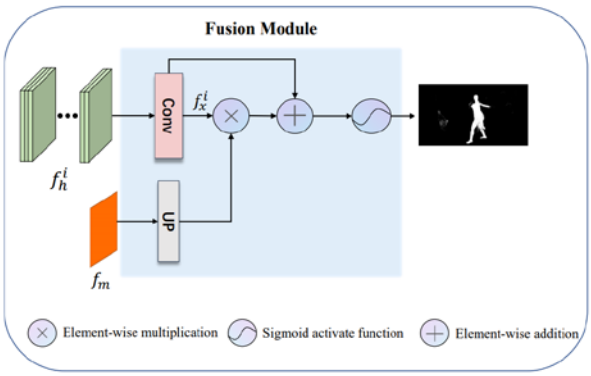

Fig. 6. Fusion Module

Based on the above discussion, we designed a dual-stream Fusion Module to combine dual-stream features in an effective way, as shown in **Fig. 6**. In addition, in order to prevent the single-stream network from overfitting the respective input information, we propose a co-enhance strategy. Through multi-task joint training, the joint loss continuously updates the parameters of the dual-stream network during the backpropagation process, and the appearance stream can be gradually concentrated on the target with salient motion, resulting in more accurate segmentation result. Conversely, the segmentation loss in the appearance stream is fed back into the motion stream to enhance the motion refinement. Thus, the entire DC-Net can further improve the segmentation ability in a co-enhance manner.

In **Fig. 6**, $f_h^i (i = 1,2,3,4)$ indicates i-th stage features of the decoder in appearance stream. f_m refers to the output of motion stream. The fusion strategy can be formulated as:

$$f_x^i = \theta_{12}(f_h^i) \tag{16}$$

$$F_{out}^i = \sigma(f_x^i \times upsample(f_m) + f_x^i) \tag{17}$$

where F_{out}^i $(i = 1, 2, 3, 4)$ indicates i-th stage output, f_h^i is i-th stage features of the decoder in the appearance stream. By the way, before the motion saliency map is sent to each stage of the appearance stream, it needs to be up-sampled ($upsample$) to the same size as the output feature map of each stage of the appearance stream.

We claim that only when both the appearance stream and the motion stream have strong segmentation capabilities, the DC-Net that merges the two streams could have well segmentation performance. In this paper, we used multi-task pipeline training to train our proposed dual-stream network in a co-enhance process. First, we initialize the encoder of appearance stream using a ResNet-50 [48] pretrained on ImageNet [52]-[53], using static-image data to fine-tune the appearance stream. Second, we use PWC-Net [47] to calculate the forward and backward optical flow on our training data, and then use [54] to convert the optical flow into three-channel images. Third, a ResNet-34 [48] model pretrained on ImageNet is used to initialize our motion stream, and then is trained on optical flow images which are synthesized on our video training data. Lastly, the two streams conduct joint training and use our proposed fusion strategy to fuse the motion saliency map and the appearance saliency map. to obtain the final segmentation mask in a co-enhance process.

3.5 Loss Function

By the means of multi-task training schema, we first train the two streams separately to take advantage of the single stream. Then dual stream joint training to further improve the overall segmentation performance.

We use binary cross entropy loss, which is commonly used loss function in the field of video object segmentation. For joint training, our network gets the final segmentation mask after the last fusion module. When training individually, each stream gets its own segment prediction. The binary cross entropy loss can be calculated as

$$L(M, G) = -\frac{1}{N}\sum_x [G_x \log(M_x) + (1 - G_x)\log(1 - M_x)] \tag{18}$$

where N refers to the total number of pixels in the input video frame. G_x represents the ground truth label at the pixel x and the M_x is the corresponding prediction at pixel x. For joint training, our total loss function as follows

$$L_{loss} = l_{main} + \sum_{i=1}^{3} \beta^i \, l_{aux}^i \tag{19}$$

where l_{main} refers the main loss corresponding to the last output in our fusion module and l_{aux} represents the auxiliary loss of the rest fusion module outputs. β^i is the weight of different loss at each fusion module, and the specific settings are $\beta^1 = 0.4,\ \beta^2 = 0.6,\ \beta^3 = 0.8$.

4. Experiments

In this part, we conducted comprehensive studies to evaluate our proposed method. For the appearance stream, we take advantage of static-image saliency datasets: DUTS [55] and MSRA10K [56]. DAVIS-2016 [14] dataset is used to train our motion stream and the whole DC-Net. From the point of view of the size of the model and the amount of computation, we adopt the existing optical flow estimation method PWC-Net [47] to estimate the forward and the backward optical flow in DAVIS-2016 dataset. The forward optical flow is calculated from the displacement of pixels from the previous frame to the current frame, and the backward optical flow is calculated from the displacement of pixels from the next frame to the current frame. We augment all the training data by performing random horizontal flip and randomly clipping the images to the size of 320 ×320. Mini-batch Stochastic gradient descent (SGD) optimizer is used to train the whole DC-Net, we set the initial learning rate is 10^{-4}, weight decay is 0.0005 and momentum is 0.9. We implement the whole DC-Net with Pytorch [57].

4.1 Datasets

The proposed DC-Net is evaluated in the following three Datasets:
DAVIS-2016 [14] is a large-scale data set used for video object segmentation, and all video frames have pixel-level annotations. There are many challenges for UVOS in DAVIS-2016, such as background occlusion, multi-target interference, appearance deformation, fast motion, motion blur.
VideoSD [15] contains 10 low-resolution video sequences under the natural scene.
Segtrack-v2 [16] is another benchmark dataset widely used for video object segmentation. There are 1,066 video frames in total, and each frame also has pixel-level annotation. The main challenges are drastic appearance deformation, uneven lighting, complex motion states, and occlusion.
We use the three general standard metrics proposed in [14], namely the region similarity J, the contour accuracy F and the temporal stability T.

4.2 Effectiveness of the proposed method

We first study the effectiveness of the Motion-cues Refine Module (MRM) and the rationality of the introduction of bidirectional optical flows. The results are shown in **Table 1**. When only using the single-direction optical flow (without MRM), we observe a significant performance drop (mean J: 70.5→56.9, mean F: 61.8→50.4 in DAVIS-2016), and the mean T evaluation index increased by 46.7%. That clearly shows the effectiveness of our Motion-cues Refine Module (MRM), which can fully leverage the motion cues of the foreground object in video sequences.

Table 1. Comparative study of single-direction optical flow and bidirectional optical flow

Measures	Single-direction optical flow (without MRM)	Bidirectional optical flow (with MRM)
J mean	56.9	70.5
F mean	50.4	61.8
T mean	87.3	40.6

Next, we conduct the ablation study on the DAVIS-2016 dataset to verify the effectiveness of the proposed network architecture and key modules. We adopt the model like U-Net [47] as

the baseline model which concatenates high-level features after up-sampling and low-level features directly.

Table 2. Ablation study on Davis-2016. (B: Motion stream with bidirectional optical flow; A: Appearance features enhancement operations (CAM, MFAM); S: Motion stream with single-direction optical flow)

Model	**J Mean**
Baseline	67.7
Basline+B	70.3
Baseline+A	74.1
Baseline+A+S	76.8
Baseline+A+B (DC-Net)	79.6

As shown in **Table 2**, the MRM significantly improved the baseline model, and the J mean value has increased from 67.7% to 70.3%. The third model (Baseline+A), which uses the appearance alone, only use the main loss which is calculated by the cross entropy between the output of fusion module in the appearance stream and the ground-truth label. As it has no motion stream, so no auxiliary loss could be used. The fourth (Baseline+A+S) and the fifth model (Baseline+A+B (DC-Net)) are all composed of appearance stream and motion stream, the difference between these two models is that the former uses single-direction optical flow while the later one uses bidirectional optical flow. Both of them are jointly trained by the main loss and the auxiliary loss. It can be seen that from both models with joint loss (the fourth and the fifth model) are better than the model only use main loss (the third model), and the J Mean increased from 74.1% to 76.8% , 79.6% respectively, which prove the effectiveness of the joint loss and validate the proposed co-enhance strategy. Thus, we attribute this to the well-designed dual-stream co-enhanced network, which includes effective feature aggregation modules and full utilization of bidirectional motion cues.

4.3 Comparative experiments on DAVIS-2016

We compare our network DC-Net with some SOTA methods on DAVIS-2016 dataset, including ARP [25], FST [22], MSG [20], KEY [26], FSEG [10], LMP [9], PDB [31], UOVOS [13], LVO [11], MotAdapt [34], LSMO [12], EPO [33].

Table 3. Quantitative results on DAVIS-2016

Measure		ARP	FST	MSG	KEY	FSEG	LMP	PDB	UOVOS	LVO	MotAdapt	LSMO	EPO	Ours
J	Mean↑	76.3	57.5	54.3	59.6	71.6	69.7	77.2	77.8	75.9	77.2	78.2	80.6	79.6
	Recall↑	89.2	65.2	63.6	67.1	87.7	82.9	90.1	93.6	89.1	87.8	89.1	95.2	92.3
	Decay↓	3.6	4.4	2.8	7.5	1.7	5.6	0.9	2.1	0.0	5.0	4.1	2.2	3.0
F	Mean↑	71.1	53.6	52.5	50.3	65.8	66.3	74.5	72.0	72.1	77.4	75.9	75.5	76.9
	Recall↑	82.8	57.9	61.3	53.4	79.0	78.3	84.4	87.7	83.4	84.4	84.7	87.9	87.0
	Decay↓	7.3	6.5	5.7	7.9	4.3	6.7	-0.2	3.8	1.3	3.3	3.5	2.4	3.3
T	Mean↓	35.9	29.3	26.3	21.0	29.5	68.8	29.1	33.0	26.5	27.9	21.2	19.3	24.5

As shown in **Table 3**, we can see that DC-Net is superior to most methods on DAVIS-2016 benchmark. Compared with the best method EPO [33], although the relevant indicators of our J mean are slightly lower, our F mean is higher than that of EPO, with an increase of 1.4%.

In **Table 3**, some of the other dual-stream based methods also using leverage both appearance and motion cues, e.g. FSEG [10], LVO [11], MotAdapt [34], UOVOS [13]. Our method has a great transcendence than all of these methods. The main reason lies in that these algorithms only use single-direction optical flow, without fully considering the motion cues of foreground objects in the video sequence. In contrast, by introducing the forward and backward optical flow and the fusion in the decoding part, our method can combine the motion information and the appearance information more effectively to obtain better feature representation. Thanks to our proposed MRM, more motion information can be exploited so that our method can achieve great performance.

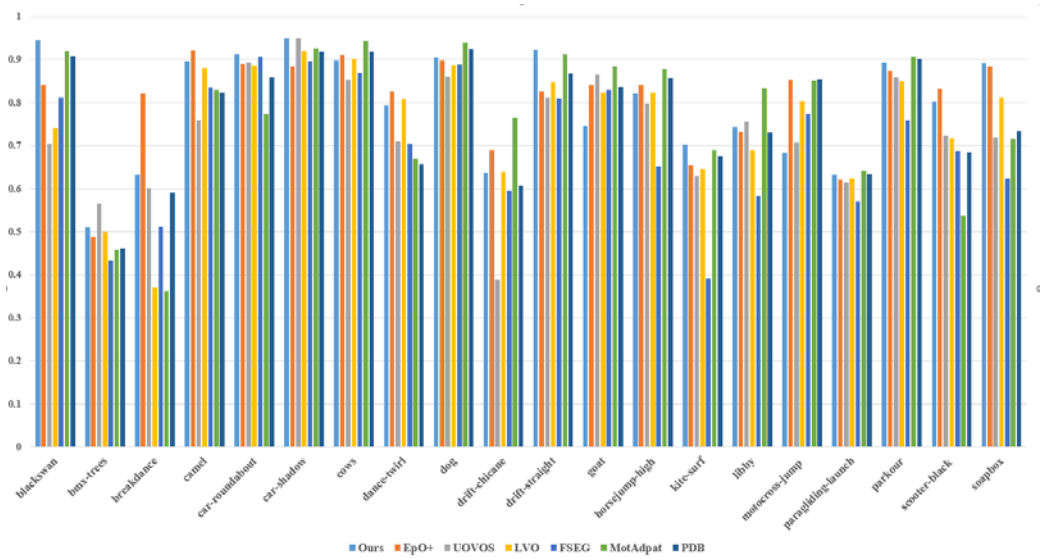

Fig. 7. Per-sequence results of region similarity on DAVIS-2016: Regional similarity results in each video sequence with six unsupervised methods, where light blue represents our results.

Fig. 7 shows the results of our regional similarity comparison on each sequence with the same methods using appearance and motion saliency or other temporal information. From **Fig. 7** we can see that our method achieved much better scores than other UVOS methods in some video sequences, such as BLACKSWAN, CAR-ROUNDABOUT, CAR-SHADOW, DRIFTCHICANE and SOAPBOX sequences. These sequences have dynamic background interference and fast object moving challenges. That proves the advanced nature of our proposed method to meet these challenges.

4.4 Evaluation on VideoSD and Segtrack-v2

For completeness, we also perform experiments on the VideoSD [15] and SegTrack-v2 [16] datasets. The results are shown in **Table 4**. DC-Net performs better (77.1% in term of mean J) than [22, 7, 58, 59, 13] on VideoSD dataset. Especially for UOVOS [13] which achieved the best results currently in existing methods, we outperform it by 11.9%. The primary challenge of this dataset is that the resolution is very low and there is motion blur. The results show that our proposed network has well adaptability in these situations.

The evaluation result of DC-Net on SegTrack-v2 [16] is shown in **Table 5.** Compared with the state-of-the-arts, DC-Net can also rank in the forefront on completely unfamiliar dataset. Removing birdfall and worm, the only sequences we perform poor, the results improve to 71.5%. In birdfall and worm video sequences, there is extremely complex background

interference, and the scale of the foreground target is very small, which makes our appearance stream unable to detect this situation well.

Table 4. Quantitative results on VideoSD over Mean J

Method	FST	SAGM	TIS	OSVOS	UOVOS	Ours
Mean J	61.6	49.7	61.6	44.7	65.2	77.1

Table 5. Quantitative results on Segtrack-v2. For the two video sequences of birdfall and worm, our method does not perform well. If these two results are removed, the overall performance will be improved to 71.6.

Method	KEY	FSEG	LVO	FST	LSMO	UOVOS	EPO	Ours
Mean J	57.3	61.4	57.3	53.5	59.1	64.3	70.9	62.8

4.5 Qualitative Results

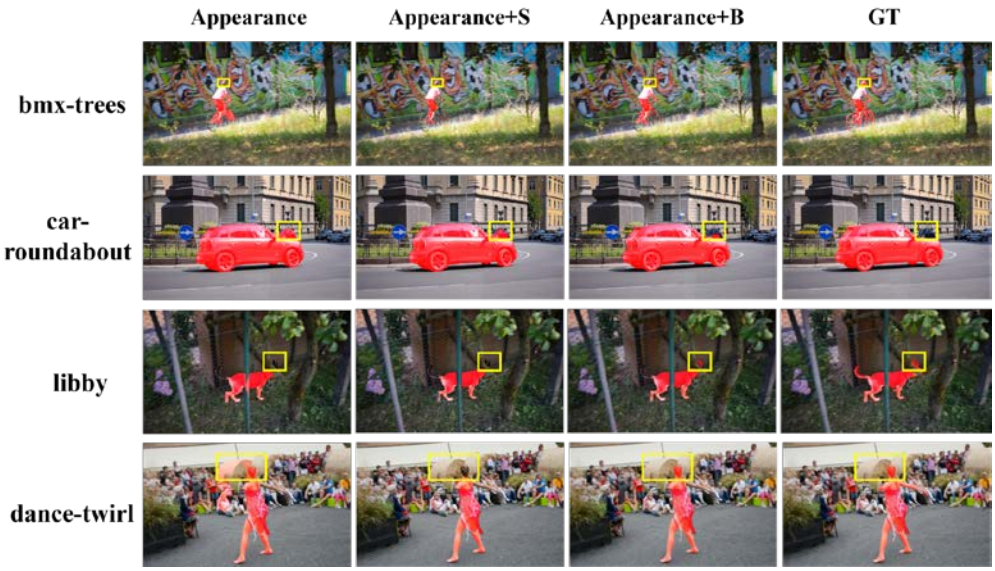

Fig. 8. Qualitative results of ablation study. *Appearance*: visual result of appearance stream. *Appearance+S*: The result obtained by using only single-direction optical flow in the motion stream. *Appearance+B*: The result of using bidirectional optical flow in the motion stream, that is, the proposed DC-Net. *GT*: The ground truth sementation corresponding to the video frame.

Fig. 8 is the qualitative results of the ablation study. From **Fig. 8**, we can see that the introduction of bidirectional optical flow can indeed deal with some challenges such as background occlusion (the head of libby in the third rocw), multiple salient objects (the small car in the second row and haystack in the fourth row), and achieve better results.

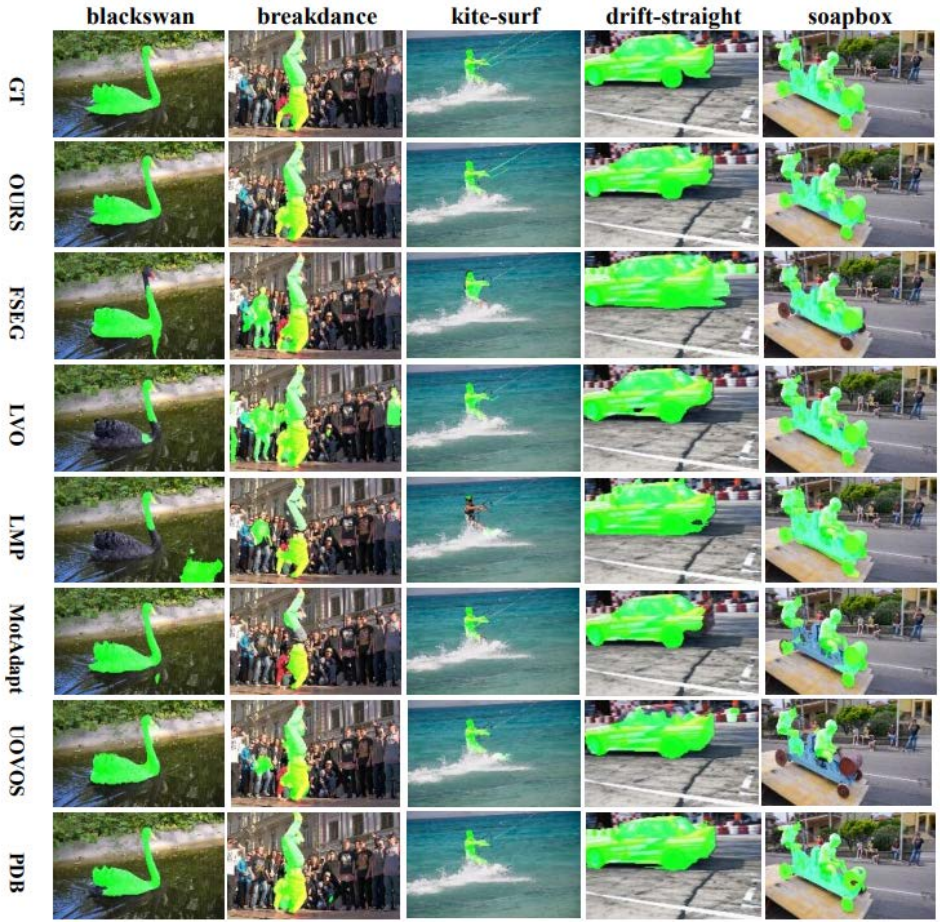

Fig. 9. Qualitative comparison with state-of-the-art methods on DAVIS-2016. The comparison of the partial results of the six unsupervised method. The first row: the groundtruth of the given video frame. The rest is the segmentation results of the given video frame of each video sequence from ours and the other six unsupervised methods.

We also provide some visual results of different mainstream algorithms to prove the advanced nature of our method. As shown in **Fig. 9**, the proposed DC-Net can deal with various challenging scenarios, including dynamic background, fast-motion, motion blur, low resolution, appearance change, interacting objects. etc. The sequences of kite-surf and breakdance, which existing dynamic background (waving water in kite-surf sequence, for breakdance sequence, there are people jumping and applauding in the background), some state-of-the-art methods like FSEG [10], LVO [11], LMP [9], UOVOS [13] regard the background area as the foreground segmentation, resulting in poor results. In drift-straight video sequence, there are scenes with fast motion and motion blur, which often leads to inaccurate motion estimation, that is, poor optical flow quality. As shown in the fourth column of **Fig. 9**, our method can achieve results close to the ground truth. Overall, our method can cope with these challenges and segment the most salient object with complete and accurate mask.

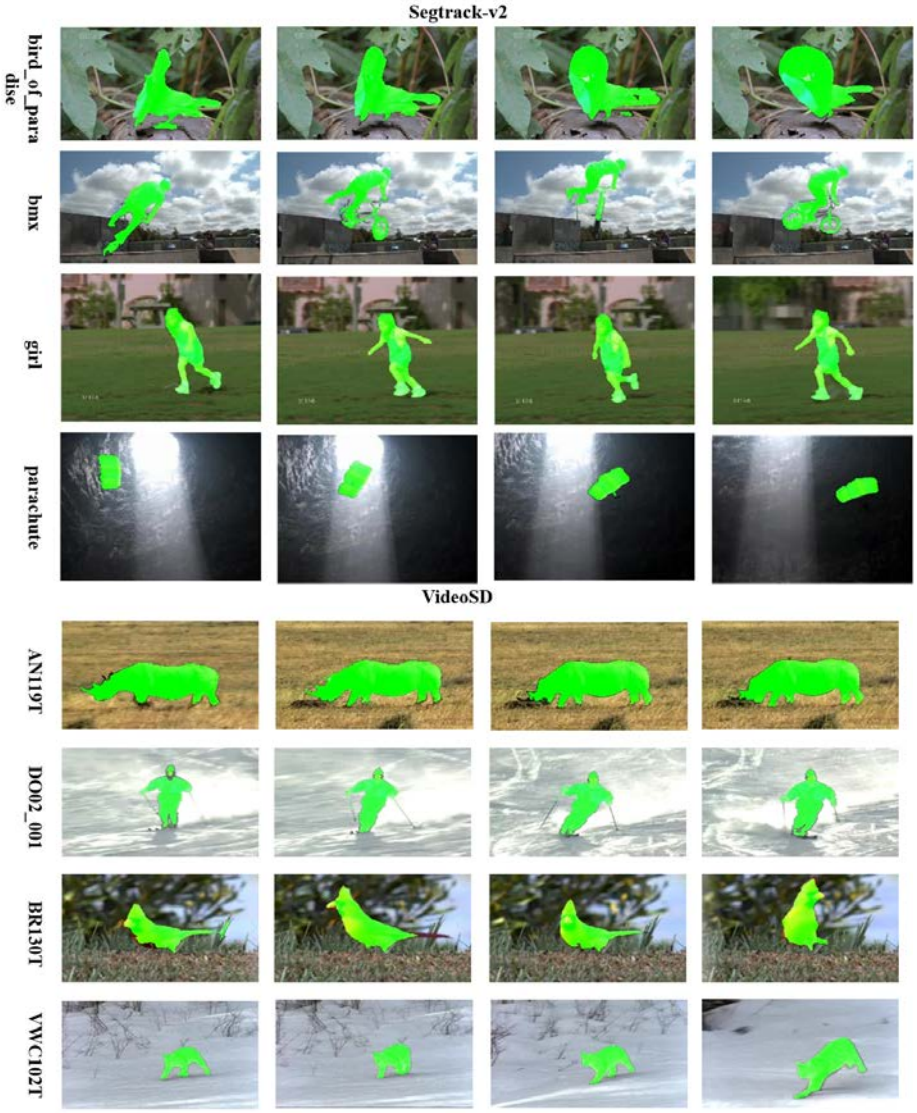

Fig. 10. Visual results on SegTrack-v2 and VideoSD dataset. The first to fourth rows are the four video sequence frames of the SegTrack-v2 dataset, and the remaining four rows are the visual segmentation results of the partial sequences of the VideoSD dataset.

Fig. 10 shows the visual results of partial sequences on VideoSD and SegTrack-v2 datasets, which are completely unfamiliar datasets, and not used in any training process for our network. VideoSD and SegTrack-v2 dataset contain many challenging cases, such as appearance variation (the bird_of_paradise and bmx sequences in SegTrack-v2, BR130T and DO02_001 sequences in VideoSD), background clutter (AN119T and BR130T sequences in VideoSD), uneven light changes (parachute in SegTrack-v2), we can see that our method can effectively handle these situations.

5. Conclusion

In this paper, a novel deep Dual-stream Co-enhanced Network (DC-Net) is proposed for UVOS. A motion cues refine module is designed to generate a more complete and accurate motion saliency map based on bidirectional optical flow. A context attention module and a multi-level features aggregation module are designed in appearance stream to integrate different level information and produce high-quality appearance saliency features. And the two streams are effectively fused to obtain the final segmentation result in a co-enhance manner. Experimental results on three datasets demonstrate that the proposed motion refinement and feature aggregation methods are highly beneficial for unsupervised video object segmentation and the proposed dual-stream co-enhanced network has achieved comparable results with some state-of-the-art methods.

Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities (Science and technology leading talent team project) (2022JBQY009), National Natural Science Foundation of China (51827813), National Key R&D Program “Transportation Infrastructure” “Reveal the list and take command” project (2022YFB2603302) and R&D Program of Beijing Municipal Education Commission (KJZD20191000402).

References

[1] Yin Li et al., "Video Object Cut and Paste," in *Proc. of SIGGRAPH '05 ACM SIGGRAPH 2005 papers*, pp.595–600, 2005. Article (CrossRefLink)

[2] W. Wang, J. Shen and F. Porikli, "Selective Video Object Cutout," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp.5645-5655, 2017. Article (CrossRefLink)

[3] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez and A. M. Lopez, "Vision-Based Offline-Online Perception Paradigm for Autonomous Driving," in *Proc. of 2015 IEEE Winter Conference on Applications of Computer Vision*, pp.231-238, 2015. Article (CrossRefLink)

[4] K. Saleh, M. Hossny and S. Nahavandi, "Kangaroo Vehicle Collision Detection Using Deep Semantic Segmentation Convolutional Neural Network," in *Proc. of 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp.1-7, 2016. Article (CrossRefLink)

[5] I. Cohen and G. Medioni, "Detecting and tracking moving objects for video surveillance," in *Proc. of Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol.2, pp.319-325, 1999. Article (CrossRefLink)

[6] Á. Erdélyi, T. Barát, P. Valet, T. Winkler and B. Rinner, "Adaptive cartooning for privacy protection in camera networks," in *Proc. of 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp.44-49, 2014. Article (CrossRefLink)

[7] Wenguan Wang, Jianbing Shen and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3395-3402, 2015. Article (CrossRef Link)

[8] D. Pathak, R. Girshick, P. Dollár, T. Darrell and B. Hariharan, "Learning Features by Watching Objects Move," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6024-6033, 2017. Article (CrossRef Link)

[9] P. Tokmakov, K. Alahari and C. Schmid, "Learning Motion Patterns in Videos," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.531-539, 2017. Article (CrossRef Link).

[10] S. D. Jain, B. Xiong and K. Grauman, "FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2117-2126, 2017. Article (CrossRef Link)
[11] P. Tokmakov, K. Alahari and C. Schmid, "Learning Video Object Segmentation with Visual Memory," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.4491-4500, 2017. Article (CrossRef Link)
[12] P. Tokmakov, C. Schmid and K. Alahari, "Learning to Segment Moving Objects," *International Journal of Computer Vision*, vol.127, pp.282-301, 2019. Article (CrossRef Link)
[13] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong and M. Kankanhalli, "Unsupervised Online Video Object Segmentation With Motion Property Understanding," *IEEE Transactions on Image Processing*, vol.29, pp.237-249, 2020. Article (CrossRef Link)
[14] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.724-732, 2016. Article (CrossRef Link)
[15] K. Fukuchi et al., "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. of 2009 IEEE International Conference on Multimedia & Expo*, pp.638-641, 2009. Article (CrossRef Link)
[16] F. Li, T. Kim, A. Humayun, D. Tsai and J. M. Rehg, "Video Segmentation by Tracking Many Figure-Ground Segments," in *Proc. of 2013 IEEE International Conference on Computer Vision*, pp.2192-2199, 2013. Article (CrossRef Link)
[17] T.Brox and J. Malik, "Object Segmentation by Long Term Analysis of Point Trajectories," in *Proc. of Computer Vision - ECCV 2010*, pp.282-295, 2010. Article (CrossRef Link)
[18] K. Fragkiadaki, G. Zhang and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1846-1853, 2012. Article (CrossRef Link)
[19] M. Keuper, B. Andres and T. Brox, "Motion Trajectory Segmentation via Minimum Cost Multicuts," in *Proc. of 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3271-3279, 2015. Article (CrossRef Link)
[20] P. Ochs and T. Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *Proc. of 2011 International Conference on Computer Vision*, pp.1583-1590, 2011. Article (CrossRef Link)
[21] P. Ochs, J. Malik and T. Brox, "Segmentation of Moving Objects by Long Term Video Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.36, no.6, pp.1187-1200, 2014. Article (CrossRef Link)
[22] A. Papazoglou and V. Ferrari, "Fast Object Segmentation in Unconstrained Video," in *Proc. of 2013 IEEE International Conference on Computer Vision*, Sydney, NSW, pp.1777-1784, 2013. Article (CrossRef Link)
[23] H. Fu, D. Xu, B. Zhang and S. Lin, "Object-Based Multiple Foreground Video Co-segmentation," in *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3166-3173, 2014. Article (CrossRef Link)
[24] Y. J. Koh et al., "Sequential Clique Optimization for Video Object Segmentation," in *Proc. of Computer Vision – ECCV 2018*, vol.11218, pp.537-556, 2018. Article (CrossRef Link)
[25] Y. J. Koh and C. Kim, "Primary Object Segmentation in Videos Based on Region Augmentation and Reduction," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7417-7425, 2017. Article (CrossRef Link)
[26] Y. J. Lee, J. Kim and K. Grauman, "Key-segments for video object segmentation," in *Proc. of 2011 International Conference on Computer Vision*, pp.1995-2002, 2011. Article (CrossRef Link)
[27] D. Zhang, O. Javed and M. Shah, "Video Object Segmentation through Spatially Accurate and Temporally Dense Extraction of Primary Object Regions," in *Proc. of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.628-635, 2013. Article (CrossRef Link)

[28] Alon Faktor and Michal Irani, “Video Segmentation by Non-Local Consensus Voting,” in *Proc. of BMVC 2014 - Proceedings of the British Machine Vision Conference 2014. British Machine Vision Association, BMVA*, 2014. Article (CrossRef Link)
[29] Yuan-Ting Hu, Jia-Bin Huang and Alexander G. Schwing. "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proc. of Computer Vision – ECCV 2018*, vol.11205, pp.813-830, 2018. Article (CrossRef Link)
[30] J. Cheng, Y. Tsai, S. Wang and M. Yang, "SegFlow: Joint Learning for Video Object Segmentation and Optical Flow," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.686-695, 2017. Article (CrossRefLink)
[31] Hongmei Song et al., "Pyramid Dilated Deeper Convlstm for Video Salient Object Detection," in *Proc. of Computer Vision – ECCV 2018*, vol.11215, pp.744-760, 2018. Article (CrossRef Link)
[32] Yikun Zhang et al., "Video Object Segmentation with Weakly Temporal Information," *KSII Transactions on Internet and Information Systems*, vol.13, no.3, pp.1434-1449, 2019. Article (CrossRef Link)
[33] M. Faisal, I. Akhter, M. Ali and R. Hartley, "EpO-Net: Exploiting Geometric Constraints on Dense Trajectories for Motion Saliency," in *Proc. of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1873-1882, 2020. Article (CrossRef Link)
[34] M. Siam et al., "Video Object Segmentation using Teacher-Student Adaptation in a Human Robot Interaction (HRI) Setting," in *Proc. of 2019 International Conference on Robotics and Automation (ICRA)*, pp.50-56, 2019. Article (CrossRef Link)
[35] Junyoung Chung et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. of NIPS 2014 Workshop on Deep Learning*, arXiv preprint arXiv:1412.3555, 2014. Article (CrossRef Link)
[36] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.2, pp.386-397, 2020. Article (CrossRefLink)
[37] G. Lee, Y. Tai and J. Kim, "Deep Saliency with Encoded Low Level Distance Map and High Level Features," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.660-668, 2016. Article (CrossRefLink)
[38] Zhaowei Cai et al., "A Unified Multi-Scale Deep Convolutional Neural Network for Fast Object Detection," in *Proc. of Computer Vision – ECCV 2016*, vol.9908, pp.354-370, 2016. Article (CrossRefLink)
[39] G. Lin, A. Milan, C. Shen and I. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5168-5177, 2017. Article (CrossRefLink)
[40] Y. Liu et al., "Richer Convolutional Features for Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.41, no.8, pp.1939-1946, 2019. Article (CrossRefLink)
[41] Matthew D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks," in *Proc. of Computer Vision – ECCV 2014*, vol.8689, pp.818-833, 2014. Article (CrossRefLink)
[42] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu and P. H. S. Torr, "Deeply Supervised Salient Object Detection with Short Connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.41, no.4, pp.815-828, 2019. Article (CrossRefLink)
[43] Zijun Deng et al., "R^3Net: Recurrent Residual Refinement Network for Saliency Detection," in *Proc. of IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp.684-690, 2018. Article (CrossRefLink)
[44] Z. Wu, L. Su and Q. Huang, "Cascaded Partial Decoder for Fast and Accurate Salient Object Detection," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3902-3911, 2019. Article (CrossRefLink)
[45] P. Zhang, D. Wang, H. Lu, H. Wang and X. Ruan, "Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.202-211, 2017. Article (CrossRefLink)
[46] O. Ronneberger, P. Fischer and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. of Medical Image Computing and Computer Assisted Intervention – MICCAI 2015*, vol.9351, pp.234-241, 2015. Article (CrossRef Link)

[47] D. Sun, X. Yang, M. Liu and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8934-8943, 2018. Article (CrossRef Link)
[48] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016. Article (CrossRef Link)
[49] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.8, pp.2011-2023, 2020. Article (CrossRef Link)
[50] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013. Article (CrossRef Link)
[51] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp.834-848, 2018. Article (CrossRef Link)
[52] Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248-255, 2009. Article (CrossRef Link)
[53] Olga Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol.115, pp.211-252, 2015. Article (CrossRef Link).
[54] Daniel J. Butler et al., "A Naturalistic Open Source Movie for Optical Flow Evaluation," in *Proc. of Computer Vision – ECCV 2012*, vol.7577, pp.611-625, 2012. Article (CrossRef Link)
[55] C. Yang, L. Zhang, H. Lu, X. Ruan and M. Yang, "Saliency Detection via Graph-Based Manifold Ranking," in *Proc. of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3166-3173, 2013. Article (CrossRef Link).
[56] M. Cheng, G. Zhang, N. J. Mitra, X. Huang and S. Hu, "Global contrast based salient region detection," in *Proc. of CVPR 2011*, pp.409-416, 2011. Article (CrossRef Link).
[57] Adam Paszke et al., "Automatic differentiation in pytorch," in *Proc. of NIPS 2017 Autodiff Workshop*, 2017. Article (CrossRef Link)
[58] B. Griffin and J. Corso, "Tukey-Inspired Video Object Segmentation," in *Proc. of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1723-1733, 2019. Article (CrossRef Link)
[59] S. Caelles, K. -K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers and L. Van Gool, "One-Shot Video Object Segmentation," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5320-5329, 2017. Article (CrossRef Link)

Hongliang Zhu received the B.E. degree from Tiangong University, Tianjin, China, in 2019. He is currently pursuing the M.S. degree from School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. His main research interests include video object segmentation and computer vision.

Hui Yin received the Ph.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China. She is currently a Full Professor of the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests include the machine vision, intelligent information processing and their application in the railway industry.

Yanting Liu received the B.E. degree from Hebei University, Baoding, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her current research interests include computer vision and pattern recognition.

Ning Chen received the B.E. degree from Beijing Information Science and Technology University, Beijing, China, in 2018. She is currently pursuing the M.S. degree from School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her main research interests include video object segmentation and computer vision.