Learning Discriminative Feature with CRF for Unsupervised Video Object Segmentation

Mingmin Zhen[1][0000−0002−8180−1023], Shiwei Li[2], Lei Zhou[1], Jiaxiang Shang[1], Haoan Feng[1], Tian Fang[2], and Long Quan[1]

[1] Hong Kong University of Science and Technology
{mzhen,lzhouai,hfengac,jshang,quan}@cse.ust.hk
[2] Everest Innovation Technology
{sli,fangtian}@altizure.com

Abstract. In this paper, we introduce a novel network, called discriminative feature network (DFNet), to address the unsupervised video object segmentation task. To capture the inherent correlation among video frames, we learn discriminative features (D-features) from the input images that reveal feature distribution from a global perspective. The D-features are then used to establish correspondence with all features of test image under conditional random field (CRF) formulation, which is leveraged to enforce consistency between pixels. The experiments verify that DFNet outperforms state-of-the-art methods by a large margin with a mean IoU score of 83.4% and ranks first on the DAVIS-2016 leaderboard while using much fewer parameters and achieving much more efficient performance in the inference phase. We further evaluate DFNet on the FBMS dataset and the video saliency dataset ViSal, reaching a new state-of-the-art. To further demonstrate the generalizability of our framework, DFNet is also applied to the image object co-segmentation task. We perform experiments on a challenging dataset PASCAL-VOC and observe the superiority of DFNet. The thorough experiments verify that DFNet is able to capture and mine the underlying relations of images and discover the common foreground objects.

Keywords: Video object segmentation, discriminative feature, CRF

1 Introduction

The research on video object segmentation (VOS), which aims to separate primary foreground objects from their background in a given video, is often divided into two categories, *i.e.*, semi-supervised and unsupervised setting. The semi-supervised VOS (SVOS) provides a mask of the first frame, which can be taken as the prior knowledge about the target in subsequent frames. By comparison, unsupervised VOS (UVOS) is in general more challenging, as it requires a further step to distinguish the target object from a complex and diverse background without prior information. In this paper, we focus on the latter challenging issue.

Recently, several works, such as COSNet [40], AGNN [61] and AnDiff [68],

model the long-term correlations between frames to explore global information inspired by the non-local operation introduced by Wang et al. [66]. However, the limitations are obvious as the computation requirement is very high, especially for AGNN [61]. Besides, the local consistency cues are overlooked, which is essential for UVOS task.

Motivated by the above observations, we propose a discriminative feature learning network, which is denoted as DFNet, to model the long-term correlations between video frames. Specifically, DFNet takes several frames from the same video as input and learns the discriminative features, which can denote the whole feature distribution of the input frames. The feature map for each frame is correlated with these discriminative features under CRF formulation, which is used to boost the smoothness and consistency of similar pixels. The proposed approach is advantageous to mine the discriminative representation from a global perspective, while at the same time helps to capture the rich contextual information within video frames. DFNet is sufficiently flexible to process variable numbers of input frames during inference, enabling it to consider more input information and gain better performance.

To verify the proposed method, we extensively evaluate DFNet on two widely-used video object segmentation datasets, namely DAVIS16 [45] and FBMS [43], showing its superior performance over current state-of-the-art methods. More specifically, DFNet ranks first on the DAVIS-2016 leaderboard with a mean IoU score of 83.4%, which is 1.7% higher than state-of-the-art method [68]. DFNet also achieves state-of-the-art results on FBMS [43] and the ViSal [62] video saliency benchmark. To further demonstrate its advantages and generalizability, we apply DFNet to image object co-segmentation task, which aims to extract the common objects from a group of semantically related images. It also gains better results on the representative dataset PASCAL VOC [7] over previous methods.

2 Related work

Unsupervised Video Object Segmentation Recently, there are many works for UVOS task, which focus on the fully convolutional neural network based models. MPNet [52], a purely optical flow-based method, discards appearance modeling and casts segmentation as foreground motion prediction, which poorly deals with static foreground objects. To better address this problem, several methods [53, 4, 48, 34] suggest adopting two-stream fully convolutional networks, which fuse the motion and appearance information for object inference. In [53], a convolutional gated recurrent unit is employed to extend the horizon spanned by optical flow based features. Li et al. [34] attempt to address this issue by employing a bilateral network for detecting the motion of background objects. RNN based methods are also a popular choice. Song et al. [49] propose a novel convolutional long short-term memory [11] architecture, in which two atrous convolution [3] layers are stacked along the forward axis and propagate features in opposite directions. COSNet [40] adopts a gated co-attention mechanism to

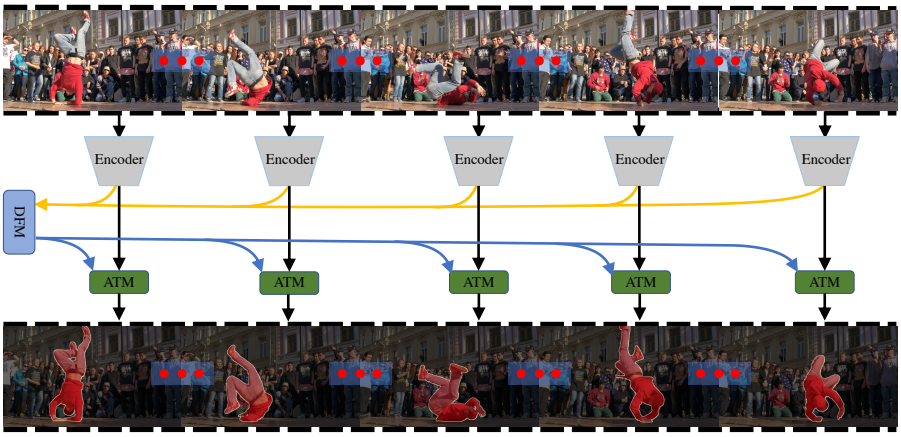

Fig. 1. Overall pipeline of the proposed method. The features are first obtained from the encoder module and goes through the discriminative feature module (DFM) to extract discriminative features. The discriminative features are then used by attention module (ATM) to recontruct a new feature map, which is used to correlate the input frames.

model the correlation of input video images. In AGNN [61], a fully connected graph is built to represent frames as nodes, and relations between arbitrary frame pairs as edges. The underlying pair-wise relations are described by a differentiable attention mechanism. To exploit the correlations of images, AnDiff [68] proposes a considerably simpler method, which propagates the features of the first frame (the "anchor") to the current frame via an aggregation technique.

Image Object Co-Segmentation Different from UVOS, the image object co-segmentation task is to extract the common object with the same semantics from a group of semantic-related images. Recent researches [17, 69] use deep visual features to improve object co-segmentation, and they also try to learn more robust synergetic properties among images in a data-driven manner. Hsu et al. [17] proposes a DNN-based method which uses the similarity between images in deep features and an additional object proposals algorithm [25] to segment the common objects. Yuan et al. [69] introduce a DNN-based dense conditional random field framework for object co-segmentation by cooperating co-occurrence maps, which are generated using selective search [55]. The very recent works [2, 35] propose end-to-end deep learning methods for co-segmentation by integrating the process of feature learning and co-segmentation inferring as an organic whole. By introducing the correlation layer [35] or a semantic attention learner [2], they can utilize the relationship between the image pair and then segment the co-object in a pairwise manner. In [30], a recurrent network architecture is proposed to address group-wise object co-segmentation.

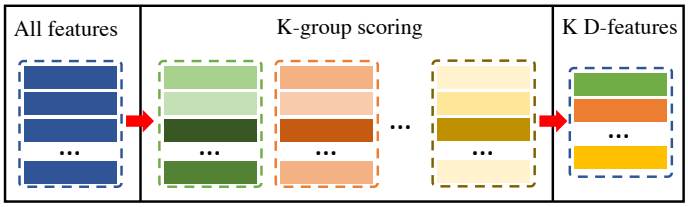

Fig. 2. Illustration of DFM. The features from input images are first reshaped into one-dimensional vectors. The K-group scoring module is adopted to score the features. Based on the K-group scores, we can obtain final K-D features. The details are presented in Section 3.2.

3 The Proposed Method

In this section, we present the proposed DFNet in detail, which is illustrated in Figure 1. We first give an overview of the whole architecture in section 3.1. Next, the discriminative feature module (DFM), which captures the global feature distribution of all input images, is elaborated in section 3.2. Then we introduce the attention module (ATM) in section 3.3, which reconstructs a new feature map modeling the long-term dependency.

3.1 Network Architecture

For the UVOS task, the target object in the given video images can be deformed and occluded, which often deteriorates the performance of estimated binary segmentation results. To recognize the target object, our method should be of two essential properties: (i) the ability to extract foreground objects from the individual frame; (ii) the ability to keep consistency among the video frames. To achieve these two goals, we correlate the features of each input image with discriminative features, which is extracted from input images selected from the same video randomly.

As shown in Figure 1, we present the proposed network architecture in detail. The proposed network takes several images as input. The shared feature encoder, which adopts the fully convolutional DeepLabv3 [3], extracted the features from the input images. The obtained feature maps are then fed into a 1×1 convolutional layer to reduce the feature map channel to 256, and the output feature maps for all input images are taken as input for the discriminative feature module (DFM), which extract the discriminative features (D-features). The input feature for each image and the D-features go through an attention module (ATM) to reconstruct a new feature map and then one 3×3 convolutional layer followed by ReLU, batch normalization (BN) layer and one 1×1 convolutional layer followed by a *sigmoid* operation are used to obtain final binary output.

More formally, given a set of input frames $\mathcal{I} = \{I_i \in \mathbb{R}^{H \times W \times 3}\}_{i=1}^{\mathcal{N}}$, we want to segment out the binary masks $\mathcal{S} = \{S_i \in \{0,1\}^{H \times W}\}_{i=1}^{\mathcal{N}}$ for all frames. The features extracted from DeepLabv3 are denoted as $\mathcal{F} = \{F_i \in \mathbb{R}^{h \times w \times c}\}_{i}^{\mathcal{N}}$, where

$h \times w$ indicates the spatial resolution of feature map and c represents the feature map channels. Since we follow the original deepLabv3, which employs dilated convolution, the output feature map F_i is $\frac{1}{8}$ smaller than the input image I_i.

3.2 Discriminative Feature Module

We learn the discriminative features from the features of all input images. Specifically, all feature maps $\mathcal{F}$ from the input images are first concatenated to form a large feature map with size $\mathcal{N} \times h \times w \times c$ and then reshaped as $F^a \in \mathbb{R}^{\mathcal{N}hw \times c}$. As shown in Figure 2, we then use a K-group scoring module to obtain K-group scores, which is used to distinguish the discriminative features from the noisy features. For each scoring group, a weight matrix $\mathcal{W}_k \in \mathcal{R}^{c \times 1}$ and F^a is multiplied to get a initial score result with size $\mathcal{N}hw \times 1$. We apply a softmax function to calculate the final scores:

$$s_i^k = \frac{exp(F_i^a.W_k)}{\sum_i^{\mathcal{N}hw} exp(F_i^a.W_k)} \tag{1}$$

where s_i^k is the score for i^{th} feature of F^a and measures the discriminability of the feature. The final discriminative feature for k^{th} scoring group is computed as $F_k^d = \sum s_i^k F_i^a$. By this way, we can obtain K discriminative features $F^d \in \mathbb{R}^{K \times c}$.

The K D-features are used to describe the feature distribution from a global perspective. The key of the D-features computation is the scoring weight $\mathcal{W}_k$. In our training step, we initialize the $\mathcal{W}_k$ by using Kaiming's initialization method [15]. For each updating iteration, we adopt the moving averaging mechanism, which is used in batch normalization (BN) [18]. After obtaining the D-feature $F_k^d(t)$ at training step t, we update the $\mathcal{W}_k$ as:

$$\mathcal{W}_k(t) = \lambda \mathcal{W}_k(t-1) + (1-\lambda) F_k^d(t) \tag{2}$$

where λ is the momentum. In our experiments, we set it to 0.5. As we train our network on a multiple-GPU machine, we also adopt the synchronized weight updating strategy motivated by synchronized BN [47]. Specifically, the images from the same video sequence are fed into the network on one GPU. Thus, we will get different D-features $F_k^d(t)$ at step t for different GPUs. For the synchronized processing, we sum up these D-features $F_k^d(t)$ across GPUs and compute the average feature $\overline{F}_k^d(t)$, which will be used in Equation 2. The updated $\mathcal{W}_k$ is synchronized on all GPUs. The whole computation is differentiable and trainable. In the inference step, the weight $\mathcal{W}_k$ is kept fixed, which is similar to BN operation.

3.3 Attention Module with CRF

To model the long-term dependency, we adopt the attention module to correlate input image and the discriminative features. For the obtained K D-features $F^d \in \mathbb{R}^{K \times c}$ and the feature map $F_i \in \mathbb{R}^{h \times w \times c}$ of input image, we follow [40,

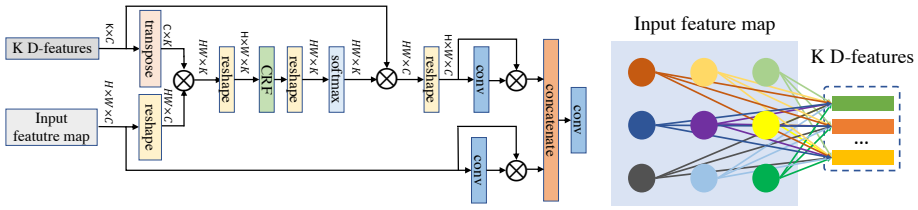

Fig. 3. (a) Illustration of ATM. The input feature map and K D-features are correlated to model long-term dependency; (b) Illustration of attention mechanism. The details are presented in Section 3.3.

61] to compute the attention matrix $P \in \mathbb{R}^{hw \times K}$ as shown in Figure 3 (a). Specifically, we obtain P from F^d and F_i as follows:

$$P = reshape(F_i)W_{att}transpose(F^d) \tag{3}$$

where $W_{att} \in \mathbb{R}^{c \times c}$ is a learnable weight matrix. The D-feature matrix F^d are tranposed with size $c \times K$ and feature map F_i is reshaped with size $hw \times c$. For the obtained attetnion matrix P, each element indicates the similarity of the corresponding feature of F_i and feature of F^d. As shown in Figure 3 (b), the lines with different colors represent the similarity between input features and K D-features. In previous attention methods [40, 61, 68], a new feature map is reconstructed based on the attention matrix by assigning K D-features to input feature map as follows:

$$F^{new} = reshape(softmax(P)F^d) \tag{4}$$

where the new feature map F^{new} is of size $h \times w \times c$.

The attention map computation can also be considered as multi-label classification problem and the assignment of D-features corresponds to a different label. Our intuition is that neighboring pixels in the same local region tend to have similar labels (K D-features), and pixels near borders or edges may have significantly different labels. We regard the reshaped attention map with size $h \times w \times K$ as fully connected pairwise conditional random fields conditioned on the corresponding image I, in which each pixel is to be assigned with a D-feature for reconstructing the new feature map.

Let $\mathbf{x} = \{x_1, x_2, ..., x_M\}$ be the label vector of M pixels in the reshaped attention map. Component x_i belongs to $\{1, 2, ..., K\}$ where K is the number of labels (D-features). The probability of the label assignment is defined in the form of Gibbs distribution as $P(\mathbf{x}|\mathbf{I}) = \frac{1}{Z}exp(-E(\mathbf{x}|\mathbf{I}))$, where $E(x)$ is the energy function which describes the cost of label assigning and Z is a normalization factor. For convenience we drop the notation of condition $\mathbf{I}$ in the followings. Following the formulation of [24], the energy function is defined as

$$E(\mathbf{x}) = \sum_{i=1}^{M} \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{5}$$

where the unary energy components $\psi_u(x_i)$ measure the cost of the pixel i taking the label x_i, and pairwise energy components $\psi(x_i, x_j)$ measure the cost of assigning labels x_i, x_j to pixels i, j simultaneously. In our formulation, unary energies are set to be reshaped attention map P, which predicts labels for pixels without considering the smoothness and the consistency of the label assignments. The pairwise energies provide an image data-dependent smoothing term that encourages assigning similar labels to pixels with similar properties.

The CRF model can be implemented in neural networks as shown in [72, 51], thus it can be naturally integrated in our network, and optimized in the end-to-end training process. After the CRF module, we can obtain a refined attention map which takes the smoothness and consistency into consideration. We follow Equation 4 to reconstruct a new feature map.

We also adopt a self-weight method to weight the new feature map F^{new} and input feature map F_i. The self-weight is formulated as follows:

$$F^{new} = F^{new} * conv(F^{new}), F_i = F_i * conv(F_i) \tag{6}$$

where we use 1×1 convolutional layer to get the weight, which indicates the importance of features in the feature map. At last, we concatenate the feature map F^{new} and F_i and feed the obtained feature map into the convolutional layers to get binary segmentation results.

4 Experiments

We first report performance on the unsupervised video object segmentation task in Section 4.1. Then, in Section 4.2, to further demonstrate the advantages of the proposed model, we test it on image object co-segmentation task. At last, we conduct an ablation study in Section 4.3 and model analysis in Section 4.4

4.1 Unsuperviesed Video Object Segmentation Task

Dataset and Evaluation Metric To evaluate UVOS task, a golden dataset **DAVIS16** is often used [40, 61, 65, 54, 49]. DAVIS16 is a recent dataset which consists of 50 videos in total (30 videos for training and 20 for testing). Per-frame pixel-wise annotations are offered. For quantitative evaluation, following the standard evaluation protocol from [45], we adopt three metrics, namely region similarity $\mathcal{J}$, which is the intersection-over-union of the prediction and ground truth, boundary accuracy $\mathcal{F}$, which is the F-measure defined on contour points in the prediction and ground truth, and time stability $\mathcal{T}$, which measures the smoothness of evolution of objects across video sequences. **FBMS** [43] is comprised of 59 video sequences. Different from the DAVIS16 dataset, the ground-truth of FBMS is sparsely labeled (only 720 frames are annotated). Following the common setting [49, 68, 48], we validate the proposed method on the testing split, which consists of 30 sequences. On the FBMS dataset, the F-measure is used as evaluation metric. We also follow [49, 68] to report saliency

Table 1. Quantitative results on the test set of DAVIS16, using the region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$ and time stability $\mathcal{T}$. For FBMS dataset, we report the F-measusre results. The best scores are marked in bold.

		DAVIS			FBMS
Method	Year	$\mathcal{J}$ Mean↑	$\mathcal{F}$ Mean↑	$\mathcal{T}$ Mean↓	F-measure
TRC [10]	CVPR12	47.3	44.1	39.1	-
CVOS [50]	CVPR15	48.2	44.7	25.0	-
KEY [29]	ICCV11	49.8	42.7	26.9	-
MSG [42]	ICCV11	53.3	50.8	30.2	-
NLC [8]	BMVC14	55.1	52.3	42.5	-
CUT [22]	ICCV15	55.2	55.2	27.7	-
FST [44]	ICCV13	55.8	51.1	36.6	69.2
ELM [26]	ECCV18	61.8	61.2	25.1	-
TIS [12]	WACV19	62.6	59.6	33.6	-
SFL [4]	ICCV17	67.4	66.7	28.2	-
LMP [52]	CVPR17	70.0	65.9	57.2	77.5
FSEG [19]	CVPR17	70.7	65.3	32.8	-
UOVOS [73]	TIP19	73.9	68.0	39.0	-
LVO [53]	ICCV17	75.9	72.1	26.5	77.8
ARP [23]	CVPR17	76.2	70.6	39.3	-
PDB [49]	ECCV18	77.2	74.5	29.1	81.5
LSMO [54]	IJCV19	78.2	75.9	21.2	-
MotAdapt [48]	ICRA19	77.2	77.4	27.9	79.0
EpO+ [6]	WACV20	80.6	75.5	19.3	-
AGS [65]	CVPR19	79.7	77.4	26.7	-
COSNet [40]	CVPR19	80.5	79.5	18.4	-
AGNN [61]	ICCV19	80.7	79.1	33.7	-
AnDiff [68]	ICCV19	81.7	80.5	21.4	81.2
Ours	ECCV20	**83.4**	**81.8**	**15.9**	**82.3**

evaluations of our method on DAVIS, FBMS and a video salient object detection dataset ViSal [62] for demonstrating the robustness and wide applicability of our method. The **ViSal** [62] dataset is a video salient object detection benchmark. The length of videos in ViSal ranges from 30 to 100 frames, and totally 193 frames are manually annotated. The whole ViSal dataset is used for evaluation. We report the mean absolute error (MAE) and the F-measure on the three datasets.

Implementation Details In the **training** step, following [40, 61, 49], we use both static data from image salient object segmentation datasets, MSRA10K [5], DUT [67], and video data from the training set of DAVIS16 to train our model. The training process is divided into two steps. First, we use the static training data to train our backbone encoder (DeepLabV3) to extract more discriminative foreground features. The learning rate is set to 0.01 and the batch size is 12. Then we use the DAVIS16 training data to train the whole model with learning rate of 0.001. The batch size is set to 8. For each video sequence, we follow [68] to select the first frame as an anchor and randomly sample one image

Table 2. Quantitative comparison results against saliency methods using MAE and maximum F-measure on DAVIS16 [45], FBMS [43] and ViSal [62]. The best scores are marked in bold. * means non-deep learning model.

			DAVIS16		FBMS		ViSal	
	Methods	Year	MAE↓	F↑	MAE↓	F↑	MAE↓	F↑
Image	Amulet [70]	ICCV17	0.082	69.9	0.110	72.5	0.032	89.4
	SRM [59]	ICCV17	0.039	77.9	0.071	77.6	0.028	89.0
	UCF [71]	ICCV17	0.107	71.6	0.147	67.9	0.068	87.0
	DSS [16]	CVPR17	0.062	71.7	0.083	76.4	0.028	90.6
	MSR [31]	CVPR17	0.057	74.6	0.064	78.7	0.031	90.1
	NLDF [41]	CVPR17	0.056	72.3	0.092	73.6	0.023	91.6
	DCL [33]	CVPR16	0.070	63.1	0.089	72.6	0.035	86.9
	DHS [37]	CVPR16	0.039	75.8	0.083	74.3	0.025	91.1
	ELD [28]	CVPR16	0.070	68.8	0.103	71.9	0.038	89.0
	KSR [60]	ECCV16	0.077	60.1	0.101	64.9	0.063	82.6
	RFCN [58]	ECCV16	0.065	71.0	0.105	73.6	0.043	88.8
Video	FGRNE [32]	CVPR18	0.043	78.6	0.083	77.9	0.040	85.0
	FCNS [63]	TIP18	0.053	72.9	0.100	73.5	0.041	87.7
	SGSP* [38]	TCSVT17	0.128	67.7	0.171	57.1	0.172	64.8
	GAFL* [62]	TIP15	0.091	57.8	0.150	55.1	0.099	72.6
	SAGE* [64]	CVPR15	0.105	47.9	0142	58.1	0.096	73.4
	STUW* [9]	TIP14	0.098	69.2	0.143	52.8	0.132	67.1
	SP* [39]	TCSVT14	0.130	60.1	0.161	53.8	0.126	73.1
	PDB [49]	ECCV18	0.030	84.9	0.069	81.5	0.022	91.7
	AnDiff [68]	ICCV19	0.044	80.8	0.064	81.2	0.030	90.4
	Ours	ECCV20	**0.018**	**89.9**	**0.054**	**83.3**	**0.017**	**92.7**

as the training example. The model is trained with binary cross-entropy loss. Network parameters are optimized via stochastic gradient descent with weight decay 0.0001. We adopt the "poly" learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with $power = 0.9$. Raw predictions are upsampled via bilinear interpolation to the size of the ground-truth masks. In the **inference** step, multiscale and mirrored inputs are employed to enhance the final performance. The final heatmap is the mean of all output heatmaps. Thresholding at 0.5 produces the final binary labels. We also follow [68] to adopt instance pruning as a post-processing method.

Experimental Results In Table 1, we evaluate DFNet against state-of-the-art unsupervised VOS methods on the DAVIS16 public leaderboard. DFNet attains the highest performance among all unsupervised methods on the DAVIS16 validation set, while also achieving a new state-of-the-art on the FBMS test set. In particular, on DAVIS16 we outperform the second-best method (AnDiff [68]) by an absolute margin of 21.7% in the region similarity $\mathcal{J}$ and 1.3% in the boundary accuracy $\mathcal{F}$. For the temporal stability $\mathcal{T}$, our method shows a more stable result over the video sequences by a large margin of 2.5 than the second-best method COSNet [40]. We also outperform state-of-the-art method

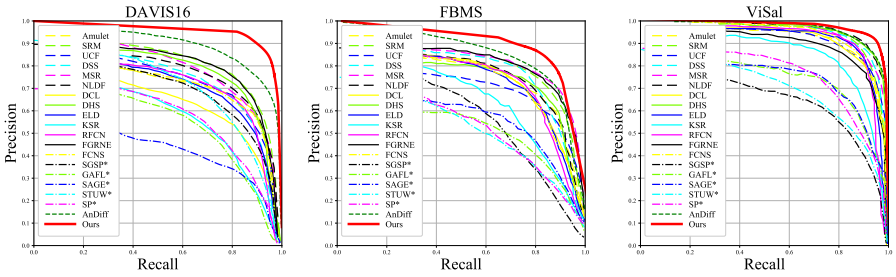

Fig. 4. Quantitative comparison against other methods using PR curve on DAVIS16 [45], FBMS [43] and ViSal [62] datasets.

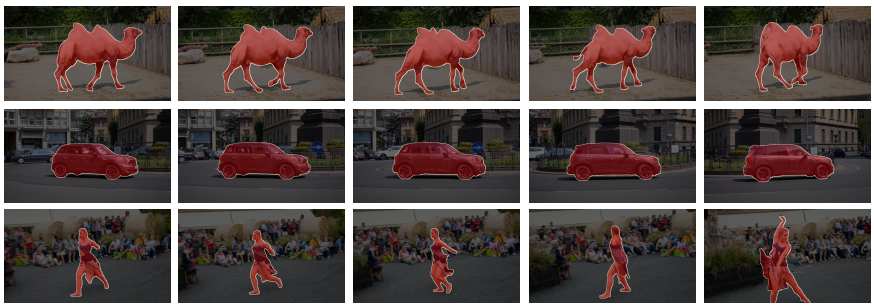

Fig. 5. The visual results generated by our approach on the DAVIS16 dataset. From the first row to the last row, the corresponding video names are *camel*, *car-roundabout* and *dance-twirl* respectively.

AnDiff [68] by 1.1% in F-measure on the FBMS dataset.

We also report the results on salient object detection for DAVIS16, FBMS and ViSal datasets as shown in Table 2. It can be observed that the proposed method improves state-of-the-art for all the three datasets for standard saliency scores, showing consistency with Table 1. The largest improvements lie in DAVIS16, where both MAE and F-measure significantly outperform previous records. Especially for the F-measure, we outperform the second-best result by a significant margin of 9.1%. The precision-recall analysis of DFNet is presented in Figure 4, where we demonstrate that our approach generally outperforms also existing salient object detection methods. DFNet achieves superior performance in all regions of the PR curve on the DAVIS validation set, maintaining significantly higher precision at all recall thresholds. On the challenging FBMS test set, DFNet shows inferior precision results than SP [39]at the recall threshold from 0.93 to 0.97 and FGRNE [32] from 0.94 to 0.95. But overall speaking, DFNet maintains a clear advantage compared with all other methods. On the ViSal dataset, it is noteworthy that the precision is higher than the other methods at nearly all recall thresholds, except for the AnDiff [68] at the threshold

Fig. 6. Qualitative comparison with state-of-the-art methods (AnDiff [68], AGNN [61] and COSNet [40]) on DAVIS16 dataset.

Table 3. The performance of object co-segmentation on the PASCAL-VOC dataset under Jaccard index and Precision. The numbers in red and green respectively indicate the best and the second best results.

Method	Faktor13 [7]	Lee15 [27]	Chang15 [1]	Hati16 [14]	Quan16 [46]	Jerri.16 [21]	Wang17 [57]	Jerri.17 [20]	Han18 [13]	Hsu18 [17]	Li19 [30]	Ours
Avg. $\mathcal{P}$	84.0	69.8	82.4	72.5	89.0	85.2	84.3	80.1	90.1	91.0	94.1	95.2
Avg. $\mathcal{J}$	0.46	0.33	0.29	0.25	0.52	0.45	0.52	0.40	0.53	0.60	0.63	0.69

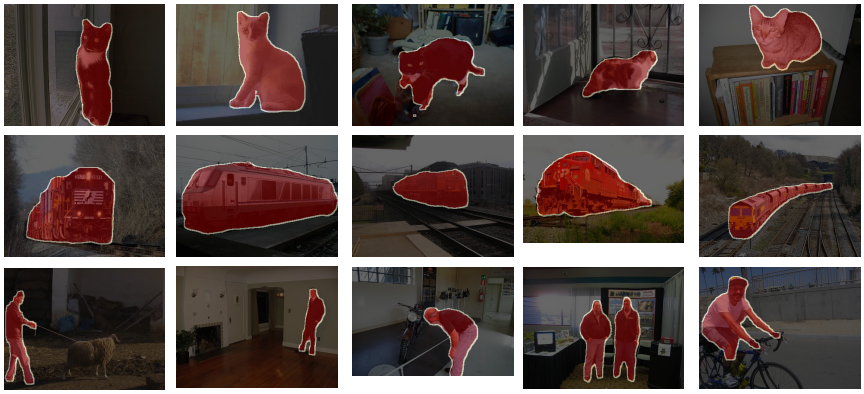

Fig. 7. The co-segment results generated by our approach on the PASCAL-VOC dataset. From the first row to the last row, the classes are *cat*, *train* and *person* respectively.

from 0.98 to 0.99. All in all, the superiority of the proposed method is verified through the comparison of the PR curves.

As shown in Figure 5, we visualize some qualitative results of the DAVIS16 dataset. We can see that the proposed method can locate the primary region or target tightly by leveraging DFM and ATM with CRF to model long-term denpendency. The primary objects from the cluttered background are segmented out correctly. We also present the visual comparison results between DFNet and COSNet [40], AGNN [61] and AnDiff [68] in Figure 6. In can be observed that the results of DFNet are more accurate and complete than the other three methods.

4.2 Image Object Co-segmentation Task

Dataset and Evaluation Metric The PASCAL-VOC [7] is a well-known dataset often used in image object co-segmentation task, which contains total 1,037 images of 20 object classes from PASCAL-VOC 2010 dataset. The PASCAL-VOC dataset is challenging and difficult due to extremely large intra-class variations and subtle figure-ground discrimination. Following previous works [7, 27, 1, 21, 30], two widely used measures, precision ($\mathcal{P}$) and Jaccard index ($\mathcal{J}$), are adapted to evaluate the performance of object co-segmentation.

Implementation Details We follow [56, 30] to train the proposed network with generated training data from the existing MS COCO dataset [36]. The learning

Table 4. Ablation study on DAVIS16 with different components used and different numbers of D-features adopted. We also compare the performance for different numbers of input images on DAVIS16 and PASCAL VOC.

Method	Baseline	+DFM&ATM	+ATM&CRF	+multiple scales	+I.Prun.
$\mathcal{J}$ mean (%)	76.7	79.5	80.4	81.1	83.4
D-features					
K	128	256	512	1024	2048
$\mathcal{J}$ mean (%)	78.3	79.0	79.4	79.7	80.4
Input images (DAVIS16)					
N^{in}	1	2	4	8	10
$\mathcal{J}$ mean (%)	79.4	80.1	80.4	80.4	80.4
Input images (PASCAL VOC)					
N^{in}	1	2	4	8	10
$Avg.\mathcal{J}$ (%)	61.4	63.5	65.0	65.4	65.4

rate is set to 0.01 and batch size is 12. For each group images, we randomly select three images as one training example. Other training setups are the same as those in previous unsupervised VOS task. After training, we evaluate the performance of our method on the PASCAL VOC dataset. When processing an image, we leverage another 4 images belonging to the same group to form a subgroup as inputs. We adopt a threshold 0.5 to generate final binary masks.

Experimental Results We compare our methods with state-of-the-art methods on the PASCAL VOC dataset. As shown in Table 3, although the objects of the PASCAL VOC dataset undergo drastic variation in scale, position and appearance, our method improves upon the second-best results [30] by margins 1.1% and 6% in terms of $\mathcal{P}$ and $\mathcal{J}$ respectively. We also present some co-segmentation results of the proposed method in Figure 7. It can be seen that our method can generate promising object segments under different types of intra-class variations, such as colors, sharps, views, scales and background clutters.

4.3 Ablation Study

To verify the effectiveness of the proposed method, we conduct ablation experiments on DAVIS16 and PASCAL VOC. As shown in Table 4, the detailed results are reported for different experimental setup. We adopt the DeepLabv3 as the baseline, which is trained on the static image dataset, and achieve 76.7% in terms of $\mathcal{J}$. After adding the proposed DFM and ATM into the network, the performance increase to 79.5%, which validates the usefulness of modeling the long-term dependency. We then adopt CRF to optimize the attention map by considering the smoothness and consistency, which improves the performance by 0.9%. Multiple-scale inference and instance pruning (I.Prun.) are also used by following [68]. At last, we obtain the highest score of 83.4% in terms of the region similarity $\mathcal{J}$, which outperforms state-of-the-art methods. By adopting different numbers of D-features, we can see that better results can be obtained

Table 5. The number of model parameters and inference time comparison with state-of-the-art methods.

Method	COSNet [40]	AGNN [61]	AnDiff [68]	Ours
# Parmeter (M)	81.2	82.3	79.3	**64.7**
Inf. Time (s/image)	0.45	0.53	0.35	**0.28**

with more discriminative features used. We also evaluate the impact of the number of input images during inference, and we report performance with different values of N^{in} on DAVIS16 and PASCAL VOC datasets. For DAVIS16, we can see the performance increases by adding more input frames from 1 to 4 and then keep stable. It can be observed that with more input images, especially from 1 to 8, the performance raises accordingly on PASCAL VOC. When more images are considered, the performance does not change obviously.

4.4 Model Analysis

In Table 5, we report the comparison with state-of-the-art methods on the number of network parameters and inference time on DAVIS16. We can observe that DFNet reduces the model complexity with fewer parameters compared with COSNet [40], AGNN [61] and AnDiff [68]. For the inference comparison, we run the public code of other methods and our code on the same machine with NVIDIA GeForce GTX 1080 Ti. The inference time includes the image loading and pre-processing time. In can be seen that our method shows a faster speed than these methods.

5 Conclusion

To model the long-term dependency of video images, we propose a novel DFNet to capture the relations among video frames and infer the common foreground objects in this paper. It extracts the discriminative features from the input images, which can describe the feature distribution from a global view. An attention module is then adopted to mine the correlations between the input images. The smoothness and consistency of the attention map are also considered, in which the attention mechanism is formulated as a classification problem and solved by CRF. The extensive experiments validate the effectiveness of the proposed method. In addition, we also apply the method to image object co-segmentation task. The quantitative evaluation of the challenging dataset PASCAL VOC demonstrates the advantage of DFNet.

Acknowledgments. This work is supported by Hong Kong RGC GRF 16206819, Hong Kong RGC GRF 16203518 and Hong Kong T22-603/15N.

References

1. Chang, H., Wang, Y.F.: Optimizing the decomposition for multiple foreground cosegmentation. Computer Vision and Image Understanding (2015)
2. Chen, H., Huang, Y., Nakayama, H.: Semantic aware attention based deep object co-segmentation. In: ACCV (2018)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017)
5. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
6. Faisal, M., Akhter, I., Ali, M., Hartley, R.: Exploiting geometric constraints on dense trajectories for motion saliency. (2019)
7. Faktor, A., Irani, M.: Co-segmentation by composition. In: ICCV (2013)
8. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC (2014)
9. Fang, Y., Wang, Z., Lin, W., Fang, Z.: Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE transactions on image processing (2014)
10. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: CVPR (2012)
11. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. ICANN (2000)
12. Griffin, B.A., Corso, J.J.: Tukey-inspired video object segmentation. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2019)
13. Han, J., Quan, R., Zhang, D., Nie, F.: Robust object co-segmentation using background prior. IEEE Transactions on Image Processing (2018)
14. Hati, A., Chaudhuri, S., Velmurugan, R.: Image co-segmentation using maximum common subgraph matching and region co-growing. In: ECCV (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)
16. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
17. Hsu, K., Lin, Y., Chuang, Y.: Co-attention cnns for unsupervised object co-segmentation. In: IJCAI (2018)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
19. Jain, S.D., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: CVPR (2017)
20. Jerripothula, K.R., Cai, J., Lu, J., Yuan, J.: Object co-skeletonization with co-segmentation. In: CVPR (2017)
21. Jerripothula, K.R., Cai, J., Yuan, J.: Image co-segmentation via saliency co-fusion. IEEE Transactions on Multimedia (2016)
22. Keuper, M., Andres, B., Brox, T.: Motion trajectory segmentation via minimum cost multicuts. In: ICCV (2015)
23. Koh, Y.J., Kim, C.S.: Primary object segmentation in videos based on region augmentation and reduction. In: CVPR (2017)

24. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
25. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: ECCV (2014)
26. Lao, D., Sundaramoorthi, G.: Extending layered models to 3d motion. In: ECCV (2018)
27. Lee, C., Jang, W., Sim, J., Kim, C.: Multiple random walkers and their application to image cosegmentation. In: CVPR (2015)
28. Lee, G., Tai, Y., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: CVPR (2016)
29. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
30. Li, B., Sun, Z., Li, Q., Wu, Y., Hu, A.: Group-wise deep object co-segmentation with co-attention recurrent neural network. In: ICCV (2019)
31. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-level salient object segmentation. In: CVPR (2017)
32. Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: CVPR (2018)
33. Li, G., Yu, Y.: Deep contrast learning for salient object detection (2016)
34. Li, S., Seybold, B., Vorobyov, A., Lei, X., Kuo, C.C.J.: Unsupervised video object segmentation with motion-based bilateral networks. In: ECCV (2018)
35. Li, W., Jafari, O.H., Rother, C.: Deep object co-segmentation. In: ACCV (2018)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
37. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: CVPR (2016)
38. Liu, Z., Li, J., Ye, L., Sun, G., Shen, L.: Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Transactions on Circuits and Systems for Video Technology (2017)
39. Liu, Z., Zhang, X., Luo, S., Meur, O.L.: Superpixel-based spatiotemporal saliency detection (2014)
40. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019)
41. Luo, Z., Mishra, A., Achkar, A., Eichel, J.A., Li, S., Jodoin, P.: Non-local deep features for salient object detection. In: CVPR (2017)
42. Ochs, P., Brox, T.: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV (2011)
43. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence (2013)
44. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
45. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
46. Rong, Q., Han, J., Zhang, D., Nie, F.: Object co-segmentation via graph optimized-flexible manifold ranking. In: CVPR (2016)
47. Rota Bulò, S., Porzi, L., Kontschieder, P.: In-place activated batchnorm for memory-optimized training of dnns. In: CVPR (2018)
48. Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., Jagersand, M.: Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: ICRA (2019)

49. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: ECCV (2018)
50. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: CVPR (2015)
51. Teichmann, M.T., Cipolla, R.: Convolutional crfs for semantic segmentation. arXiv preprint arXiv:1805.04777 (2018)
52. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: CVPR (2017)
53. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)
54. Tokmakov, P., Schmid, C., Alahari, K.: Learning to segment moving objects. IJCV (2019)
55. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision (2013)
56. Wang, C., Zha, Z.J., Liu, D., Xie, H.: Robust deep co-saliency detection with group semantic. In: AAAI (2019)
57. Wang, C., Zhang, H., Yang, L., Cao, X., Xiong, H.: Multiple semantic matching on augmented n -partite graph for object co-segmentation. IEEE Transactions on Image Processing (2017)
58. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: ECCV (2016)
59. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: ICCV (2017)
60. Wang, T., Zhang, L., Lu, H., Sun, C., Qi, J.: Kernelized subspace ranking for saliency detection. In: ECCV (2016)
61. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: ICCV (2019)
62. Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing (2015)
63. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing (2018)
64. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
65. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019)
66. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
67. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013)
68. Yang, Z., Wang, Q., Bertinetto, L., Hu, W., Bai, S., Torr, P.H.S.: Anchor diffusion for unsupervised video object segmentation. In: ICCV (2019)
69. Yuan, Z.H., Lu, T., Wu, Y.: Deep-dense conditional random fields for object co-segmentation. In: IJCAI (2017)
70. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: CVPR (2017)
71. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: ICCV (2017)
72. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: ICCV (2015)

73. Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., Kankanhalli, M.: Unsupervised online video object segmentation with motion property understanding. IEEE Transaction on Image Processing (2019)