The 2020 DAVIS Challenge on Video Object Segmentation - CVPR 2020 Workshops

Video Object Segmentation using Guided Feature and Directional Deep Appearance Learning

Mustansar Fiaz[1], Arif Mahmood[2], Soon Ki Jung[1]
[1]Kyungpook National University, Republic of Korea, [2]Information Technology University, Pakistan
{mustansar, skjung}@knu.ac.kr, arif.mahmood@itu.edu.pk

Abstract

In this work, we focus on semi-supervised Video Object Segmentation (VOS) problem, where an object mask is provided in the initial frame and VOS algorithm has to segment that object in the rest of the video frames. VOS is a challenging task due to object appearance variations, illumination changes, occlusion, background clutter and various distractions. Many online VOS methods have been proposed however, most of these methods limit their real-world applications due to computationally expensive online fine-tuning. On the contrary, many cost efficient template-based and propagation-based approaches suffer from degraded performance due to object appearance drifts. In order to tackle those issues, we propose a guided feature learning with directional deep appearance learning for VOS. First, we introduce guided feature modulation to capture the video context information based on target mask. Secondly, a directional matching module is utilized to learn pixel-wise semantic embedding. Third, a directional appearance model is integrated to represent the target and the background cues on a spherical embedding space. Finally, we propose a guided pooling decoder to learn the global and the local context information during refinement. The proposed network is trained offline and does not require fine-tuning. Our algorithm achieved an overall J and F score of 64.9 on the DAVIS 2020 test-challenge data and 60.9 on the DAVIS 2020 test-dev dataset.

1. Introduction

Video Object Segmentation is a labeling task to segment the specific target objects from the background in every frame of a sequence. VOS is an expanding research field due to wide range of applications such as video summarizing, video understanding, video editing, and action recognition. In this work, we handle VOS in a semi-supervised manner, where the groundtruth mask of target objects are provided in the initial frame and objective is to segment the target from backgrounds in the rest of the frames [7]. VOS is a challenging problem as target objects change their appearances dramatically due to occlusion and background distractors.

In the past decade, various VOS algorithms have been proposed to handle aforementioned challenges to fine-tune the model parameters to learn target appearances. This strategy has largely limited real-world applications due to slow speed [1, 2, 11]. In contrast, template matching and propagation-based strategies avoid online optimization. These approaches may suffer from mismatching problem due to temporal consistency and drift issue due to background disctractors [4, 13]. Many fine-tuning-free VOS algorithms have been proposed to learn pixel-wise embedding learning in Euclidean space [6, 5, 10]. These approaches require high-computation because of similarity matching in Euclidean space. During feature refinement, the decoder of VOS algorithm may also lose contextual information due to upsampling or convolutional layers [9].

In this work, we propose a network architecture to efficiently learn the guided feature information in a encoder-decoder architecture. We introduce a guided feature modulation module to learn guided feature information for better discrimination by benefiting from modulation activation. We utilize a pixel-wise semantic embedding to learn static target/background cues by utilizing a global directional matching module. On the contrary to AGAME [6], our directional generative appearance model efficiently learns dynamic target/background cues from the subsequent frames on a spherical embedding. Proposed directional appearance module estimates strong discriminative cues in a single forward pass to avoid online learning. We also propose a guided pooled decoder to learn global and local contextual information during feature refinement. Proposed VOS framework is differentiable and is trained in an end-to-end offline learning manner. We evaluate our VOS over DAVIS 2020 test-challenge and test-dev datasets.

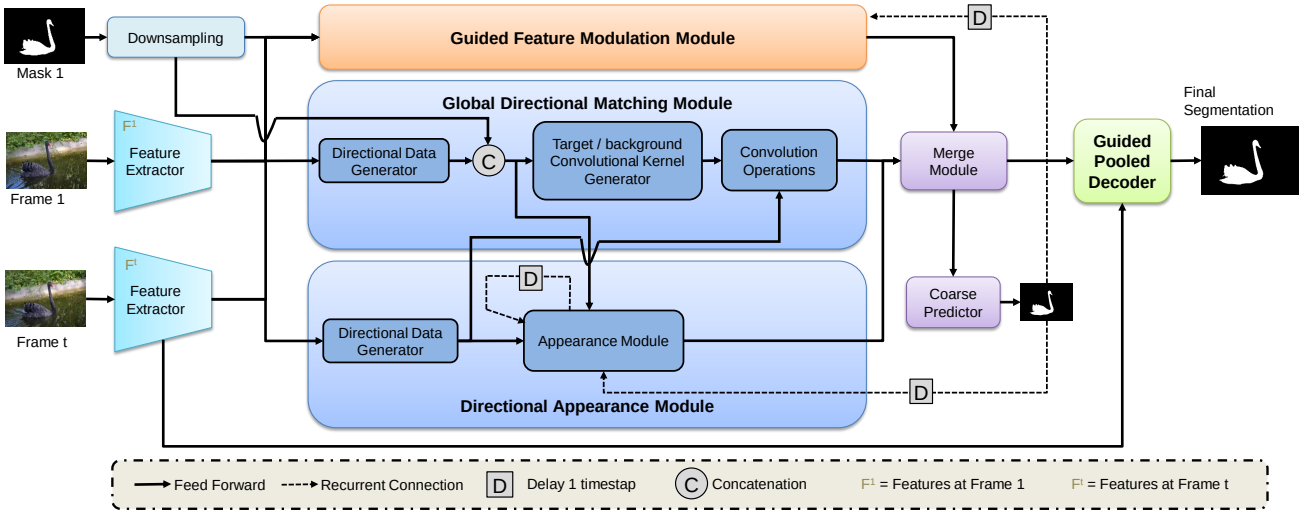

Figure 1. The illustration of proposed video object segmentation framework. The extracted features and resized mask from the first frame are forwarded to Guided Feature Modulation Module (GFMM), Global Directional Matching Module (GAMM), and Directional Appearance Module (DAM). During inference, features from the current frame are forwarded to GFMM, GDMM, and DAM. The outputs from these modules are combined in merge module and passed to coarse predictor to estimate the coarse segmentation encoding. This coarse encoding is feeded back to GFMM and DAM for next frame segmentation estimation. The merge module output with shallow features are forwarded to Guided Pooled Decoded (GPD) for refinement and to produce final segmentation result.

2. Proposed Method

This work aims to propose a model update free VOS algorithm by learning the guided context features, static pixel-wise semantic embedding from the first frame, dynamic target/background cues from subsequent frames, and guided pooled decoder in one-shot learning. Similar to DDEAL [14], our VOS framework is composed of feature extraction, Guided Feature Modulation Module (GFMM), Global Directional Matching Module (GDM), dynamic Directional Appearance Module (DAM) module, merge module, coarse predictor, and Guided Pooled Decoder (GPD) module as shown in Fig. 1. Our feature extractor ResNet extracts features from the first frame and input frame, and forward to GFMM, GDMM, and DAM. The initial mask from the first frame is resized and also forwarded to these modules. The outputs from aforementioned modules are concatenated and fused using two convolutional layers in merge module and forwarded to coarse predictor which has one convolutional layer to estimate coarse segmentation mask. This coarse segmentation mask is re-used in the next frame by GFMM and DAM. The output from merge module along with shallow features are also fed to GPD to estimate the final segmentation.

2.1. Guided Feature Modulation Module(GFMM)

The objective of GFMM is to learn the video context information in a modulation activation manner. GFMM learns guided feature information such that it preserves the semantic information. Proposed GFMM architecture is shown in Fig. 2. Out GFMM takes object masks from first and previous frame, and extracted features at frame one and t. The object masks and feature are forwarded to guided Feature Modulation Block (FMB) which returns guided features. The object masks are forwarded to convolutional layers to construct α and β. The input features are normalized and element-wise multiplied with α and element-wise added with β in a modulation manner. The outputs from FMB blocks and masks from first and previous frames are concatenated and forwarded to merge module after convolutional layer to reduce the channels.

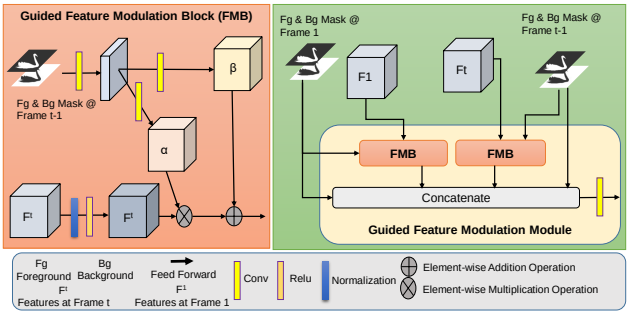

Figure 2. The illustration of Guided Feature Modulation Module (GFMM) on right and guided Feature Modulation Block (FMB) on left.

2.2. Global Directional Matching Module (GDMM)

Our GDMM learns static cues in a directional embedding manner from first frame by directional feature match-

ing. Similar to DDEAL [14], GDMM performs matching in spherical embedding space by computing cosine similarity between the directional features from first frame and current frame. Let $F^0 \in \mathcal{R}^{1xCxHxW}$ be the features extracted from first frame are the directional feature vectors in sphere. Suppose $M^0 \in \mathcal{R}^{1x1xHxW}$ is the resized target mask from first frame. The target kernel vectors $T_{hw,c,1}$ and background kernel vectors $B_{hw,c,1}$ are computed as:

$$\begin{aligned} T_{hw,c,1} &= M^0_{1,1,h,w}.F^0_{1,c,h,w}, \\ B_{hw,c,1} &= (1 - M^0_{1,1,h,w}).F^0_{1,c,h,w}. \end{aligned} \tag{1}$$

Each kernel vector in $T_{hw,c,1}$ and $B_{hw,c,1}$ represents the weighted feature for each position in F^0. The target cosine distance $P_{1,k,h,w}$ between F^t and $T_{hw,c,1}$ is computed as:

$$P_{1,hw,h,w} = \sum_{c=1}^{C} (F^t_{1,c,h,w}.T^{hw}_{hw,c,1,1}), P \in \mathcal{R}^{1xHWxHxW}. \tag{2}$$

Similarly, background cosine distance $Q_{1,hw,h,w}$ is calculated. Finally, the global directional matching $p_{1,1,H,W}$ and $q_{1,1,H,W}$ is achieved by maximizing $P_{1,HW,H,W}$ and $Q_{1,HW,H,W}$ respectively.

2.3. Dynamic Directional Appearance Module (DAM) module

Our dynamic DAM module is based on von Mises-Fisher (vMF) directional appearance model to estimate the target/background cues in a spherical embedding space. Our directional appearance model returns the posterior class probabilities for target/background for discrimination and updates for each frame. Our DDAM structure is close to DDEAL [14]. Let g be the directional vectors, the vMF distribution of a (p-1) dimensional sphere in $\mathcal{R}^p$ with $L_2 - norm$ is $f_p(g;\mu,\kappa) = (\frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}) exp(\kappa \mu^T g)$, where $\|\mu\|$ =1, $\kappa \geq 0$, and I_v be the Bessel function of p dimensionality with order v.

The directional vectors g_l are extracted from features maps F from the current frame at spatial location l. Each class-conditional density with μ and κ for vMF distribution is $p(g_l|z_l = k) = f_p(g_l;\mu_k,\kappa_k)$. The directional variable g_l is assigned to a discrete component $z_l = k$. In practice, we employed four model components i.e, 1 and 3 for directional target features while 0 and 3 are used for directional background features. The base components are computed from the masks from first frame while supplementary components are computed from previous estimated masks. The κ parameter is a trainable parameter in our model. The parameter μ is estimated by maximum likelihood as $\mu^i_k = \frac{\sum_l \alpha^i_l g^i_l}{\|\sum_l \alpha^i_l g^i_l\|}$, where $\alpha^i_{l,k} \in \{0,1\}$ be $i-th$ soft-label to assign directional feature g^i_l to a specific component k. The parameter μ^i_k is linearly updated with learning rate λ as $\mu^i_k = (1-\lambda)\mu^{i-1}_k + \lambda \mu^i_k$.

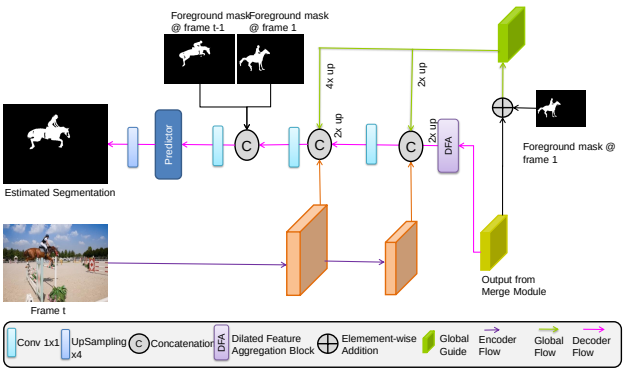

Figure 3. Illustration of proposed Guided Pooled Decoder (GPD). Low-level features are merged using global guide along with Dilated Feature Aggregation (DFA) to obtain high-level features. Foreground maps from first frame and previous frames are also combined during feature refinement. Finally, features are fed to predictor to estimate the final estimation. The final output is obtained by upsampling.

2.4. Guided Pooled Decoder (GPD) module

We propose a GPD is to generate high-resolution feature maps by capturing global and local context information. Proposed GPD explicitly preserve the salient object features at different layers and global target information from the input feature maps during refinement. We intend to provide a global guidance information throughout the feature refinement process from coarse-level to fine-grained level. For that, we introduce global guide along with a global flow to deliver the high-level semantic information at different layers (as illustrated in Fig. 3). We also introduce a Dilated Feature Aggregation (DFA) module, similar to [3], that merges the features extracted with different dilation rates.

3. Experiments

3.1. Network Training Details

We implemented the proposed framework in PyTorch by utilizing ResNet101 as backbone feature extractor, pretrained on ImageNet benchmark for object classification task. The model is end-to-end trainable based on a single ground-truth mask from the first frame and $n-th$ frame from a sequence. The model has to estimate the segmentation mask for the $n-th$ frame. We apply cross-entropy loss function to compute the final segmentation loss and auxiliary loss for coarse segmentation. We trained our network in two stages. In stage one, we set input image size to 240×432 which are normalized by computing the mean and the standard deviation. We set batch size of 8 with 12 frames from a video and trained for 120 epochs. During stage two, we used original input image size and set batch size 4 with 8 frames. We optimize our model

for 100 epochs using Adam optimizer by adding losses. The initial learning rate is 10^{-5} and it decreases exponentially till 10^{-2} with weight decay 10^{-5}. We trained our model over DAVIS2017-train [8] and YouTube-VOS-train [12] datasets.

3.1.1 Results on DAVIS 2020 Challenge

We used three measure parameters such as region similarity (J), contour accuracy (F), and global score (G) [7] to evaluate our VOS method. We achieved 57.9%, 63.9%, and 60.9% on DAVIS 2020 test-Dev in terms of J, F, and G. We also evaluate our on DAVIS 2020 test-Challenge and secured 8-th position. Our approach secured 72.1, 67.3, and 64.9 scores in terms of J, F, and G respectively. Our VOS performs segmentation at 0.06 seconds per frame on DAVIS 2020 test-Challenge dataset.

We perform ablation study over DAVIS 2020 test-dev to validate the proposed GFMM, GDMM, DAM, and DFA modules. The J, F, and G scores are shown in Table. 1.

Table 1. Ablation study performed over DAVIS 2020 test-Dev dataset.

Variants	J-mean	F-mean	G-mean
Ours-without GFM	55.4	61.4	58.4
Ours-without GDM	55.3	60.8	58.0
Ours-without DAM	54.1	58.7	56.7
Ours-without GPD	56.3	61.9	59.1
Ours	57.9	63.9	60.9

4. Conclusions

We propose a VOS algorithm to segment targets from backgrounds without model update. Proposed GDMM learns semantic information in a modulation manner from each frame. GDMM matches the semantic information from first frame for each frame. DAM gives the strong cues for mask estimation. Global guided pooled decoder captures both the global and local context information during refinement. Our method achieved an overall score 64.9 on DAVIS 2020 test-Challenge dataset at 0.06 seconds/frame.

Acknowledgement: This study was supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005)

References

[1] L. Bao, B. Wu, and W. Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2018. 1

[2] S. Caelles, K.-K. Maninis, J. P-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 1

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[4] Y. Chen, J. P-Tuset, A. Montes, and L. V Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1198, 2018. 1

[5] H. Ci, C. Wang, and Y. Wang. Video object segmentation by learning location-sensitive embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–516, 2018. 1

[6] J. Johnander, M. Danelljan, E. Brissman, F. S. Khan, and M. Felsberg. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2019. 1

[7] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 1, 4

[8] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[9] Z. Tian, T. He, C. Shen, and Y. Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3126–3135, 2019. 1

[10] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 1

[11] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. In *The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, volume 5, 2017. 1

[12] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 4

[13] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018. 1

[14] Y. Yin, D. Xu, X. Wang, and L. Zhang. Directional deep embedding and appearance learning for fast video object segmentation. *arXiv preprint arXiv:2002.06736*, 2020. 2, 3