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

CFBI+: Collaborative Video Object Segmentation by Multi-Scale Foreground-Background Integration

Zongxin Yang, Yuhang Ding, Yunchao Wei, and Yi Yang
ReLER, Centre for Artificial Intelligence, University of Technology Sydney
zongxin.yang@student.uts.edu.au, dyh.ustc.uts@gmail.com, {yunchao.wei,yi.yang}@uts.edu.au

Abstract

In this paper, we investigate the principles of embedding learning between the given reference and the predicted sequence to tackle the challenging semi-supervised video object segmentation. Unlike previous practices that only explore embedding learning using pixels from foreground object (s), we consider background should be treated equally. Thus, we propose a Collaborative video object segmentation by multi-scale Foreground-Background Integration (CFBI+) approach, which is an enhanced version of our previous CFBI [13]. Our CFBI+ implicitly imposes the feature embedding from the target foreground object and its corresponding background to be contrastive, promoting the segmentation results accordingly. With the feature embedding from both foreground and background, our CFBI+ performs the matching process between the reference and the predicted sequence using a multi-scale strategy, making the CFBI+ robust to various object scales. In our experiments and the DAVIS-2020 Challenge, we follow a standard training setting, i.e.*, pre-training on COCO and fine-tuning on Youtube-VOS 2019 and DAVIS 2017. Without any bells and whistles, our method achieves new state-of-the-art $\mathcal{J}\&\mathcal{F}$ performance of* 82.8% *and* 77.5% *on Validation and Testing sets of DAVIS 2017 respectively, while keeping an efficient run-time (about* $5FPS$ *on 480P videos). By simply applying multi-scale & flip strategies during the inference stage, our single model achieves* 81.9% *on Testing and* 82.2% *on Challenge of DAVIS 2017.*

1. Introduction

Video Object Segmentation (VOS) is a fundamental task in computer vision with many potential applications, including interactive video editing [5] and augmented reality [6]. In this paper, we focus on semi-supervised VOS, which targets on segmenting a particular object instance across the entire video sequence based on the object mask given at the first frame.

Early VOS works (*e.g.*, [4]) rely on fine-tuning with the first frame in evaluation, which heavily slows down the inference speed. Recent works (*e.g.*, [10, 7]) aim to avoid fine-tuning and achieve better run-time. In these works, STMVOS [7] introduces memory networks to learn to read sequence information and outperforms all the fine-tuning based methods. However, STMVOS relies on simulating extensive frame sequences using many large image datasets (*e.g.*, [3]) for training. The simulated data significantly boosts the performance but makes the training procedure elaborate. Without simulated data, FEELVOS [10] adopts a semantic pixel-wise embedding together with a global (between the first and current frames) and a local (between the previous and current frames) matching mechanism to guide the prediction. The matching mechanism is simple and fast, but the performance is not comparable with STMVOS.

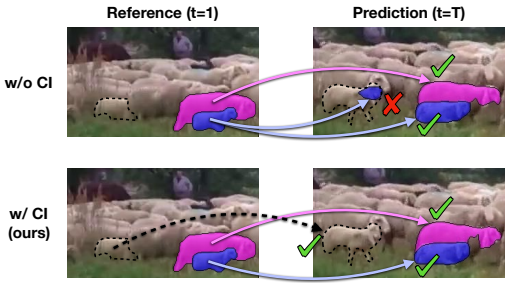

Figure 1: CI means collaborative integration. There are two foreground sheep (pink and blue) in the video. In the top line, the contempt of background matching leads to a confusion of the prediction of sheep. In the bottom line, we relieve the confusion problem by introducing background matching (dot-line arrow).

Even though the efforts mentioned above have made significant progress, current works pay little attention to the feature embedding of background in videos and only focus on exploring robust matching strategies for the foreground object (s). Intuitively, it is easy to extract the foreground from a video when precisely removing all the background. Moreover, modern video scenes commonly focus on many similar objects, such as the cars in car racing, the people in a conference, and the animals on a farm. For these cases, the contempt of integrating foreground and background embeddings traps VOS in an unexpected background confusion problem. As shown in Fig. 1, if we focus on only the

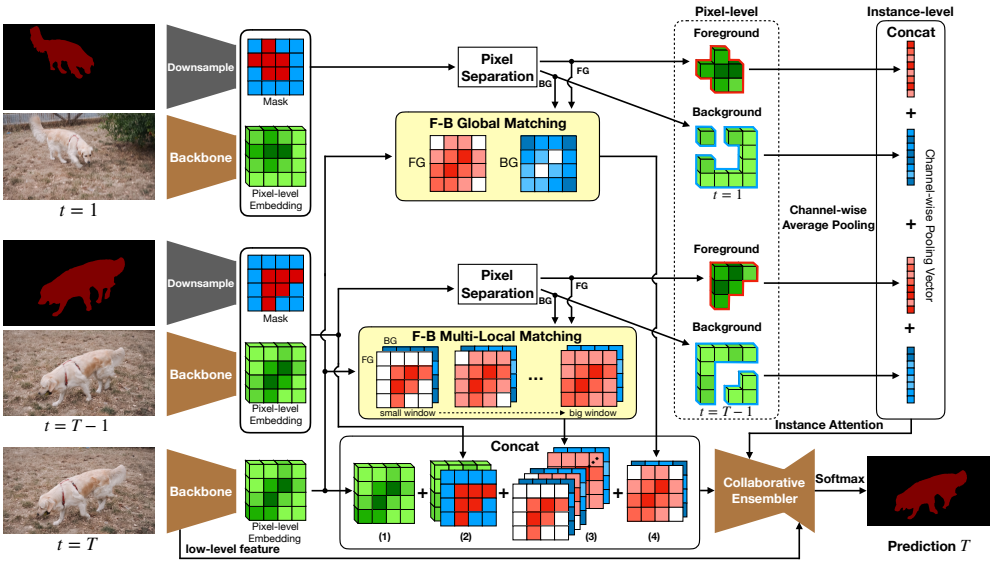

Figure 2: An **overview** of the single scale version of CFBI+, *i.e.*, CFBI. F-G denotes Foreground-Background. We use red and blue to indicate foreground and background separately. The deeper the red or blue color, the higher the confidence. Given the first frame, previous frame, and current frame, we firstly extract their pixel-wise embedding by using a backbone network. Second, we separate the first and previous frame embeddings into the foreground and background pixels based on their masks. After that, we use foreground-background pixel-level matching and instance-level attention to guide our collaborative ensembler network to generate an accurate prediction.

foreground matching like [10], a similar and same kind of object (sheep here) in the background is easy to confuse the prediction of the foreground object. Such an observation motivates us that the background should be equally treated compared with the foreground so that better feature embedding can be learned to relieve the background confusion.

Based on the above motivation, our recent work proposed a novel framework for Collaborative video object segmentation by Foreground-Background Integration (CFBI) [13]. Different from the above methods, we not only extract the embedding and do match for the foreground target in the reference frame, but also for the background region. Besides, our framework extracts two types of embedding (*i.e.*, pixel-level and instance-level embedding) for each video frame to cover different scales of features. Similar to FEELVOS, we employ pixel-level embedding to match all the details of objects with the same global and local mechanism. However, the pixel-level matching is not sufficient and robust to match those objects with larger scales and may bring unexpected noises due to the pixel-wise diversity. Thus we introduce instance-level embedding to help the segmentation of large-scale objects by using attention mechanisms. The strategies proposed above can significantly improve the quality of the learned collaborative embeddings for conducting VOS while keeping the network simple yet effective. Our CFBI (single model result) ranked 3rd in both Track 1 and 2 of the 2nd Large-scale Video Object Segmentation Challenge [12, 2].

In this paper, we develop an enhanced version of CFBI, *i.e.*, Collaborative video object segmentation by Multi-scale Foreground-Background Integration (CFBI+), to improve the robustness of predicting objects with different scales. Instead of doing pixel-level matching on a single scale as CFBI, our CFBI+ matches pixels on three different spatial scales. Besides, the higher the resolution of scale, the less the channels of feature we use to compute a matching distance. Hence, CFBI+ is able to outperform CFBI while keeping a comparable inference speed.

Without any bells and whistles, CFBI+ achieves new state-of-the-art $\mathcal{J}\&\mathcal{F}$ results of 82.8% and 77.5% on Validation and Testing sets of DAVIS-2017 [8] respectively, while keeping an efficient run-time (about $5FPS$). By applying multi-scale & flip strategies during the inference stage, our single model achieves 81.9% on Testing and 82.2% on Challenge. The latter significantly outperforms the winner (76.7%) [1] of DAVIS Challenge last year.

2. Method

2.1. Revisiting CFBI: A Single-Scale Version

To overcome the problems raised by the background confusion problem as well as promote the robustness of different scales of objects, we proposed CFBI, as shown in Figure 2. First, beyond learning feature embedding from foreground pixels, our CFBI additional consider the embedding learning from background pixels for collaboration. Such a learning scheme will encourage the feature embedding from the target object and its corresponding background to be contrastive, promoting the segmentation results accordingly. Second, with the collaboration of pixels from the foreground and background, we further conduct the embedding matching from both pixel-level and instance-level. For the pixel-level matching, we improve the

robustness of the local matching under various object moving rates. For the instance-level matching, we design an instance-level attention mechanism, which can efficiently augment the pixel-level matching. More details can be found in [13], and we only show some keys to the collaborative pixel-level matching in this paper.

2.1.1 Collaborative Pixel-level Matching

For the pixel-level matching, we adopt a global and local matching mechanism similar to [10] for introducing the guided information from the first and previous frames, respectively.

To incorporate background information, we firstly redesign the pixel distance of [10] to further distinguish the foreground and background. Let B_t and F_t denote the pixel sets of background and all the foreground objects of frame t, respectively. We define a new distance between pixel p of the current frame T and pixel q of frame t in terms of their corresponding embedding, e_p and e_q, by

$$D_t(p,q) = \begin{cases} 1 - \frac{2}{1+exp(||e_p - e_q||^2 + b_B)} & \text{if } q \in B_t \\ 1 - \frac{2}{1+exp(||e_p - e_q||^2 + b_F)} & \text{if } q \in F_t \end{cases}, \quad (1)$$

where b_B and b_F are trainable background bias and foreground bias. We introduce these two biases to make our model be able further to learn the difference between foreground distance and background distance.

Foreground-Background Global Matching. Let $\mathcal{P}_t$ denote the set of all pixels (with a stride of 4) at time t and $\mathcal{P}_{t,o} \subseteq \mathcal{P}_t$ is the set of pixels at time t which belongs to the foreground object o. The global foreground matching between one pixel p of the current frame T and the pixels of the first reference frame (*i.e.*, $t = 1$) is,

$$G_{T,o}(p) = \min_{q \in \mathcal{P}_{1,o}} D_1(p,q). \quad (2)$$

Similarly, let $\overline{\mathcal{P}}_{t,o} = \mathcal{P}_t \backslash \mathcal{P}_{t,o}$ denote the set of relative background pixels of object o at time t, and the global background matching is,

$$\overline{G}_{T,o}(p) = \min_{q \in \overline{\mathcal{P}}_{1,o}} D_1(p,q). \quad (3)$$

Foreground-Background Multi-Local Matching. We propose to apply the local matching mechanism on different scales and let the network learn how to select an appropriate local scale, which makes our framework more robust to various moving rates of objects.

Formally, let $K = \{k_1, k_2, ..., k_n\}$ denote all the neighborhood sizes and $H(p,k)$ denote the neighborhood set of pixels that are at most k pixels away from p in both x and y directions, our foreground multi-local matching between the current frame T and its previous frame $T-1$ is

$$ML_{T,o}(p,K) = \{L_{T,o}(p,k_1), L_{T,o}(p,k_2), ..., L_{T,o}(p,k_n)\}, \quad (4)$$

where

$$L_{T,o}(p,k) = \begin{cases} \min_{q \in \mathcal{P}^{p,k}_{T-1,o}} D_{T-1}(p,q) & \text{if } \mathcal{P}^{p,k}_{T-1,o} \neq \emptyset \\ 1 & \text{otherwise} \end{cases}. \quad (5)$$

Here, $\mathcal{P}^{p,k}_{T-1,o} := \mathcal{P}_{T-1,o} \cap H(p,k)$ denotes the pixels in the local window (or neighborhood). And our background multi-local matching is

$$\overline{ML}_{T,o}(p,K) = \{\overline{L}_{T,o}(p,k_1), \overline{L}_{T,o}(p,k_2), ..., \overline{L}_{T,o}(p,k_n)\}, \quad (6)$$

where

$$\overline{L}_{T,o}(p,k) = \begin{cases} \min_{q \in \overline{\mathcal{P}}^{p,k}_{T-1,o}} D_{T-1}(p,q) & \text{if } \overline{\mathcal{P}}^{p,k}_{T-1,o} \neq \emptyset \\ 1 & \text{otherwise} \end{cases}. \quad (7)$$

Here similarly, $\overline{\mathcal{P}}^{p,k}_{T-1,o} := \overline{\mathcal{P}}_{T-1,o} \cap H(p,k)$.

2.2. From CFBI to CFBI+

To further improve the robustness of predicting objects with different scales, we develop a multi-scale version of CFBI, *i.e.*, CFBI+. Instead of doing pixel-level matching on a single scale (with a stride of 4) like CFBI, our CFBI+ matches pixels on three different spatial scales (with a stride of 4, 8, 16). Besides, the higher the resolution of scale, the less the channels of feature we use to compute a matching distance. In detail, the channels are 16, 32, and 64 for the scales with a stride of 4, 8, and 16, respectively. Hence, CFBI+ is able to outperform CFBI, while keeping a comparable inference speed (about 5 FPS).

3. Experiment

We evaluate our method on Validation, Testing and Challenge of DAVIS-2017 [8]. The evaluation metric is the $\mathcal{J}$ score, calculated as the average IoU between the prediction and the ground truth mask, and the $\mathcal{F}$ score, calculated as an average boundary similarity measure between the boundary of the prediction and the ground truth, and their average value ($\mathcal{J}\&\mathcal{F}$).

We follow the same setting in CFBI [13] to train our model only on public datasets. In details, we firstly pretrain our backbone on ImageNet [9] and COCO [3]. And then, we finetune our full model on YouTube-VOS [11] and DAVIS-2017 in an end-to-end way. When evaluating on Validation, we train with Train of DAVIS-2017. When evaluating on Testing or Challenge, we train with both Train and Validation of DAVIS-2017.

3.1. Single-Scale

Our single-scale version, *i.e.*, CFBI, achieves 81.9% and 74.8% on Validation and Testing of DAVIS-2017, respectively. **Ablation Study.** As shown in Table 1, we first

P	I	Avg	$\mathcal{J}$	$\mathcal{F}$
✓	✓	74.9	72.1	77.7
✓		73.0	69.9	76.0
	✓	72.3	69.1	75.4
		70.9	68.2	73.6

Table 1: Ablation of background embedding on Validation of DAVIS-2017 ($\mathcal{J}\&\mathcal{F}$). P and I separately denote the pixel-level matching and instance-level attention.

analyze the influence of removing the background embedding while keeping the foreground only. Without any background mechanisms, the result of our method heavily drops from 74.9% to 70.9%. This result shows that it is significant to embed both foreground and background features collaboratively. More analyses and the ablation study of other proposed components can be found in [13].

3.2. Multi-Scale

By introducing a multi-scale strategy for pixel-level matching, our CFBI+ (single model) achieves **82.8%** and **77.5%** on Validation and Testing of DAVIS-2017 respectively. In other words, we achieve a new state-of-the-art performance on DAVIS-2017 without any post-processing. If we apply a multi-scale and flip strategy in evaluation, we can boost our single model result to **81.9%** on Testing, and achieve **82.2%** on Challenge, which significantly outperforms the winner (OSS [1], 76.7%) of DAVIS-2019 Challenge in last year.

4. Conclusion

We propose a novel framework for video object segmentation by combining collaborative foreground-background integration with multi-scale matching and achieves new state-of-the-art results on DAVIS-2017. We hope CFBI and CFBI+ will serve as a solid baseline and help ease the future research of VOS and related areas.

Methods	F	S	Avg	$\mathcal{J}$	$\mathcal{F}$
Validation Split					
FEELVOS [10]			71.5	69.1	74.0
PReMVOS [4]	✓		77.8	73.9	81.7
STMVOS [7]		✓	81.8	79.2	84.3
CFBI [13]			81.9	79.1	84.6
CFBI+			**82.8**	**80.1**	**85.5**
Testing Split					
FEELVOS [10]			57.8	55.2	60.5
PReMVOS [4]	✓		71.6	67.5	75.7
STMVOS [7]		✓	72.2	69.3	75.2
CFBI [13]			74.8	71.1	78.5
CFBI+			**77.5**	**74.3**	**80.7**
CFBI+^{MS}			**81.9**	**78.7**	**85.2**
Challenge Split					
OSS [1]	✓		76.7	72.7	80.6
CFBI+^{MS}			**82.2**	**79.5**	**85.0**
CFBI+^{MS*}			**83.8**	**81.1**	**86.5**

Table 2: The quantitative evaluation on DAVIS-2017 [8]. F, S, and ^{MS} separately denote fine-tuning at test time, using simulated data in the training process and using a multi-scale and flip strategy in evaluation. * denotes performing model ensemble.

References

[1] N. Wang S. Wang X. Zhang S. Liu S. Gao K. Lu D. Zhang L. Shen Y. Wang Y. Xu B. Wang, C. Zheng. Object-based spatial similarity for semi-supervised video object segmentation. *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2019. 2, 4

[2] Qianyu Feng, Zongxin Yang, Peike Li, Yunchao Wei, and Yi Yang. Dual embedding learning for video instance segmentation. In *ICCV Workshops*, 2019. 2

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 1, 3

[4] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, pages 565–580, 2018. 1, 4

[5] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*, 2020. 1

[6] King Ngi Ngan and Hongliang Li. *Video segmentation and its applications*. Springer Science & Business Media, 2011. 1

[7] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1, 4

[8] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 3, 4

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[10] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, pages 9481–9490, 2019. 1, 2, 3, 4

[11] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 3

[12] Zongxin Yang, Peike Li, Qianyu Feng, Yunchao Wei, and Yi Yang. Going deeper into embedding learning for video object segmentation. In *ICCV Workshops*, 2019. 2

[13] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. *arXiv preprint arXiv:2003.08333*, 2020. 1, 2, 3, 4