Memory Aggregated CFBI+ for Interactive Video Object Segmentation

Chen Liang[1], Zongxin Yang[2], Jiaxu Miao[2], Yunchao Wei[2] and Yi Yang[2]

[1]Zhejiang University [2]University of Technology Sydney

leonnnop@gmail.com, {zongxin.yang, jiaxu.miao}@student.uts.edu.au,
{yunchao.wei, yi.yang}@uts.edu.au

Abstract

In this paper, we propose a novel framework for tackling the interactive video object segmentation. To deal with the gradually increasing scribble information, our framework applies two independent networks for conducting user interaction and temporal propagation. For the former part, we adopt an inside-outside single-object coarse-to-fine structure augmented with a pyramid scene parsing module for aggregating global contextual information (IOI-Net). For the temporal propagation part, to record the informative knowledge from previous interaction rounds, the proposed model (MCFBI-Net) adopts a simple yet effective memory aggregation mechanism based on the Collaborative video object segmentation by Multi-scale Foreground-Background Integration (CFBI+) method, which fully utilizes the rich information from both foreground pixels and background pixels. Besides, we introduce the High Confidence Filter and the Background Random Drop Mechanism in this paper to improve the robustness in discovering challenging objects. Our approach took the 2nd place according to $\mathcal{J}\&\mathcal{F}$@60s and the 3rd place with AUC score on interactive track in DAVIS Challenge on Video Object Segmentation 2020.

1. Introduction

Video Object Segmentation (VOS) plays a significant role in various potential applications, including video editing, augmented reality, and self-driving cars, which aims at separating a foreground object from a video sequence. In this paper, we try to explore a better VOS approach in an interactive setting. Comparing with semi-supervised approaches, which need a fully annotated mask at the first frame of a video sequence, interactive VOS (iVOS) is more user-friendly and takes less time in annotation acquirement since only a few scribbles to complete the prediction of the entire video sequence. In this scheme, segmentation outputs are gradually refined by earning scribbles on the falsely predicted regions. The whole work-flow to achieve round-based interactive VOS could be found in the DAVIS-2020 challenge [1]. Inspired by some recent works [5, 10, 9, 7], in this paper, we propose a novel framework for iVOS by dividing the whole task into two separate parts: 1) Inside-outside single-object coarse-to-fine interactive network (IOI-Net); 2) Memory aggregated CFBI+ unsupervised propagation network (MCFBI-Net).

With IOI-Net, we take the sparse multi-object scribbles and generate the fully segmented mask for the annotated frame. In detail, for each object in a video sequence, we firstly extract a rough region-of-interest(ROI) and ignore all the other parts of the input. Then, the scribbles are classified into inside scribbles that belong to the current object, and the remaining outside scribbles. Both parts are concatenated and fed into a coarse-to-fine architecture network for segmentation mask generation. For the propagation part, based on the CFBI+ [8], we introduce a simple yet effective memory aggregation mechanism into the base network by maintaining a global memory unit. The mechanism helps the entire network continuously accumulate informative knowledge of the users' interaction, which further boosts the robustness for the object instances with a great variety of appearances. Furthermore, to deal with the accumulation of errors caused by an *inaccurate* interactive segmentation mask, we design a high confidence pixel filter to reduce the impact of inaccurate global information. Since the background usually contains less information, we randomly drop a portion of background pixels. Combined with Background Random Drop, this approach further reduces the number of reference pixels to ease the computation workload and accelerate the framework.

By accurately exploiting and gradually accumulating reference information between rounds, our approach could further improve the framework performance on object disappearance and occlusion state. Without any bells and whistles, our approach achieves $\mathcal{J}\&\mathcal{F}$@$60s$ performance

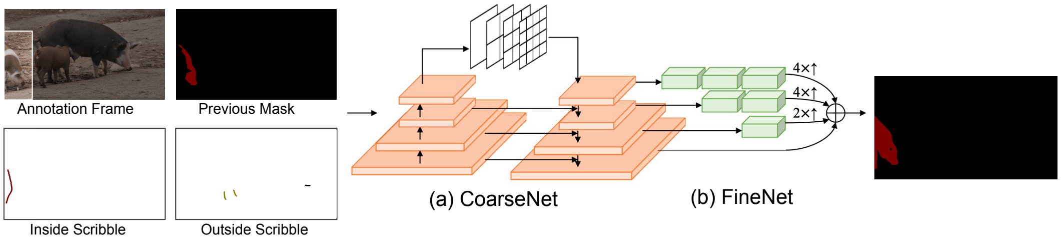

Figure 1. An overview of IOI-Net. Our model only considers one target object for each prediction. The prediction process of the piglet highlighted in *Annotation Frame* at 2 to 7 interaction round is illustrated. There are only 0 and 1 in scribble masks, even though we have drawn scribbles in color to facilitate observation. For the first interaction, only inside scribble is provided, since no previous mask or outside scribble is available.

of 76.2% and AUC performance of 75.4% on the test-dev set of DAVIS-2017[6]. In DAVIS Challenge on Video Object Segmentation 2020[1], our approach took the 2nd place according to $\mathcal{J}\&\mathcal{F}$@$60s$ and the 3rd place with AUC score on interactive track.

2. Method

In our approach, we divide the whole interactive video object segmentation task into two separate parts: An interactive handler for user annotated frame, and then a propagation handler is applied to temporally propagate the generated mask to the entire video.

2.1. Interactive Handler

Rough Region-of-interest(ROI) Extraction. To avoid the interference caused by nearby similar objects to the prediction, we take a loose restricted area as a region of interest(ROI) and crop the ROI from all channels of input to be fed into the coarse-to-fine network. According to the experimental results, we choose double as the final scale factor. In other words, the rough ROI is twice in size of annotated scribble.

Inside-outside Single-object Coarse-to-fine Interactive Network (IOI-Net). Inspired by [10], we employ a cascaded structure similar to [2]. The first sub-network, CoarseNet applies an FPN-like[4] architecture. With lateral connections between deeper and earlier layers, the semantic information is fused with low-level details. Then, a pyramid scene parsing module [11] is applied at the deepest layer for enriching the representation with global contextual information. The second sub-network, FineNet takes a coarse prediction from the CoarseNet and aggressively fuses the feature extracted by the CoarseNet across different levels, which helps recover the missing boundary details in the generated mask. A single object is processed per prediction, for multi-object scenarios, different objects are respectively fed into the network. As illustrated in Fig.1, the red pig is separated from the full mask. All the obtained results are merged according to the confidence pixel by pixel. Only cropped RGB image and inside scribble mask are feed into the network for the first interaction, both previous mask and outside scribble mask are set to zero. For the later interactions, we extract inside and outside scribbles by comparing the scribble mask with the current processing object segmentation mask, as shown in Fig.1.

2.2. Propagation Handler

Memory Aggregated CFBI+ Unsupervised Propagation Network (MCFBI-Net). In the propagation part, we treat the segmentation mask of the user annotated frame generated by IOI-Net as a reference mask and leverage the MCFBI-Net to propagate the informative knowledge of the reference mask and the previous frame to the current frame using the pixel embedding. Following CFBI+[8], we employ the global and local matching map as the soft cues of the reference frame and the previous frame, respectively. Both foreground and background are treated equally for pixel-level matching and instance-level attention to guide our collaborative ensembler network generating an accurate prediction.

Different from the semi-supervised VOS who obtains a fully annotated frame, the interactive setting only provides few scribbles in each round. The produced global matching map in each round is relatively insufficient for entirely object segmentation. To accumulate the informative knowledge from the previous interactions, similar to [5], we propose a simple yet effective memory aggregation mechanism that improves the robustness in discovering challenging objects greatly. As shown in Fig.2, a global memory unit is employed. Formally, let $M_g \in R_{n,o,h,w}$ denotes the global map memory, where n, o, h, w denotes the total number of video frames, target object, height and width of embedding feature maps, respectively. G_t denotes the extracted global matching map at time step t. We initialize M_g with all 1 in the first interaction and update M_g later by preserving the minimum value of each pixel. For the round of r and the

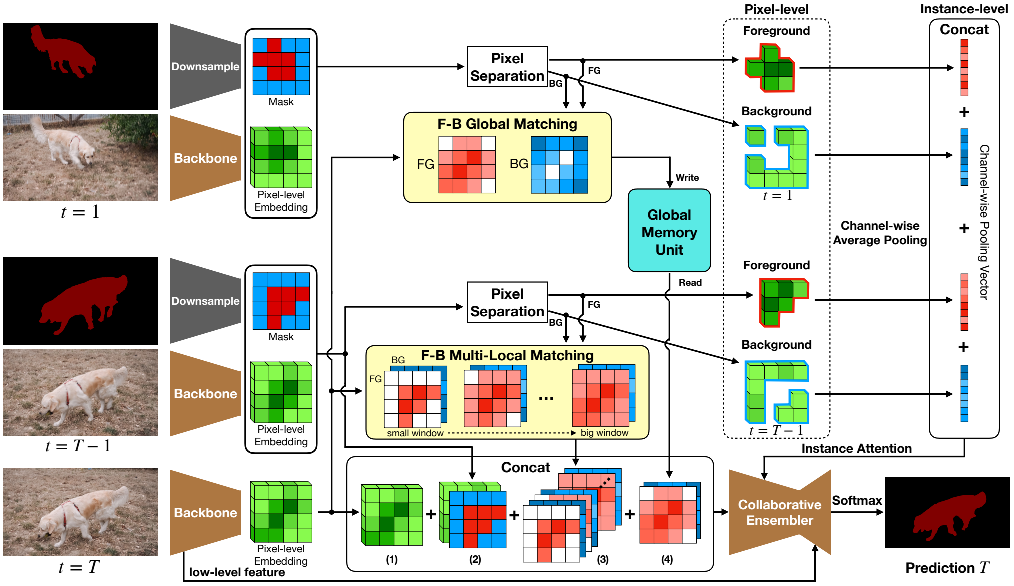

Figure 2. An overview of MCFBI-Net with the single scale version of CFBI+[8], *i.e.*, CFBI [9]. Similar to [8], we use red square or F to indicate foreground and blue square or G is background relatively. The deeper the red or blue color, the higher the confidence. A global memory unit(GMU) is employed in MCFBI, which accumulates informative knowledge among interactions. We initialize GMU with all 1 in the first interaction and update itself with a matching map generated from F-G Global Matching unit by preserving the minimum value of each pixel. The updated global map memory could be read for the collaborative ensembler.

frame at time t, $M_g^{t,r}$ is updated by

$$M_g^{t,r} = min(M_g^{t,r-1}, G_t, r)$$

The updated global map memory $M_g^{t,r}$ could be read for collaborative ensembler.

High Confidence Pixel Filter & Background Random Drop. In the actual propagation process, we find out that foreground pixels generally have higher confidence than background for interactive segmentation mask. In other words, the background pixels in the reference image is generally *inaccurate* for the propagation network, which directly leads to the misleading feature map and the chaotic propagation masks since the background pixels are equally treated in our network. Therefore, we design a high confidence filter with a threshold T = 0.7. Only pixels with a probability greater than T or less than $1 - T$ are taken into consideration. Moreover, we randomly drop out about 90% background pixels since nearby background pixels usually carry similar features. This guarantees a further increase in the speed of the network with relatively little performance loss.

3. Experiment

3.1. DAVIS Challenge 2020

We evaluate the proposed framework on the DAVIS-2017 test-dev dataset [6] with the evaluation metrics of the area under the curve (AUC), Jaccard, and boundary at 60 seconds ($\mathcal{J}\&\mathcal{F}@60s$). The IOI-Net is trained only on Pascal VOC [3] and then finetuned on the DAVIS-2017 training set. Since there is few scribble data available for training, we use the minimum spanning tree algorithm to fit the points in the ground-truth mask and generate corresponding scribbles as a supplement to the training set. For the propagation part, the same setting in CFBI+[8] is applied to train our MCFBI- Net only on public datasets.

In the challenge, each method is expected to generate the segmentation mask within 30 seconds per object for each interaction and ranks according to AUC and $\mathcal{J}\&\mathcal{F}@60s$ in an interactive manner. The proposed algorithm ranks 2nd with the performances of 76.2% $\mathcal{J}\&\mathcal{F}@60s$ and 3rd according to the AUC score of 75.4% without any bells and whistles.

3.2. Ablation Study

The Effectiveness of Memory Mechanism. We conduct ablation studies using the DAVIS-2017[6] test-dev dataset to validate the effectiveness of our proposed memory mechanism. we compare our method with and without the global memory unit, and the result is shown in Fig.3. As the number of interaction rounds increases, the performance of the framework gradually improves. Since more informative knowledge is accumulated, the baseline model with the global memory mechanism outperforms the baseline model from the second interaction round and finally achieves a

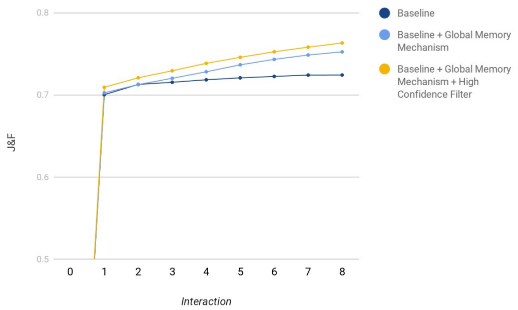

Figure 3. $\mathcal{J}\&\mathcal{F}$ performance on the test-dev set in DAVIS-2017 according to the number of interaction rounds. *baseline* denotes the framework that only applies IOI-Net and CFBI+ without memory unit, High Confidence Filter or Background Random Drop mechanism. '+' denotes we add the corresponding auxiliary mechanism when prediction.

$\mathcal{J}\&\mathcal{F}$ of 75.2% at the last interaction round. Limited by the running time, we only update 20 frames in each interaction round, the final result may be better if the time-consuming problem could be overcome. The specific numerical results could be found in Table 1.

The Effectiveness of High Confidence Filter & Background Random Drop. As shown in Fig.3 and Table 1, we evaluate the baseline model with High Confidence Filter and Background Random Drop mechanism. The final results are illustrated in the yellow line. Comparing with the baseline model which only applies CFBI+[8] for propagation, the former model boosts the $\mathcal{J}\&\mathcal{F}$ to 70.9% with an increase of about 1% at the first interaction round without memory mechanism. In the subsequent interaction rounds, the memory unit starts to work and results in a preservation of performance improvement.

4. Conclusion

In this paper, we propose a novel framework for interactive Video Object Segmentation. The entire framework consists of two separate parts: an inside-outside single-object coarse-to-fine network for user-interactive segmentation mask generation and a memory aggregated CFBI+ propagation network to separate objects from the video sequence in an unsupervised manner. More importantly, we employ the High Confidence Filter and Background Random Drop to further reduce the impact of an *inaccurate* interactive segmentation mask and boost the performance to a higher position. The proposed framework takes the 3rd place according to AUC and 2nd according to $\mathcal{J}\&\mathcal{F}@60s$ on the Interactive Scenario of the DAVIS Challenge 2020 on Video Object Segmentation.

M	HCF	BRD	$\mathcal{J}\&\mathcal{F}$
✓	✓	✓	0.763
✓	-	-	0.752
-	-	-	0.724

Table 1. Ablation study of memory mechanism and auxiliary improvement on test-dev set of DAVIS-2017 ($\mathcal{J}\&\mathcal{F}$) after final interaction. M denotes the memory mechanism. HCF and BRD separately denote the High Confidence Filter and Background Random Drop.

References

[1] Davis challenge on video object segmentation 2020 - interactive challenge. https://davischallenge.org/challenge2020/interactive.html, 2020. 1, 2

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 3

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[5] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. *arXiv preprint arXiv:2003.13246*, 2020. 1, 2

[6] 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

[7] C. Kim Y. Heo, Y. J. Koh. Interactive video object segmentation using sparse-to-dense networks. *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2019. 1

[8] Zongxin Yang, Yuhang Ding, Yunchao Wei, and Yi Yang. Cfbi+: Collaborative video object segmentation by multi-scaleforeground-background integration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2020. 1, 2, 3, 4

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

[10] Shiyin Zhang, Junhao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2