An Effective Multi-level Backbone for Video Object Segmentation

Daizong Liu [1*], Dongdong Yu[2], Minghui Dong[2], Lei Ma[2], Jie Shao[2], Jian Wang[2],
Changhu Wang[2], and Pan Zhou[1]
[1]Huazhong University of Science and Technology
[2]ByteDance AI Lab

Abstract

Semi-supervised video object segmentation (VOS) has made a great progress in recent years. Although previous works adapt abundant concepts (like memory and tracking) into VOS task, they deeply rely on the quality of image feature extracted by ResNet or Deeplab. In this paper, we carry out an investigation on the selection of extractor backbones, and propose a novel multi-level backbone to generate much higher spatial resolution representations. The core idea of the multi-level backbone is the usage of an "up-and-down" structure, which repeats "down" and "up" sampling steps to enable high spatial resolution. To avoid information losing, we propose to aggregate features across different levels to strengthen the information flow. Such structure can be easily adapted into any traditional backbone like ResNet. We find that this backbone is suitable for VOS task as it can generate much fine-grained representation, and bring the improvement of 2-3 points compared to traditional backbones on DAVIS dataset. We evaluate the improved backbone with a memory network on the DAVIS 2020 test-challenge set and achieve the $\mathcal{J}\&\mathcal{F}$ mean score of 72.2%.

1. Introduction

Semi-supervised video object segmentation (VOS) aims to segment one or more interested objects from background in a video according to the ground-truth pixels of the given objects in the first frame. Earlier methods such as [1, 14] achieve target adaption by fine-tuning a deep neural network on the first frame. Recently, instead of training deep models with multiple tricks, more and more methods [10, 15, 18] adapt abundant concepts (like memory and tracking) into VOS task and achieve significant performances. However, they deeply rely on the quality of image feature extracted by traditional backbone like ResNet. No matter what the features were used for (for distance matching or correlation), a higher resolution feature can provide more precisely details for accurate segmentation.

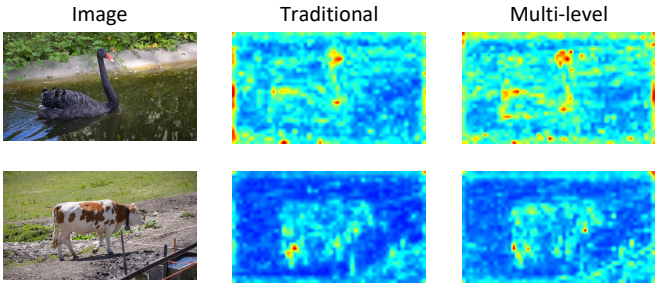

Figure 1. Comparison of the features extracted by different backbones, where ours provides more boundary details.

In this paper, we aims to develop an effective backbone to generate higher resolution features for VOS task. As up sampling [2] and deconvolution [16, 13, 19] are generally appended after the backbone networks to increase the spatial resolution of deep features, we combine such "up" and "down" stream a a "up-and-down" block, and repeat this "up-and-down" block to build our final multi-level backbone. Specifically, each level is a simple light-weight network and contains its own down sampling and up sampling path. The feature maps between the levels remain a high resolution. To avoid information losing during the down and up steps, we aggregate the feature maps across different levels to strengthen the final information flow. Such multi-level structure enables high spatial resolution and can be adapted to any traditional backbone. Figure 1 shows the comparison on feature resolution of different backbones, and our multi-level backbone provides more accurate boundary details. We evaluate our backbone with a memory network on the DAVIS 2020 test-challenge set and achieve a the $\mathcal{J}\&\mathcal{F}$ mean score of 72.2%. Ablation study is also carried out on DAVIS dataset.

2. Multi-level Backbone

We illustrate our multi-level backbone in Figure 2. For each "down and up" block, it has its own down and up sam-

*This work was performed while Daizong Liu worked as an intern at ByteDance AI Lab.

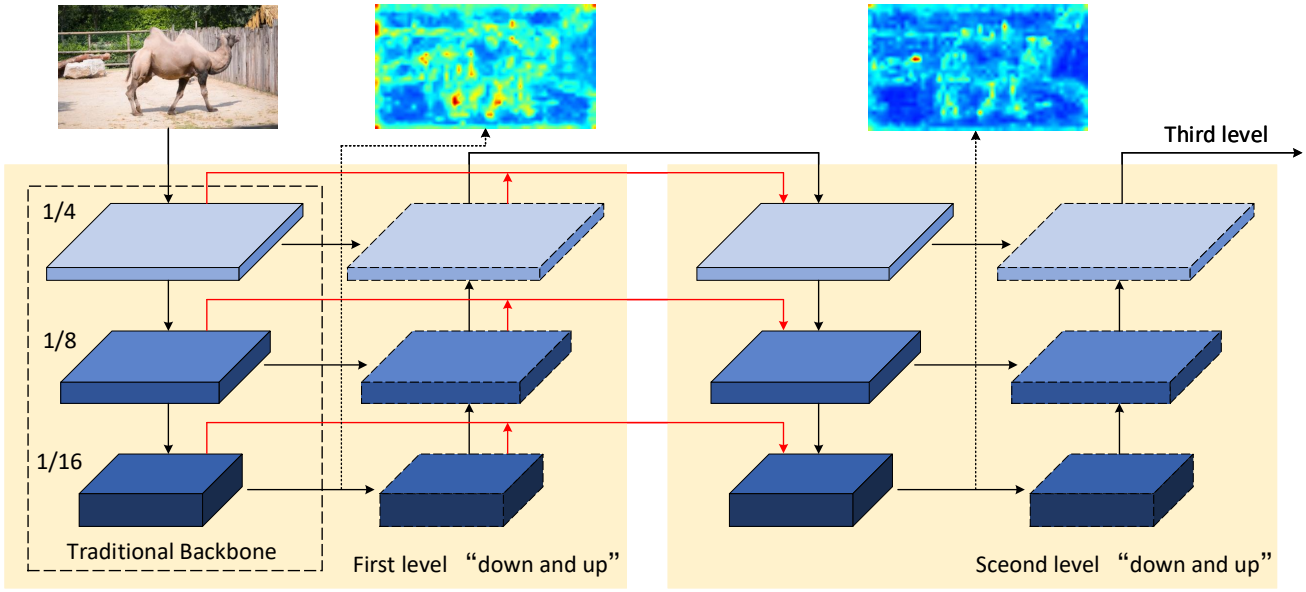

Figure 2. Overview of our proposed multi-level backbone. It is composed of multiple "down and up" blocks which enable higher spatial resolution. A cross level aggregation strategy (red line) is adopted between adjacent stages. (Best view in color)

pling path, where the down sampling is the same to the traditional backbone. We repeat such block multiple times to build the multi-level backbone. As this equal-channel-width design results in a relatively poor performance since a lot of information will be lost after every down sampling. It is reasonable since we aim to extract more representative features in the down sampling process and the lost information can hardly be recovered in the up sampling. To this end, we introduce a aggregation strategy to strengthen the information flow for increasing the capacity of down sampling unit.

2.1. "Down and Up" Block

At down sampling path, we follow the down sampling strategy of traditional backbone to embed the image into different scales. At up sampling path, we upsample the embeded features with bilinear interpolation. To keep previous low-dimension features with the same scale, we add them directly and send it to the next upsample layer.

2.2. Cross-level Feature Aggregation

As our multi-level structure is vulnerable by the information losing during repeated up and down sampling. To mitigate this issue, we propose a cross level feature aggregation strategy to propagate multi-level features from earlier levels to current one for strengthening.

As shown in Figure 2, for each scale unit of current down sampling procedure, two separate information flows (red lines) are introduced from down and up sampling units in the previous level by additional 1×1 convolution layers. With this design, the current level can take full advantage of prior information to extract more discriminative features.

3. Experiments

In this section, we evaluate ResNet50 based multi-level backbone with a memory network.

3.1. Dataset and Evaluation Metrics

We evaluate our models on DAVIS 2017 test-dev set and DAVIS 2020 test-challenge set, with mean intersection-over-union ($\mathcal{J}$), mean contour accuracy ($\mathcal{F}$) and their global mean value ($\mathcal{G}$).

3.2. Training Procedure

The model is first pre-trained on a simulation dataset generated from static image data (Pascal VOS [4, 5], COCO [9], MSRA10K [12], ESCCD [3]), and then trained for real-world videos through the main training. For fine-tuning, we apply random translation and scaling on the first frame with lucid augmentation [7].

3.3. Implemental Details

We implement our multi-level backbone based on ResNet50 with a memory network [10] in Pytorch [11].For the training, we adopt Adam [8] optimizer with learning rate $1e-5$ for pre-training and $5e-6$ for main-training and fine-tuning. The input size for the network is made to a fixed 384×384, and we use cross-entropy loss function. All the experiments are conducted on 4V100 GPUs on a server, where the batch size is set to 4 on each GPU.

Figure 3. Qualitative examples of our method on DAVIS 2020 test-challenge set, where the images are sampled at the average intervals for each video. From top to bottom, the sequences are "boxing", "choreography", "e-bike", "kids-turning", and "running" on the DAVIS2020 test-challenge set. Different objects are highlighted as different colors..

3.4. Analysis

We first investigate the performances of different training strategies. Details are shown in Table 1. It suggests that pre-training is necessary for this network, and YouTube-VOS [17] dataset can help the model converge to a better feature space.

We also investigate the performances of different backbones. As shown in Table 2, we first compare the results between ResNet50 [6] and ResNest50 [20], and find that ResNet50 performs better. We apply multi-level structure on ResNet50, and the result shows that it has 2.8% improvement with the help of the multi-level structure.

At last, we shows the tricks we used in DAVIS challenge 2020 in Table 3. We finetune the full-trained model on each object of a video with lucid [7] augmentation. This step boosts most with improvement of 4.1%. As we find that different model has different performances on different videos, we ensemble the full-trained model, pre-trained model and finetuned model to generate the final segmentation results.

3.5. Qualitative Results

Qualitative results are shown in Figure 3. We can find that our multi-level backbone provides better boundary details for segmentation, and it is robustness for object missing and occlusion.

Table 1. The results of Pre-training, Main-training and Full-training with ResNet50 based Multi-level backbone on DAVIS2017 test-dev set.

Training Method	$\mathcal{G}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
Pre-training only	61.3	59.4	63.3
(with YouTube-VOS)	63.4	60.9	65.9
(with Main-training)	61.9	60.0	63.7
Main-training only	55.5	54.3	56.6
Full-training	**66.7**	**64.4**	**69.0**

Table 2. The results of different backbones without finetuning on DAVIS2017 test-dev set.

Backbone	$\mathcal{G}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
ResNest50	63.9	61.8	66.0
ResNet50	64.4	62.8	66.1
Multi-level+ResNet50	**66.7**	**64.4**	**69.0**

Table 3. The results on DAVIS2020 test-challenge set.

Method	boost	$\mathcal{G}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
Full-training	-	66.4	63.8	69.0
+ finetuning	4.1	70.5	68.0	72.9
+ ensemble	1.7	**72.2**	**69.8**	**74.6**

4. Conclusion

In this work, we investigate the quality of the traditional backbone in VOS task. We find that higher quality of image feature can provide better guiding clues for the final segmentation. To this end, we develop an effective multi-level backbone to generate higher spatial resolution features. This structure is suitable for VOS and bring about 2-3 points improvement. Moreover, it can be easily adapted to any traditional backbone for future work. We achieve the $\mathcal{J}\&\mathcal{F}$ mean score of 72.2% on DAVIS challenge 2020.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 221–230, 2017.

[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 (CVPR)*, pages 7103–7112, 2018.

[3] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Cision*, 88(2):303–338, 2010.

[5] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 991–998, 2011.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, 127(9):1175–1197, 2019.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 740–755, 2014.

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9226–9235, 2019.

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[12] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2015.

[13] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5674–5682, 2019.

[14] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.

[15] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2019.

[16] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[17] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[18] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 314–323, 2019.

[19] Dongdong Yu, Kai Su, Jia Sun, and Changhu Wang. Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[20] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.