Supplementary Material to "DMVOS: Discriminative Matching for real-time Video Object Segmentation"

Peisong Wen[1,2], Ruolin Yang[3,4], Qianqian Xu[1], Chen Qian[4]
Qingming Huang[1,2,5,6], Runming Cong[7], Jianlou Si[4] *
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing University of Posts and Telecommunications, Beijing, China
[4]SenseTime, Beijing, China
[5]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China
[6]Peng Cheng Laboratory, Beijing, China
[7]Institute of Information Science, Beijing Jiaotong University, Beijing, China
wps_@mail.nankai.edu.cn,yangruolin@bupt.edu.cn,xuqianqian@ict.ac.cn
{qianchen,sijianlou}@sensetime.com,qmhuang@ucas.ac.cn,rmcong@bjtu.edu.cn

1 NETWORK ARCHITECTURE DETAILS

We describe more details on the Siamese encoder and the fast decoder of our model. An illustration is provided in Fig.1.

Encoder. The Siamese encoder is a fully convolutional network adopted from ResNet-50 [2] and pre-trained on ImageNet [3]. We remove the FC layers, and replace the last downsampling layer with a dilated convolutional layer to preserve more fine-grained information. In order to reduce the amount of calculation, a 1×1 convolutional layer and a 3×3 convolutional layer are connected to the end of the encoder, reducing the output feature dimension to 256. The features from the last layer are employed for the instance center offset prediction and the correlation calculation. We then apply a 2×2 average pooling to the template feature for less computational complexity.

Decoder. After the fusion feature map $\mathbf{M}_{fusion}$ is extracted, we apply a pyramid decoder to generate the final segmentation. The decoder is mainly built with Multi-Scale Blocks [8] and Residual Blocks [2]. We also utilize Squeeze Blocks to decreases the computational complexity. Details on the Squeeze Block and the Multi-Scale Blocks are shown in Fig.1. The features from the stage-2, stage-3 and stage-4 of the encoder are fed to the decoder to introduce low-level features. Note that the normalization layers are instance normalization layers [7] in our decoder.

2 MORE INFERENCE DETAILS

The input image of the network is cropped and resized based on the approximate position of the object. Inspired by related work on video object tracking [1], we utilize temporal smoothing to prevent prediction jitter. Denote the width and the height of the predicted mask at timestamp t by $\tilde{w_t}$ and $\tilde{h_t}$. The bounding box for cropping is updated as follows:

$$
\begin{aligned}
w_t &= 0.5 \times w_{t-1} + 0.5 \times 1.5 \times \tilde{w_t} \\
h_t &= 0.5 \times h_{t-1} + 0.5 \times 1.5 \times \tilde{h_t} \\
w_0 &= 1.5 \times \tilde{w_0} \\
h_0 &= 1.5 \times \tilde{h_0}
\end{aligned}
\tag{1}
$$

where w_t and h_t are the updated width and height. Afterward, the image patches are resized to 480×854.

3 VIDEO COMPARISONS ON DIFFERENT METHODS

In the attached video file *Comparisons.mp4*, we provide qualitative comparison with two state-of-the-art VOS methods, including RANet [8], PReMVOS [4], and RGMP [9]. The videos are sample from the DAVIS [5],[6] benchmark. The results of other methods are obtained through the official code.

4 FAILURE CASES

We analyze the shortcomings of our model by showing some failure cases in the video file *FailureCase.mp4*. The instance center offset prediction relies on the instance center obtained in the previous frame, and noise may be generated when the target is largely occluded. In addition, in the multi-object segmentation task, processing each target individually cannot make full use of the information between the targets, and overlaps will occur. We look forward to addressing these two issues in future work.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*. Springer, 850–865.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

*Corresponding author.

[4] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. 2018. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*. Springer, 565–580.
[5] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 724–732.
[6] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).
[7] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).
[8] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. 2019. Ranet: Ranking attention network for fast video object segmentation. In *IEEE International Conference on Computer Vision*. 3978–3987.
[9] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. 2018. Fast video object segmentation by reference-guided mask propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7376–7385.

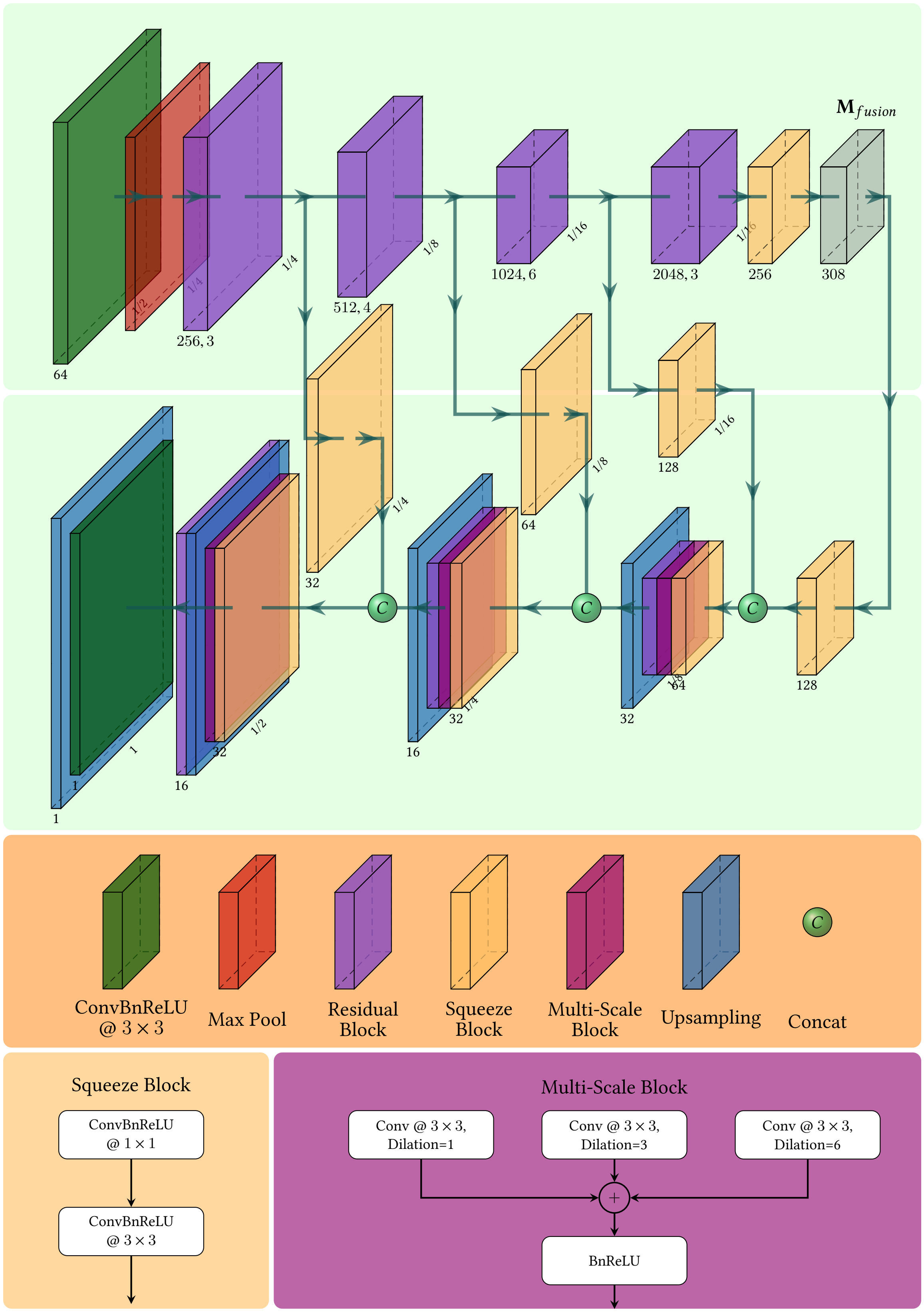

Figure 1: An illustration of the encoder and the decoder architecture.