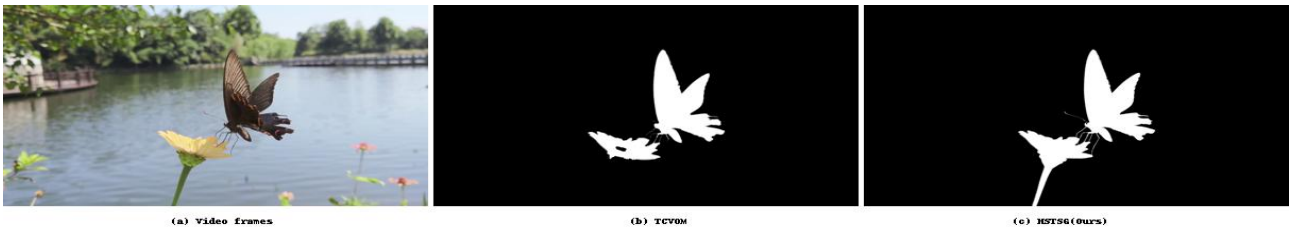
Video Object Matting via Hierarchical Space-Time Semantic Guidance
Supplementary Material

Yumeng Wang[1,2,*], Bo Xu[1,*], Ziwen Li[1], Han Huang[1], Cheng Lu[3], and Yandong Guo[1†]
[1]OPPO Research Institute, [2]Northwestern Polytechnical University, [3]Xmotors
yandong.guo@live.com



Animated comparison. Given a RGB video sequence (left), one of the recent state-of-the-art VOM approach - TCVOM [3] (center), fails to produce accurate alpha mattes. However, our proposed HSTSG VOM (right) is able to maintain the robustness of the temporal coherence for the alpha prediction. *We suggest readers view this animated figure by Adobe Reader.*

1. Overview

We provide additional details in this supplementary Material. In Section 2, we elaborate details of training and the setting of the hyper-parameters. In Section 3, we explain the loss functions in the training stage. In Section 4, we show more visual comparisons on several real-world video sequences.

2. Training details

More training details and the setting of hyper-parameters are listed in Table 1. The implementation of the hierarchical space-time semantic guidance video object matting (HSTSG) is based on the public PyTorch toolbox and trained on the Tesla V100 GPU. We first train our HSTSG on the DIM [2] dataset for 1.0×10^5 iterations. After initializing the model with the pre-trained weights, we proceed to train the HSTSG on the VideoMatting108 dataset (VMD) [3] and DVM dataset [1] for 2.0×10^5 iterations.

Parameter	Value
Optimizer Adam	Adam
Weight decay	1.0×10^{-7}
Learning rate	1.0×10^{-5}
Batch Size	4
Input image size	512×512
Number of Pre iterations	1.0×10^5
Number of VMD iterations	2.0×10^5
Number of DVM iterations	2.0×10^5
Loss weight λ_1	1
Loss weight λ_2	1
Loss weight λ_3	3
Loss weight λ_4	1
Loss weight λ_5	1

Table 1. Training implementation details and hyper-parameter setting.

3. Loss Functions

We utilize the binary cross-entropy (BCE) function to supervise the semantic mask prediction, as follow:

$$L_m = \mathrm{BCE}(m,\hat{m}) \tag{1}$$

where $\hat{m}$ is the predicted semantic mask and m is the Gaussian transformed thumbnail of the ground truth alpha matte. For the alpha prediction, we compute the pyramid Laplacian loss L_{lap}, alpha matte loss L_a and foreground loss L_f

*Yumeng Wang and Bo Xu contribute equally.
†The corresponding author.

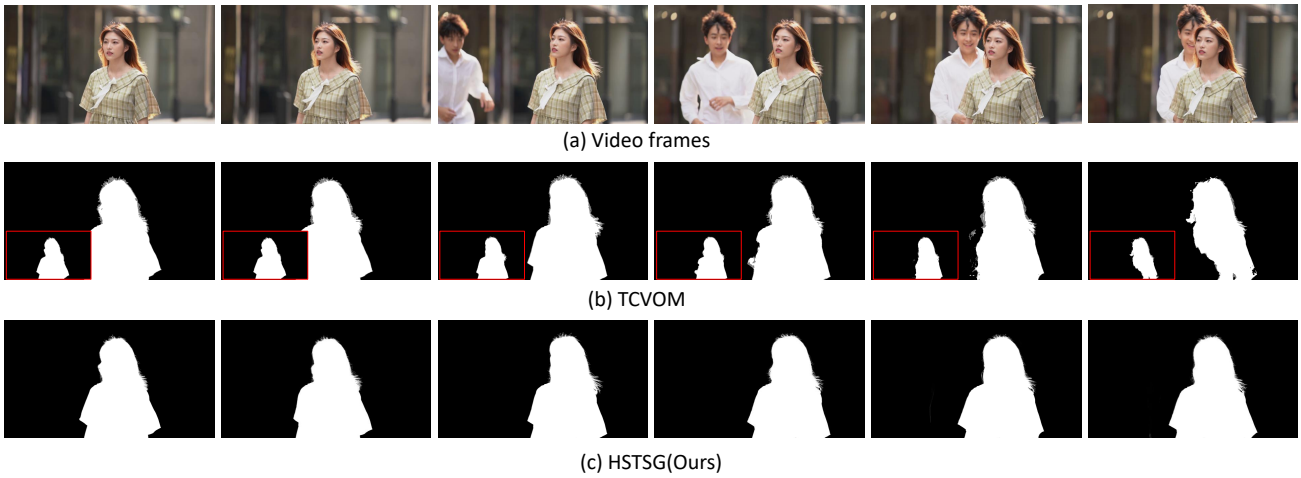


Figure 1. Visual comparisons on the real-world high-resolution video 1. Red boxes represent trimaps generated by TCVOM [3].

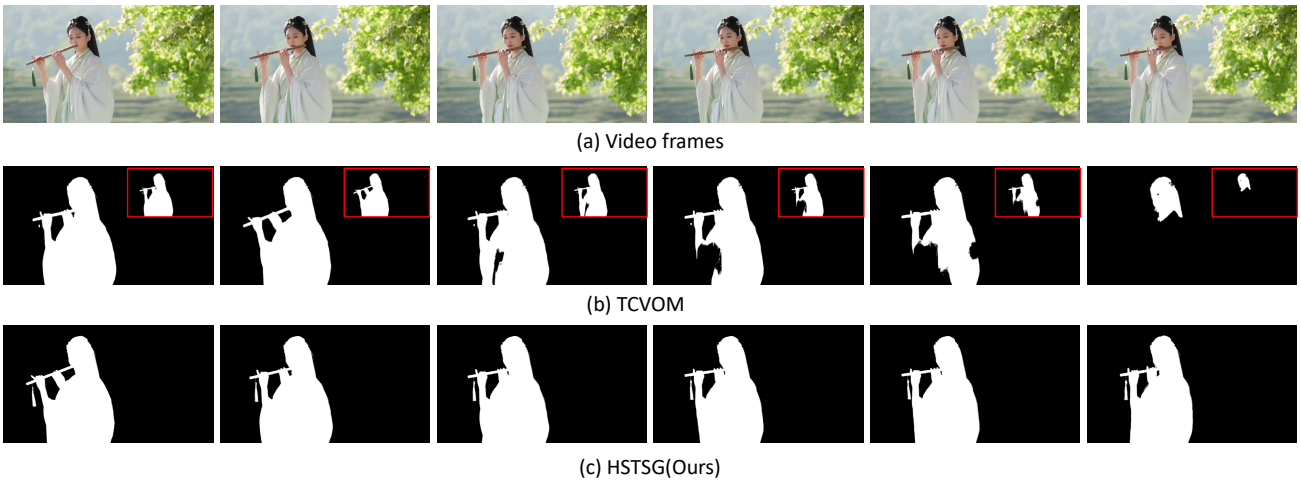


Figure 2. Visual comparisons on the real-world high-resolution video 2.

as follows:

$$L_a = \|a - \hat{a}\|_1 \tag{2}$$

$$L_f = \left\|f - \hat{f}\right\|_1 \tag{3}$$

$$L_{lap} = \sum_{i=1}^{5} 2^{i-1} \|a - \hat{a}\|_1 \tag{4}$$

where $\hat{a}$ and a represent the predicted and the ground truth alpha mattes separately. $\hat{f}$ and f represent the predicted and the ground truth foregrounds. To learn more refined alpha matte, we only measure pixels where $0 < a < 1$ by L_1 loss:

$$L_a{}^* = \|(a - \hat{a}) * (0 < a < 1)\|_1 \tag{5}$$

The total matting loss L_{total} is:

$$L_{total} = \lambda_1 L_m + \lambda_2 L_a + \lambda_3 L_m + \lambda_4 L_m + \lambda_5 L_{lap} \tag{6}$$

where λ_i is listed in Table 1.

4. More visual comparisons on real-world video sequences

We display more representative visualizations on real-world high-resolution video sequences in Figure 1, 2 and 3. Given the limited number of user-annotated trimaps, TCVOM [3] is limited by insufficient robustness in the temporal coherence of its trimap propagation. Visual comparisons also demonstrate the robustness of our proposed hierarchical space-time semantic guidance mechanism in temporal coherence. Meanwhile, our cross-frame attention refinement network can also enhance the temporal smoothness of predicted alpha mattes.

References

[1] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE/CVF Con-*

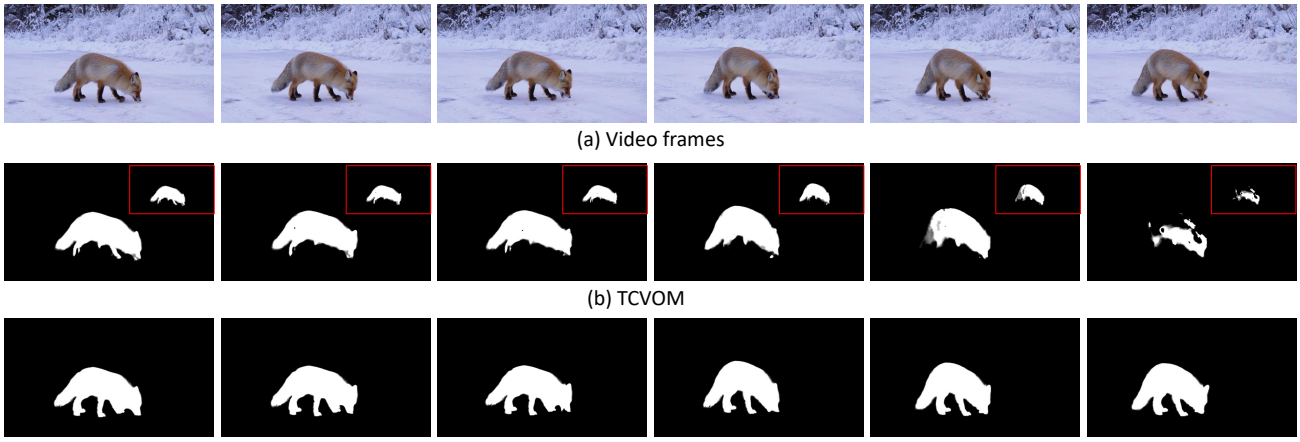


Figure 3. Visual comparisons on the real-world high-resolution video 3.

ference on Computer Vision and Pattern Recognition, pages 6975–6984, 2021.

[2] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017.

[3] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5128–5137, 2021.