End-to-end Video Matting with Trimap Propagation

Wei-Lun Huang
National Taiwan University
r09944040@csie.ntu.edu.tw

Ming-Sui Lee
National Taiwan University
mslee@csie.ntu.edu.tw

1. More Experiments

The comparison of performance and efficiency. Fig. 1 illustrates the speed and performance of different methods, where RVM has 3.74M parameters with 86.62 FPS. FTP-VM performs reasonably well on various datasets and achieves a good balance between performance and efficiency. Moreover, FTP-VM is able to matte different objects while RVM only works on humans.

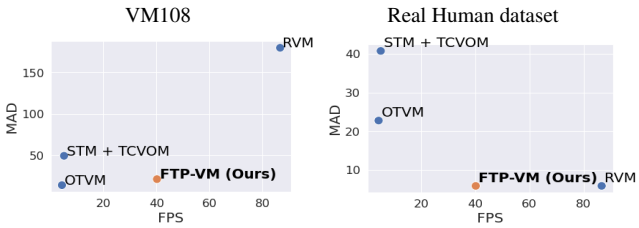


Figure 1. The comparison of speed and performance.

Ablation on hyperparameters of loss functions. Tab. 1 displays the ablation studies of γ in $\mathcal{L}_{\text{focal}}$ (Eq. 3), the weight of the correct and wrong classes in $\mathcal{L}_{\text{consis}}$ (Eq. 4), and the weight of $\mathcal{L}_{\text{tc}}$ (Eq. 9).

Ablation	Hyperparam	MSE ↓	MAD ↓	Grad ↓	dtSSD ↓	Conn ↓
γ in $\mathcal{L}_{\text{focal}}$	0	9.35	27.68	5.00	2.72	16.85
	2	**5.19**	**20.74**	**3.92**	**2.46**	**12.13**
	5	6.75	25.27	5.00	2.65	15.01
weight of classes in $\mathcal{L}_{\text{consis}}$	0.8 : 0.1	**4.79**	21.07	4.18	2.54	12.29
	0.5 : 0.25	5.19	**20.74**	**3.92**	**2.46**	**12.13**
weight of $\mathcal{L}_{\text{tc}}$	1	7.64	25.55	4.53	2.51	14.37
	5	**5.19**	**20.74**	**3.92**	**2.46**	**12.13**
	10	10.73	31.15	5.35	2.59	18.82

Table 1. The ablation studies of hyperparameters.

Ablation on Bottleneck Fusion Module. Fig. 2 shows the architecture of the bottleneck fusion module. In addition to the memory matching, CBAM [6] and PPM [8] are adopted to aggregate the features. In Tab. 2, all settings have a similar processing speed and number of parameters. However, the training fails as the performance degrades after removing any one of them. The results demonstrate that feature aggregation modules are essential in memory matching.

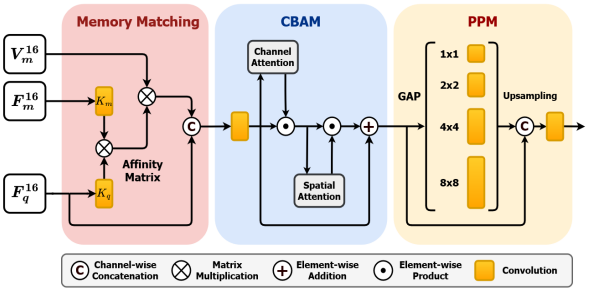


Figure 2. The bottleneck fusion module.

Method	MSE ↓	MAD ↓	Grad ↓	dtSSD ↓	Conn ↓
w/o CBAM	62.47	86.10	55.00	7.58	49.50
w/o PPM	8.53	26.23	5.23	2.69	15.07
w/ both	**5.19**	**20.74**	**3.92**	**2.46**	**12.13**

Table 2. Ablation on bottleneck fusion module.

Different Memory Trimap Width. In this experiment, memory trimaps with different dilation kernels $k \times k$ are given to evaluate the robustness of the proposed models, where k is set to be 11, 25, 41, and 81, respectively. The larger the dilation kernel, the thicker the gray strokes in the resultant trimap. Tab. 3 shows that the results with $k = 11$, 25 and 41 are similar. However, when $k = 81$, the results are degraded with a noticeable margin because this super coarse trimap contains small foreground regions (white area). As illustrated in Fig. 3, the matting result with $k = 81$ is unsatisfactory.

Fig. 4 presents the performance of different kernel sizes in terms of MAD as varying the trimap updating period. The shorter the updating period, the better the results expected. We can see that the results of trimaps with kernel size 81×81 are unstable, especially when the trimaps are updated more frequently. It explains that the aggressive coarse trimaps interrupt the temporal coherence and introduce noise in the unknown regions. On the contrary, precise trimaps provide much more pleasing results.

Scaling Factor k	MSE ↓	MAD ↓	Grad ↓	dtSSD ↓	Conn ↓
11	5.16	20.19	3.92	2.47	11.87
25	5.20	20.74	3.92	2.47	12.14
41	5.50	21.87	4.22	2.56	12.75
81	8.30	30.28	6.19	3.03	17.61

Table 3. Results of different k values.

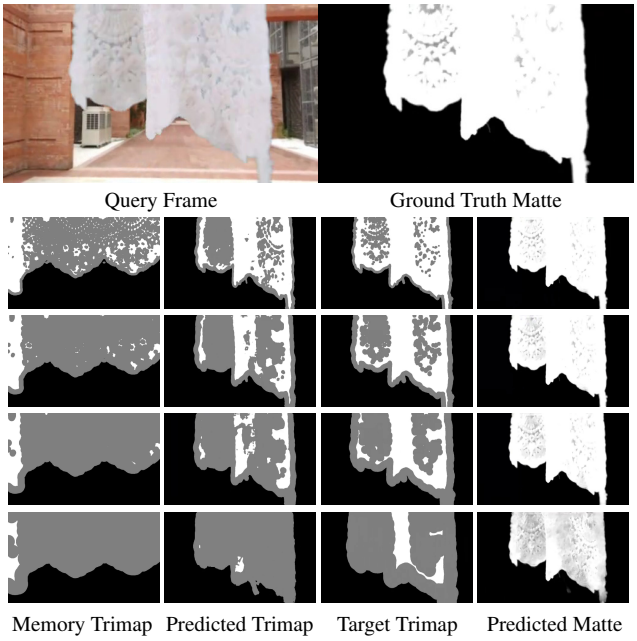


Figure 3. Qualitative comparison of different kernel size. The rows from top to bottom are k =11, 25, 41 and 81, respectively.

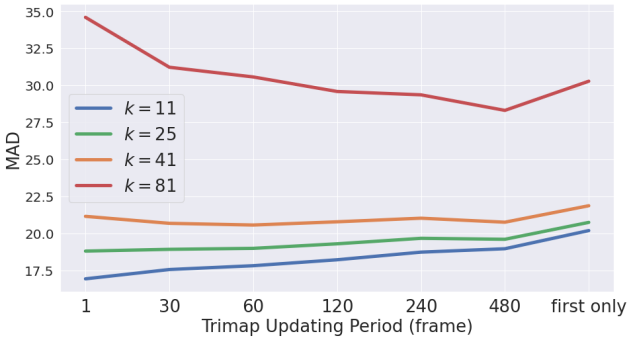


Figure 4. MAD of different trimap updating periods with different kernel sizes.

2. High-Resolution Videos

Following RVM [4], FGF (Fast Guided Filter) [1, 2], a technique for edge-aware filtering, is utilized to process high-resolution videos. All the input frames are downsampled and passed through the proposed model to produce the mattes in low resolution. FGF computes local linear coefficients of the mattes and applies the coefficients to the original images to produce high-resolution results. Instead of directly passing the original frames through the network, the reasons why FGF is adopted are listed as follows. First, FGF can be utilized flexibly without any training process. Besides, as a resolution gap between training and testing might exist, downsampling the inputs and applying FGF solve this issue successfully. Last but not least, the progress can be sped up and the computational cost is reduced owing to the smaller input resolution.

Extending experiments with various downsampling scales s are conducted on VM108 in HD and the corresponding evaluations are detailed in Tab. 4. Three scales are tested where FGF is applied to cases with $s = 0.5$ and 0.25. Grad and dtSSD of the original resolution are better than the other two settings due to unavoidable artifacts caused by FGF. It can be observed from Fig. 5, the hair details are well-captured with $s = 1$. For MAD, MSE and Conn, those metrics reveal that the model attempts to focus on the details in the left background of Fig. 5 with original resolution, which causes undesired errors. Considering the trade-off between the overall performance and the processing speed, the proposed model is feasible in interactive applications.

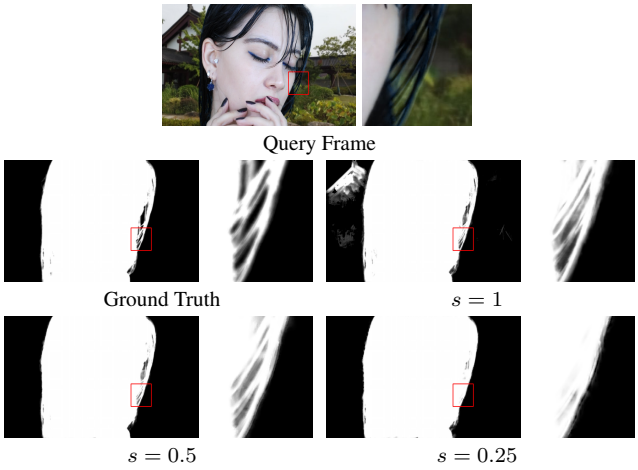


Figure 5. Qualitative results on VM108 (HD).

To examine the effects of FGF, the experiment is further conducted on VM240k in 4K. The downsampling scales are set to 0.25 and 0.125, which can attenuate the effect of FGF. Since the computation cost of 4K is too high, Conn is not evaluated here. Tab. 5 displays that MSE, MAD and dtSSD are almost the same after applying FGF. Only Grad has a considerable improvement because FGF is an edge-aware filter. Although FGF enhances the details powerfully, the noise is also intensified simultaneously, such as the boundary of the hair in Fig. 6.

3. Videos in the Wild

The proposed model, FTP-VM, not only works nicely for human videos but is also applicable to videos of

s (Downsampling scale)	FPS ↑	MSE ↓	MAD ↓	Grad ↓	dtSSD ↓	Conn ↓
1	6.57	6.57	23.13	**9.23**	**2.79**	48.03
0.5	42.78	**5.36**	**20.98**	10.93	2.85	**42.57**
0.25	**106.93**	9.63	27.55	20.05	3.57	56.51

Table 4. Results of VM108 validation set [7] with the resolution of HD (1920 × 1080).

s (Downsampling scale)	Apply FGF [1]	FPS ↑	MSE ↓	MAD ↓	Grad ↓	dtSSD ↓
0.25		44.06	**0.65**	4.91	19.47	**1.77**
	✓	41.29	**0.65**	**4.90**	**16.91**	1.88
0.125		**111.08**	2.48	7.83	43.92	2.86
	✓	97.29	2.27	7.63	29.96	2.67

Table 5. Results of VM240k validation set [3] with the resolution of 4K (3840 × 2160).

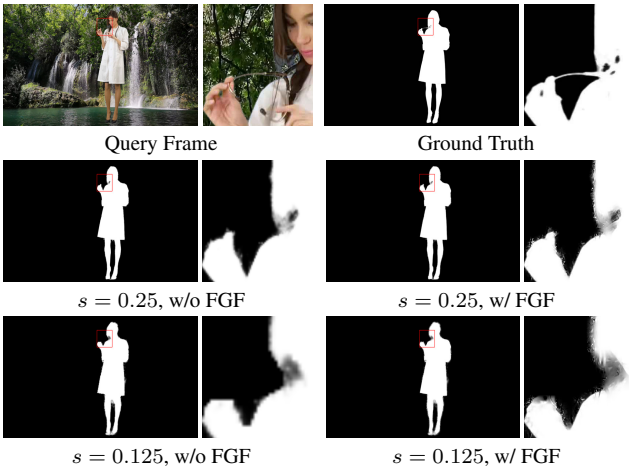
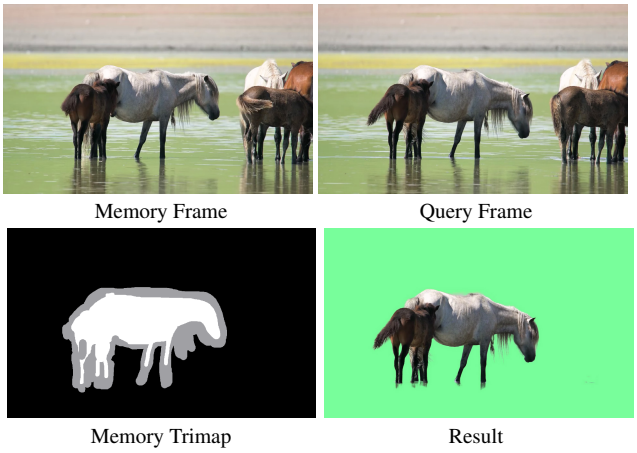


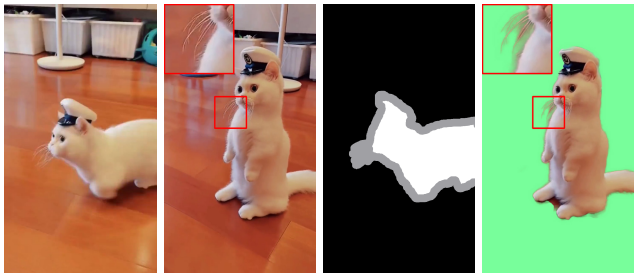
Figure 6. Qualitative results on VM240k (4K).

various objects like animals and animated characters. Two examples are demonstrated in Fig. 7. The first one is a professional video of 1920 × 1080, which is stable and contains details of horses. The other is a video acquired with a hand-held device containing blur and blocky artifacts due to poor compression. It is clear to see that the matting results are satisfactory in both cases. Fig. 7a shows that the matting result is not affected by other similar objects, while Fig. 7b confirms that the details of the whiskers can be captured even if the input quality is not good.

Fig. 8 displays the results of an animation video which is an unseen video type in the dataset. An animation seldom contains semi-transparent objects. On the contrary, objects in the animation often have precise edges. Capturing the target is not complicated in most cases. However, the results shown in Fig. 8 find that OTVM cannot matt the target character correctly. These experimental results confirm the robustness and applicability of the proposed FTP-VM in various scenarios.



(a) An examplar video with high quality.



(b) An examplar video with low quality.

Figure 7. Results of animal videos in different scenarios.

References

[1] Kaiming He and Jian Sun. Fast guided filter. *arXiv preprint arXiv:1505.00996*, 2015. 2, 3

[2] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 2

[3] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution

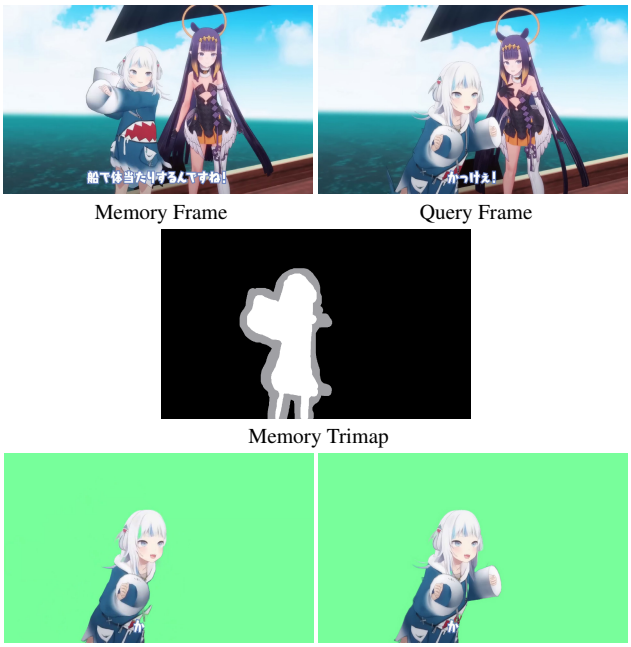


Figure 8. Qualitative comparison on the animation.

background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 3

[4] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 2

[5] Hongje Seong, Seoung Wug Oh, Brian Price, Euntai Kim, and Joon-Young Lee. One-trimap video matting. In *European Conference on Computer Vision*, 2022. 4

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[7] 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. 3

[8] 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. 1