Video Matting via Consistency-Regularized Graph Neural Networks

Anonymous ICCV submission

Paper ID 9687

1. Effect of Updates with T Iterations on CRGNN

In Section 3.1 of the manuscript, we discuss that the feature aggregation step and node state update step are alternatively executed by T iterations. We conduct experiments to study the effect of T. Table 1 shows the results when T is set as 1, 3 and 5 (we cannot use more iterations due to memory issues). Overall, we obtain better results when more iterations are used for the CRGNN model. The performance of the message passing converges at $T = 3$.

#Iterations	MSE	SAD	Gradient	Connectivity	MESSDdt
T=1	5.887	36.16	29.89	33.67	0.385
T=3	5.722	34.44	28.21	30.59	0.330
T=5	5.720	34.39	28.34	30.69	0.336

Table 1: Effect of update iteration T on the CRGNN.

2. Ablation Study on the Real-World Dataset

In Section 5.2 of the manuscript, we conduct the ablation study on the composited dataset to analyze the effect of each essential component of the proposed method. Table 2 shows the ablation study results on the real-world dataset. Overall, the GNN model, consistency regularization and discriminator can enhance the performance and the proposed deformable convolution based aggregation method performs better than the non-local aggregation method.

		MSE	SAD	Gradient	Connectivity	MESSDdt
Variants	Baseline	31.68	9.922	120.6	80.23	3.991
	+GNN	28.51	8.849	100.7	78.67	3.546
	+Consistency	27.37	8.642	95.43	76.88	3.392
	+Discriminator	**26.32**	**8.340**	**92.40**	**76.31**	**3.150**
Non-local aggregation		29.65	9.867	115.6	79.85	3.882

Table 2: Ablation study of the variants of the proposed network on the real dataset. 'Baseline' means the image-level model without using the GNN. '+' means the progressive connection of different modules.

3. Visual Results

Visual results for ablation study. The novel components of the proposed CRGNN are the (1) GNN based inter-frame relationship modeling module, (2) consistency regularization, and (3) adversarial learning scheme. In the manuscript, we show the the quantitative results for the ablation study about the effectiveness of these components. Here, we present the visual results for the ablation study.

To analyze the contribution of each component of our CRGNN, we introduce a baseline model by removing the inter-frame relationship. That is, the image-level baseline model using the encoder-decoder structure similar to [3]. Each video frame is forwarded into our baseline model frame by frame. As shown in Figure 1(a), GNN generates better details compared

to the image-level model in the first row, which benefits from the introduction of multiple frames in enhancing the temporal coherence.

To analyze the effectiveness of the consistency scheme, we provide the results with and without prediction consistency in Figure 1(b). Compared to the results without utilizing the alpha, foreground and frame consistency, utilizing the consistency regularization can generate better results. The performance gain can be attributed to the better feature representation enhanced by the consistency regularization.

In addition, Figure 1(c) shows that the introduction of the discriminator can further improve the performance based on the consistency regularization, which benefits from the advantages of the discriminator to distinguish if the image belongs to the composited image or the real one.

Failure cases. One limitation of the proposed CRGNN is that it may include the background noise when foreground and background share much similarity for the transparent objects, as shown in Figure 2. This is not unexpected because the contrast between foreground and background is weak such that it is even difficult for humans to precisely differentiate between the foreground and background.

Visual comparisons with state-of-the-art methods. We provide more visual examples in Figure 3, 4, 5 and 6 on the composited dataset and real dataset. The trimap of the composited dataset is generated based on the alpha matte. The alpha matte of the real dataset is not available, which is generated based on the erosion-dilation of the segmentation map predicted by the DeepLabv3 [1]. Compared to the existing image based methods DIM [5], LF [6], CAM [2], IM [3] and video based method BM [4], our model generate better boundary details and can suppress the background noise better. BM [4] uses an extra background as the input of the network, we do not provide the results of BM [4] on the real-world dataset because the background cannot be acquired.

(a)

(b)

(c)

Figure 1: Visual comparisons for different variants.

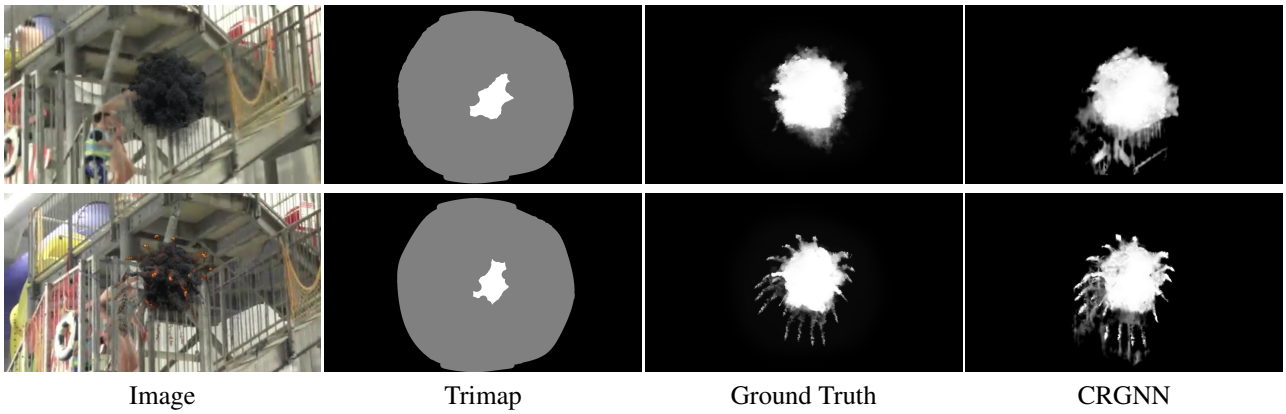

Figure 2: Failure cases.

Figure 3: Visual comparisons on the composited human matting dataset.

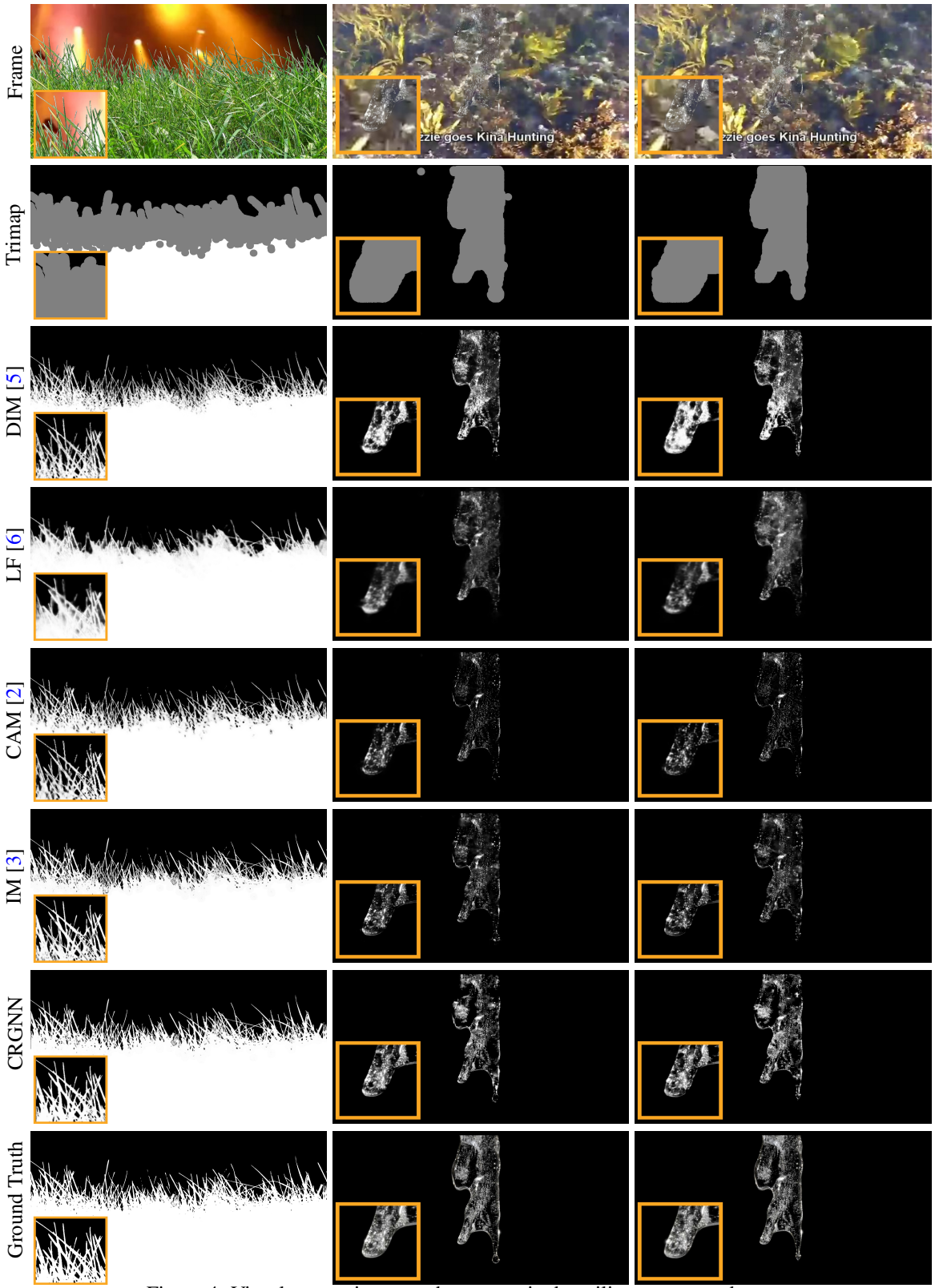

Figure 4: Visual comparisons on the composited auxiliary category dataset.

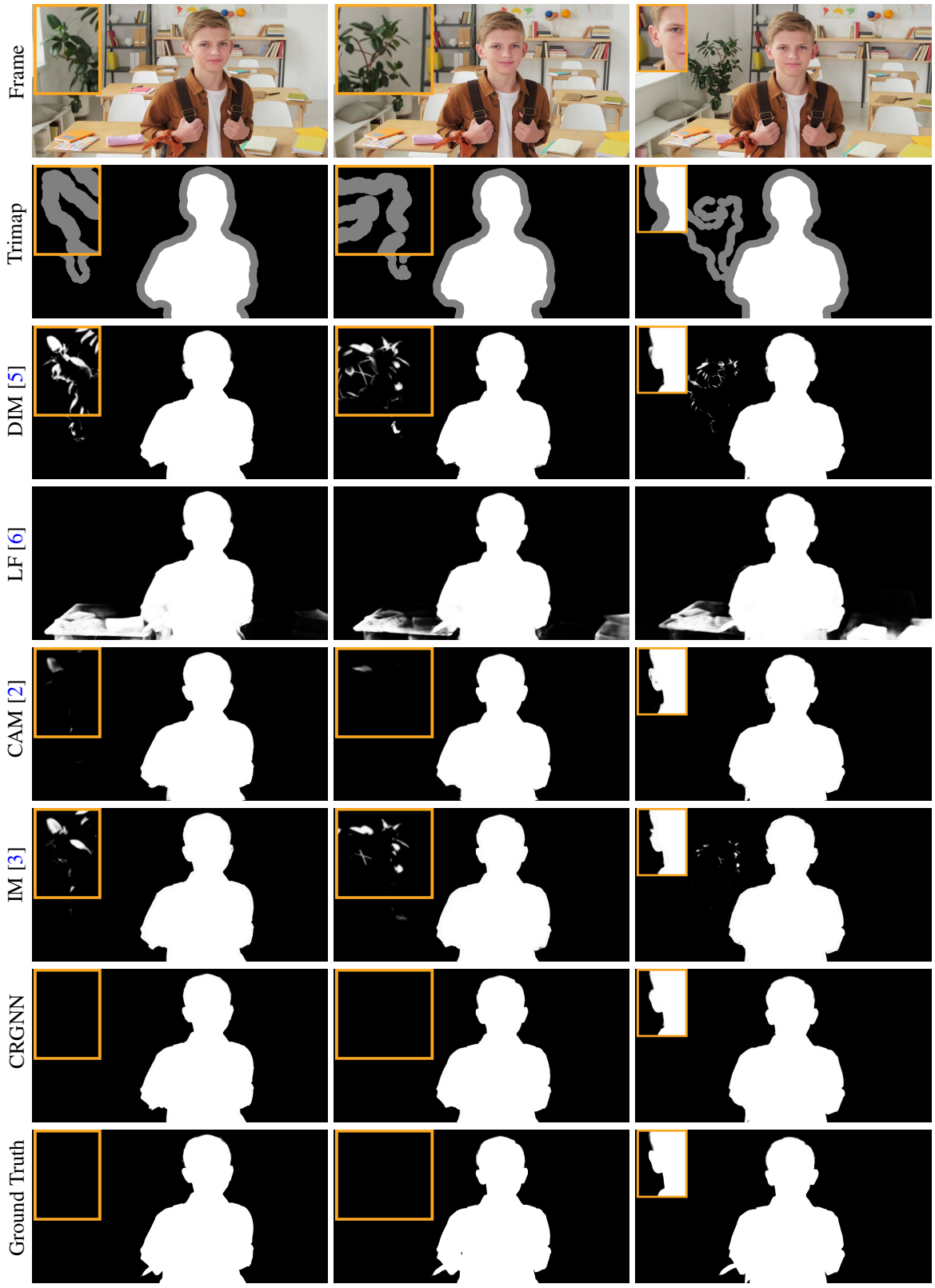

Figure 5: Visual comparisons on the real-world dataset.

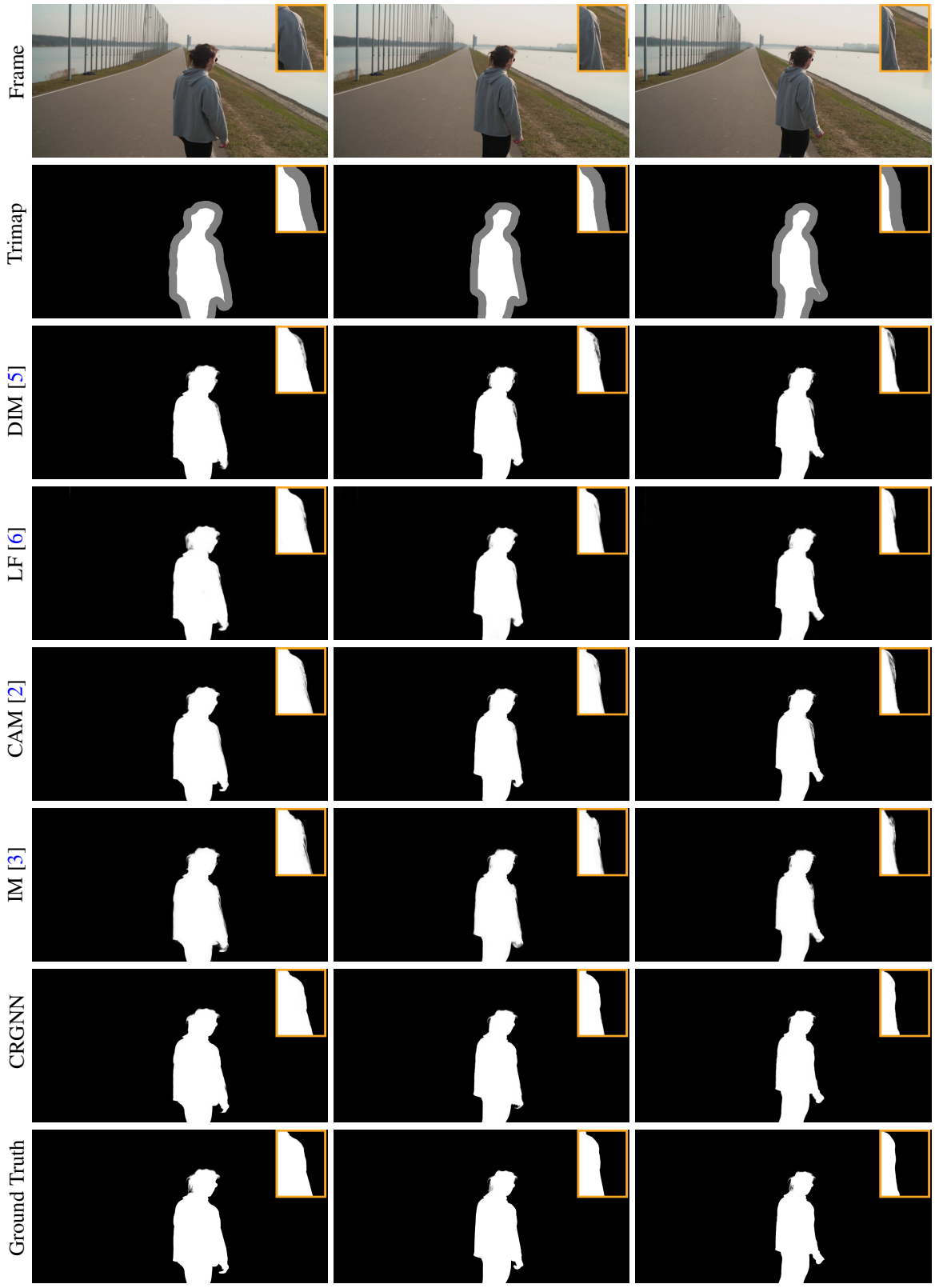

Figure 6: Visual comparisons on the real-world dataset.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[2] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019. 2, 4, 5, 6, 7

[3] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019. 1, 2, 4, 5, 6, 7

[4] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 2, 4

[5] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017. 2, 4, 5, 6, 7

[6] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, 2019. 2, 4, 5, 6, 7