information among the whole video and help recover the missing predictions by the feature aggregation. Compared to the video-based method BM [34], the proposed method achieves better performance because the CRGNN assisted by the deformable feature aggregation can fully mine the interactions between frames.

Results on the real dataset. To further verify the efficacy of the proposed method, we evaluate the results on the proposed real-world dataset. The quantitative results are shown in Table 1b. We see that our CRGNN performs best among all methods, which demonstrates the efficacy of our core idea of formulating the video matting as the combination of GNN and consistency regularization technique.

Qualitative results. Figure 5 and 6 show the visual results on the composited and real video datasets. From these results, we can clearly see that the proposed method predicts more subtle details of the frames, such as the grass in the second column of Figure 5 and suppresses the background better as shown in the second column of Figure 6. These further substantiate the superiority of the proposed method for the video matting task.

		MSE	SAD	Gradient	Connectivity	MESSDdt
Variants	Baseline	10.21	90.23	130.7	67.23	15.31
	+GNN	9.480	78.38	123.2	62.81	13.45
	+Consistency	9.260	73.21	115.4	60.75	12.69
	+Discriminator	9.223	73.49	112.1	58.49	12.23
Number of nodes	#5	9.230	74.62	115.7	58.53	12.30
	#7	9.228	73.77	115.2	58.50	12.27
Non-local agg.	-	9.954	89.45	128.9	65.68	13.56

Table 3: Ablation study on the variants of the proposed network. 'Baseline' means the image-level model without using the GNN. '+' means the progressive connection of different modules.

5.2. Ablation Study

We perform an ablation study to investigate the effect of each essential component of the proposed method.

Effectiveness of the proposed graph neural network. To analyze the contribution of our CRGNN, we introduce a baseline model by removing the inter-frame relationship, that is, the image-level baseline using the encoder-decoder structure similar to [28]. Each video frame is forwarded into our baseline model frame by frame. As shown in the second row of Table 3, GNN indeed brings significant performance improvements compared to the image-level model in the first row, which benefits from the introduction of multiple frames in enhancing the temporal coherence.

Effectiveness of the consistency regularization strategy. To investigate the effectiveness of the consistency scheme, we provide the results with and without prediction consistency in Table 3. Compared to the results without utilizing the alpha, foreground and frame consistency (the second row), utilizing the prediction consistency can generate better result, (e.g. MSE: 9.260 v.s. 9.480). The performance gain is derived from the better feature representation enhanced by the consistency regularization.

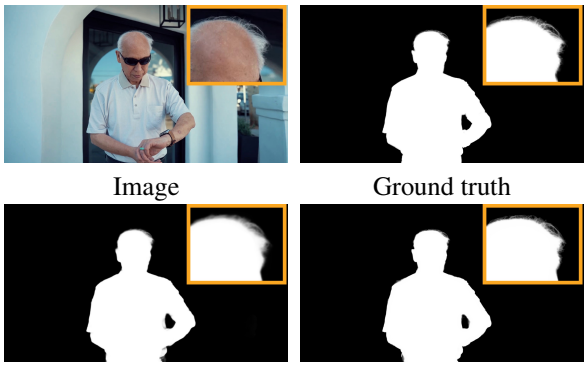

Figure 7: Visual comparison of deformable aggregation and the non-local aggregation on the real dataset.

Effectiveness of the adversarial learning scheme. The fourth row in Table 3 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.

Comparison of different number of nodes. We report the performance using the different number of nodes during the test stage. As shown in Table 3, increasing the number of nodes generates comparable results.

Comparison with the non-local structure. The non-local structure [45] has been widely used for feature aggregation on various tasks, such as video object segmentation [31] and object detection [47]. Features are aggregated by enumerating all possible positions in the embedding space. As shown in Table 3, the proposed method can generate better results comparing to utilize the non-local structure for aggregation.

6. Conclusion

In this paper, we focus on enhancing the temporal coherence for matting in videos. Different from the previous methods built on the image matting models, we propose to maintain the temporal consistency by fully exploiting the inter-frame relationship among the whole video. We use a graph neural network to relate adjacent frames with the aid of annotated synthesized video matting datasets. To generalize the proposed model from synthesized videos to real-world videos, we propose a regularization scheme to enforce the consistency on the alpha, foreground and predicted frames. In addition, we annotate a real-world dataset with alpha mattes to evaluate the efficacy of the proposed method. Extensive experiments on the synthesized and real datasets show the proposed CRGNN model performs favorably against the state-of-the-art methods.

7. Acknowledgements

This work is supported in part by the NSF CAREER Grant #1149783.