	MSE	SAD	Gradient	Connectivity	MESSDdt
DIM [48]	10.69	79.87	74.54	72.75	7.676
IM [28]	9.216	81.56	64.97	63.72	5.595
IM* [28]	5.734	54.31	43.82	44.68	3.297
LF [49]	20.61	113.0	168.7	108.2	13.90
CAM [23]	20.97	145.5	147.5	116.2	9.867
BM [34]	13.57	90.15	130.8	84.85	7.388
Ours	**3.770**	**45.77**	**30.80**	**33.81**	**2.475**

(a) Composited dataset.

	MSE	SAD	Gradient	Connectivity	MESSDdt
DIM [48]	13.32	98.92	129.1	88.56	17.48
IM [28]	10.91	95.07	120.0	73.05	14.45
IM* [28]	13.84	97.09	136.9	84.57	17.89
LF [49]	29.61	141.4	168.5	131.7	32.58
CAM [23]	11.62	101.0	123.9	78.21	14.93
Ours	**9.224**	**73.50**	**112.1**	**58.49**	**12.23**

(b) Real dataset.

Table 1: Quantitative results on the two human matting datasets. To better show the performance difference, the numbers for the above measures have been scaled up or scaled down. The scaling factors of the five measures from left to right are 1000, 0.01, 0.01, 0.01, 1000. IM* means we re-train IM using the proposed dataset. The best results are in **bold**.

	MSE	SAD	Gradient	Connectivity	MESSDdt
DIM [48]	25.03	402.1	167.4	407.4	16.47
IM [28]	37.30	582.8	115.3	597.1	16.67
LF [49]	49.25	478.7	339.0	466.3	25.07
CAM [23]	25.95	461.3	92.97	468.6	11.70
Ours	**20.65**	**378.8**	**87.54**	**365.0**	**10.41**

Table 2: Results on the auxiliary category dataset.

videos and the annotations are carefully manually created using Adobe After Effects and Photoshop. Figure 4 shows some examples from the proposed datasets.

Composited video dataset. Because of the increasing interest in human matting on videos, we propose a composited dataset with the human category (composited human matting dataset). We also provide a dataset with categories except for the human (auxiliary category dataset) to verify the generalization of our model on both the human category and other categories. Videos in these two datasets are annotated against the green screen or simple background. Because of the simplicity of the backgrounds, it is easy to generate high-quality alpha mattes and the corresponding foregrounds for each video. For the human matting dataset, there are 20 training videos (6312 frames) and 10 test videos (3807 frames). For the auxiliary category dataset (e.g., cat, plant), 20 training videos (3983 frames) and 10 test videos (1722 frames) are provided. To enlarge the diversity of the dataset, each foreground video is composited with varied backgrounds using the groundtruth alpha mattes.

Real video dataset. To measure the performance of natural videos, we also collect a real-world human matting dataset with 19 videos. The alpha and foreground are manually annotated at every 10 frames with a frame rate of 30 fps for each video, which in total results in 711 frames being labeled.

5. Experiments

We use the data augmentation scheme to increase the diversity of the input data. First, we randomly crop the image and trimap pairs centered on pixels in the unknown regions with varied resolutions (e.g. 480×480, 640×640, 960×960) and resize them to 480×480 due to the memory constraint. We also utilize random rotation, scaling, shearing as well as the vertical and horizontal flipping for the affine transformation. Our model is first pretrained on the image matting dataset [48] and then finetuned using the labeled composited data and unlabeled real data. For the image matting dataset, we use the random affine transformation to generate a short video clip with 3 frames to imitate the motion flow of the objects. Because it is hard to generate the pseudo trimap with the category like transparency, we only utilize the proposed graph neural network on the auxiliary category dataset and adopt the full model on the human matting dataset for training and inference. In the test stage, the trimaps for all datasets are generated from the ground-truth alpha mattes by thresholding and the unknown region is dilated with the kernel size 25.

We adopt the similar encoder and decoder structures introduced in [28]. We remove the last two pooling layers so the output size of the encoder is 1/8 of the input image. The decoder D_a and D_f for predicting the alpha and foreground have same structures except for the prediction layer. The output channels for the prediction layer to predict the alpha and foreground are set to 1 and 3. For the discriminator, we adopt the structure proposed in PatchGAN [24]. All the weights in objective L_{adapt} and L_{adv} used to balance different losses are set to 1. The number of vertices K and the number of iteration step T are set to 3. The running speed is about 1 fps on one single Nvidia 2080 Ti GPU.

5.1. Comparative Results

Evaluation metrics. To show the effectiveness of the proposed method, we evaluate the results on five popular metrics, including the SAD, MSE, Gradient [33], Connectivity [33] and temporal coherence (MESSDdt) [17]. These metrics can be used to evaluate the accuracy of the alpha matte for every single frame and the temporal coherence within a video. The first four metrics are widely used for image-level matting evaluation. However, long-range videos own more features compared to the image. One key feature is the temporal coherence which means the objects move among different frames should be consistent for better human perceptibility.

Results on the composited dataset. We first evaluate the proposed algorithm and state-of-the-art methods on the pro-

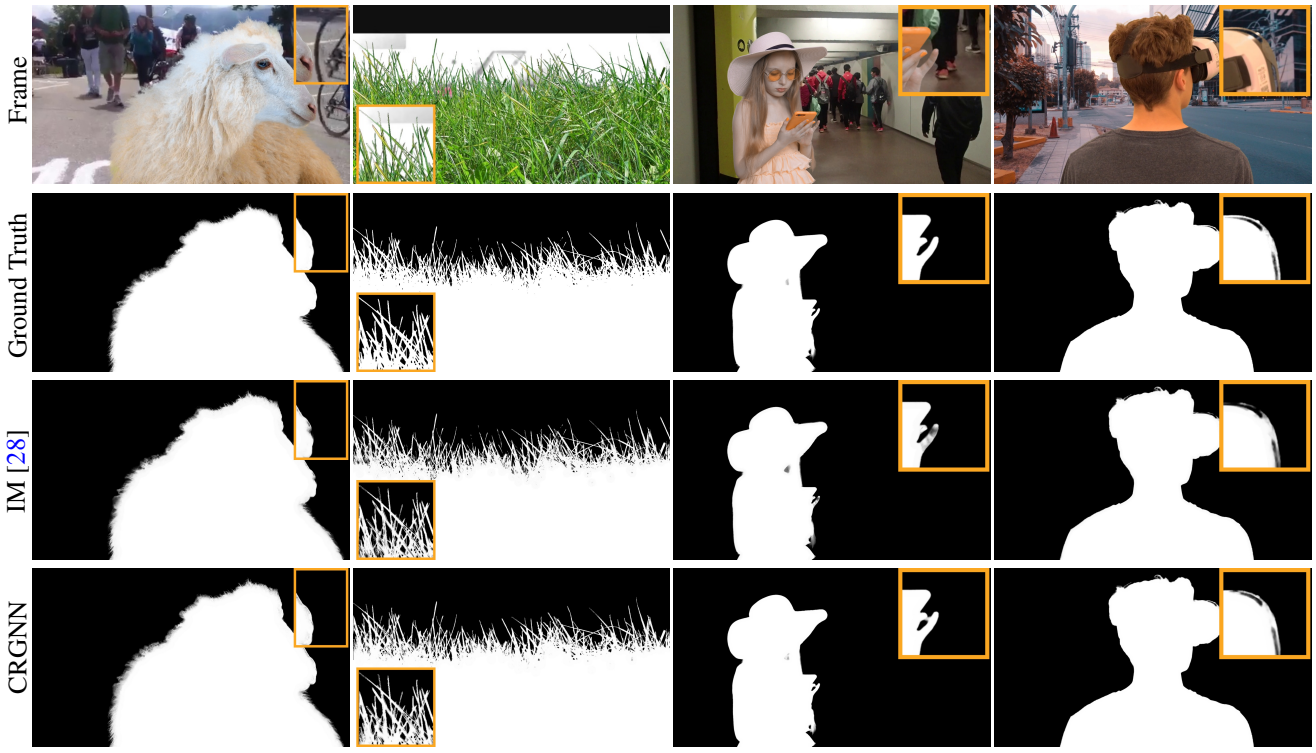

Figure 5: Visual comparison on the composited dataset.

Figure 6: Visual comparison on the real dataset.

posed composited human matting dataset and auxiliary category dataset. We include the existing image based matting methods [48, 49, 28, 23] and video based method [34]. It can be observed from Table 1a and Table 2 that the proposed method achieves better performance compared to all other methods evaluated on all five metrics. Compared to the image-based methods, the performance gain is derived from the utilization of the CRGNN, which leverages multi-frame

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.

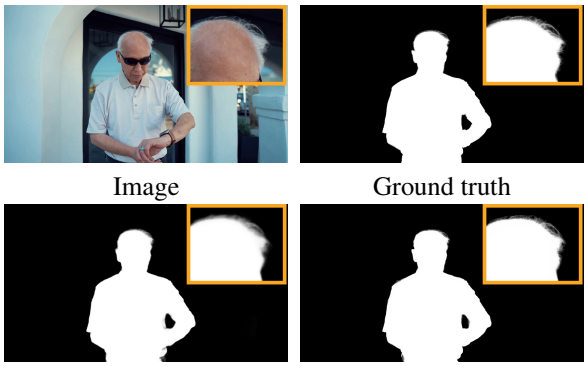

Figure 7: Visual comparison of deformable aggregation and the non-local aggregation on the real dataset.

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.