	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-