Virtual Multi-Modality Self-Supervised Foreground Matting for Human-Object Interaction
Supplementary Material

Bo Xu[1], Han Huang[1], Cheng Lu[2], Ziwen Li[1,3] and Yandong Guo[1,*]
[1]OPPO Research Institute, [2]Xmotors, [3]University of California, San Diego
yandong.guo@live.com

The supplementary material covers: technical details of training, hyper-parameters, our network architecture, the trimap generation method for trimap-based algorithms on the UFM75K data sets, more representative visualizations.

1. Technical Details

1.1. Implementation Details

The training details and all hyper-parameters are outlined in Table 1. After pre-training on the Adobe dataset [4], we first train the dual FP network on labeled data (LFM40K) for 10 epochs (Sup^1). Then we continue the training of FP and activate the training of CL module which is supervised by the deviation probability map of each FP network, for another 30 epochs (Sup^2). To present ablation study, we completed self-supervised training 3 times, each time under different setting: completed CL self-supervision (L_{cs} and L_{dc}), one single supervision L_{cs} and one single supervision L_{dc}, each for 25 epochs.

Architectures of SFPNet and DFPNet in dual foreground Prediction (FP) network are shown in Figure 1. SFPnet receives raw RGB image I, interaction heatmap H and segmentation mask S as inputs, while DFPnet replaces S with the depth map D. In each single FP network (SFPNet or DFPNet), we encode the input image (concatenated with the interaction heatmap) and the extra virtual modality (S or D) respectively. Then the image feature vector and the extra virtual modality feature vector are fused within 3 convolution layers, followed by a decoder that outputs the estimated alpha matte. We select top-k pixels with the highest estimated errors in the deviation probability map as centers to define 16×16 patches in the predicted alpha matte. Each selected patch is concatenated with its RGB region and then fed into the refinement network. We report the network architecture details in Table 2 to 4.

Parameter	Value
Optimizer	Adam
Learning rate	1.0×10^{-4}
Number of Pre epochs	20
Number of Sup^1 epochs	10
Number of Sup^2 epochs	30
Number of CL epochs	25
Number of CL(w/o L_{dc}) epochs	25
Number of CL(w/o L_{cs}) epochs	25
Number of refinement epochs	20
Batch Size	4
Loss weight λ_a	1
Loss weight λ_{com}	0.5
Loss weight λ_{cl}	1
Loss weight λ_{cs}	6
Probability threshold τ	0.5
Input image/extra modality size	512 $\times$512

Table 1: Implementation details and hyper-parameter setting.

1.2. Trimap Generation for Human-Object Interactive Scene

To evaluate and benchmark our algorithm on unlabeled data (UFM75K), we need to form trimaps for trimap-based algorithms. We follow [3] to create the pseudo trimap and apply [1] to produce human segmentation S_h. Then we label each pixel with person-class probability > 0.95 as foreground F_h, < 0.05 as background B_h, and the rest as unknown area U_h. Similarly, we manually circle areas to cover the interacted objects and correspondingly label them as unknown area U_o. Finally, the human-object interactive trimap can be generated by Eq. 1:

$$\begin{aligned} F_{h\cup o} &= F_h \\ U_{h\cup o} &= U_h \cup U_o - F_{h\cup o} \\ B_{h\cup o} &= I_{area} - F_{h\cup o} \cup U_{h\cup o} \end{aligned} \tag{1}$$

where $F_{h\cup o}$, $B_{h\cup o}$, $U_{h\cup o}$ denote foreground, background, and unknown area of the interaction trimap, I_{area} denote the whole area of raw image I.

1.3. Interaction Category

The set of human-object interaction and corresponding object categories are listed in Table 5. We follow the defini-

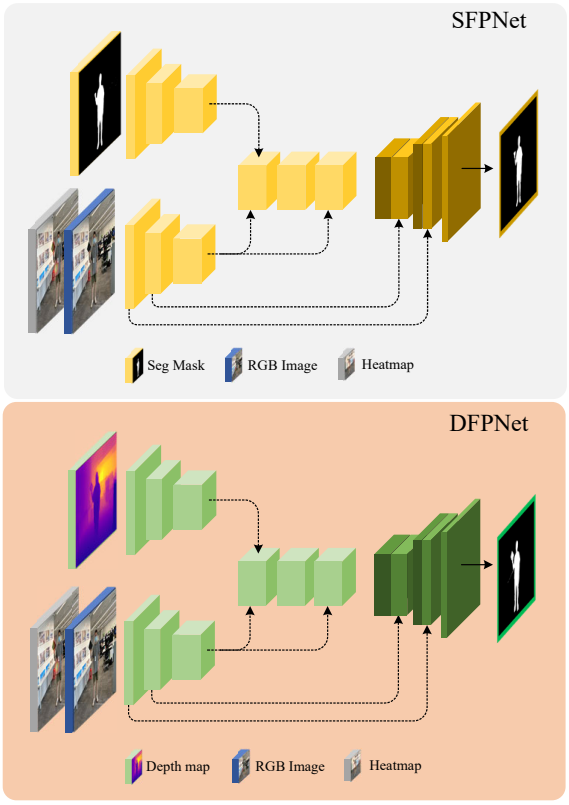

Figure 1: Architecture of the dual FP network.

Encoder (Image/vitual modality)	Output size
Conv+BN+ReLU	64 ×512 × 512
Conv+BN+ReLU	128 ×256 × 256
Conv+BN+ReLU	256 ×128 × 128
Fusion	**Output size**
Conv+BN+ReLU	256 ×128 × 128
Conv+BN+ReLU	256 ×128 × 128
Conv+BN+ReLU	256 ×128 × 128
Deconv+BN+ReLU	128 ×256 × 256
Conv+BN+ReLU	128 ×256 × 256
Deconv+BN+ReLU	64 ×512 × 512
Conv+BN+ReLU	64 ×512 × 512
Conv+Tanh	1 ×128 × 128

Table 2: Network architecture for SFPNet and DFPnet.

tion of interaction as in [2] and label 20 interaction classes to cover most human-object interaction application scenarios. In addition, human may have multiple interactions with a given object. LFM40K dataset is the first large-scale and high-quality annotated human-object interactive matting dataset with diverse scenarios, which can facilitate other researchers in this area.

Encoder (CL^m_{enc})	Output size
Conv+BN+ReLU	32 ×256 × 256
Conv+BN+ReLU	64 ×128 × 128
Conv+BN+ReLU	128 ×64 × 64
Conv+BN+ReLU	256 ×64 × 64
Decoder (CL_{dec})	**Output size**
Deconv+BN+ReLU	128 ×64 × 64
Conv+BN+ReLU	128 ×64 × 64
Deconv+BN+ReLU	64 ×128 × 128
Conv+BN+ReLU	64 ×128 × 128
Up-sample(2)	64 ×256 × 256
Conv+BN+ReLU	32 ×256 × 256
Up-sample(2)	32 ×512 × 512
Conv+BN+ReLU	32 ×512 × 512
Conv	1 ×512 × 512

Table 3: Network architecture for the complementary learning (CL) module.

Network	Output size
Conv+BN+ReLU	k ×4 × 16 × 16
Conv+BN+ReLU	k ×8 × 16 × 16
Conv+BN+ReLU	k ×8 × 16 × 16
Conv+BN+ReLU	k ×4 × 16 × 16
Conv+Tanh	k ×1 × 16 × 16

Table 4: Network architecture for the foreground refinement (RN) module.

2. More Visualization Results

Visual Results on Adobe dataset. In Figure 2 we show more visual comparisons on the Adobe test benchmark [4]. Benefiting from the virtual multi-modality (depth and segmentation) and self-supervision (complementary learning), our method can capture semantic information of the human-object interactive images with higher degree of completeness.

Visual Results on UFM75K dataset. We also display more representative visualizations of our proposed VMFM on the UFM70K dataset. As illustrated in Figure 3 to 4, visual comparisons on real images further demonstrate the effectiveness and generalization of our algorithm in multiple scenarios. For future work, our foreground matting method may extend to cover a variety of real-world applications, *e.g.* providing high-definition foreground mattes for 3D photo production, background replacement in live scene and film and TV show production, image editing, and creation by a personal computer or a mobile phone. We can also apply similar techniques to better interpret semantic information in the 3D body reconstruction under human-object interactive scenes.

Figure 2: Visual comparisons on the Adobe test benchmark.

Limitation. As shown in Figure 5, our method is limited under the scenario that human and the background obscure each other. The algorithm performance is bounded in this case because the foreground and background are sophistically blended, which remains challenging for other state-of-the-art algorithms as well. However, we are working on solution to this kind of matting use case in future research.

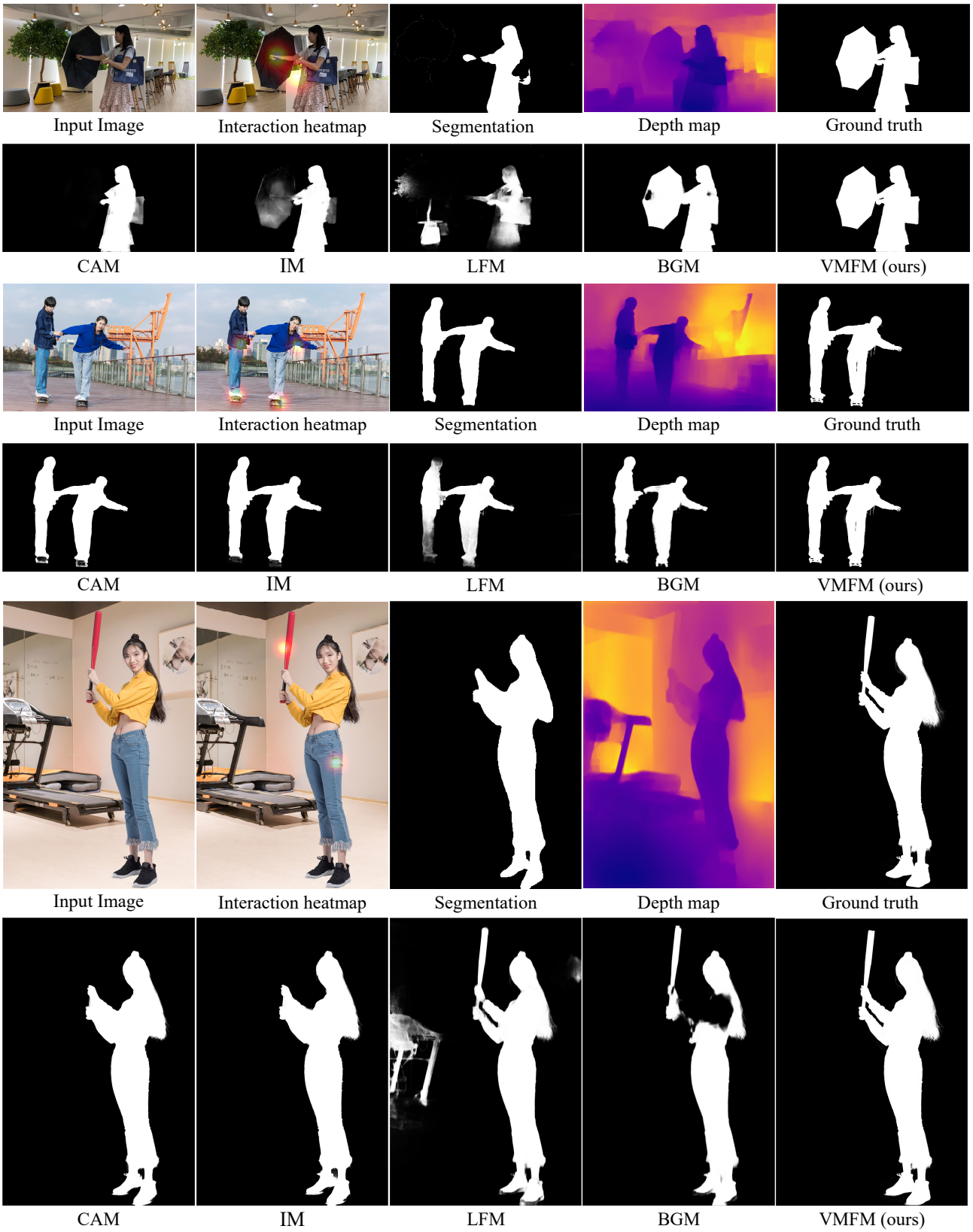

Figure 3: Visual comparisons on UFM75K test set.

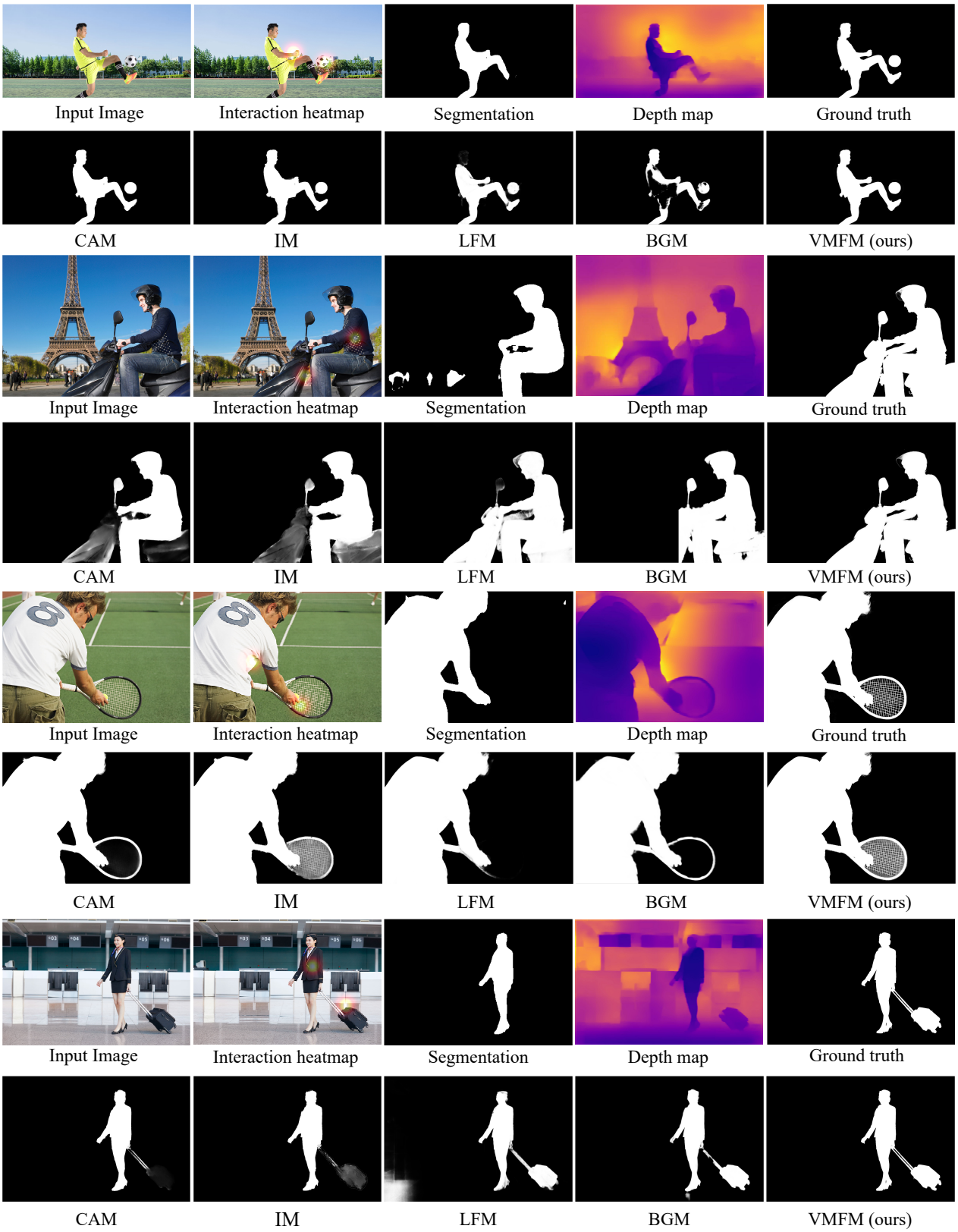

Figure 4: Visual comparisons on UFM75K test set.

Image Alpha matte

Figure 5: Failure cases.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[3] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2291–2300, 2020.

[4] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2970–2979, 2017.

Interaction class	Objects	
	LFM40K	UFM75K
carry	handbag, backpack, umbrella, box	luggage, sports ball, hairdryer, box handbag, backpack, umbrella
catch	sports ball, frisbee	sports ball, frisbee
cut	scissors, knife	fork, scissors, knife
drink	bottle	wine glass, cup, bowl
eat	hot dog, sandwich, banana	apple, orange, hot dog, cake carrot, pizza, donut
hold	pen, cup, pad, box, flower	box, ball, phone, hairdryer computer, bottle, flute, fan
hit	tennis racket, baseball bat	tennis racket, baseball bat, sports ball
jump	snowboard, skateboard, skis, surfboard	snowboard, skateboard, skis, surfboard
kick	sports ball	sports ball
lay	bench, dining table	bench, hammock, bed, couch,chair
read	book, pad	book, pad
ride	motorcycle	bicycle, motorcycle
sit	bench, chair	couch, bed, dining table, suitcase
skateboard	skateboard	skateboard
ski	ski	ski
snowboard	snowboard	snowboard
surf	surfboard	surfboard
talk on phone	cell phone	cell phone
throw	sports ball, frisbee	sports ball, frisbee
work on computer	laptop	laptop

Table 5: Interaction category and corresponding objects.