Ultrahigh Resolution Image/Video Matting with Spatio-Temporal Sparsity
Supplementary Materials

1. Datasets

To enhance research on natural image matting and human matting, previous works have contributed several valuable datasets. Table 1 summarizes these publicly available matting datasets. Existing datasets suffer from limited quantity [14, 17, 20], low resolution [1, 17, 20] or highly imbalance of human landmark (e.g., upper bodies only) [1]. Thus, in this paper we collect an ultrahigh-resolution human matting dataset, including a training set **HHM50K** and a testing set **HHM2K**, with some visual examples in Figure 1 and a distribution in Figure 2.

Collection. We collect a large number of high-resolution human images from the Internet, including Google[1] [3], Pexels[2] [4] and YFCC100M [19]. To generate alpha annotations, we apply the following pipeline. First a trimap-free matting method [9] is used to extract coarse alphas and thus automatic trimaps can be obtained from dilating these resulting coarse alphas. Afterward, a trimap-based matting method [7] is applied twice to refine alpha mattes. Finally, we use Photoshop to further refine the machine-labeled alpha mattes to meet the high benchmark quality requirement.

In addition, we include images from several public human segmentation datasets [2, 5, 6, 16, 17, 21] for enriching our training samples. However, some of their images are of low resolution. We first upgrade the quality of their images by applying denoising and state-of-the-art super-solution methods. Then we follow the aforementioned pipeline to obtain high-quality annotations for these images instead of using the original low-resolution annotations.

To guarantee the quality of machine-labeled alpha mattes, we manually inspect and correct the generated matte samples. Finally, our dataset is composed of 50,000 training images and 2,000 testing images with a distribution in Figure 2.

Experiments. Table 2 shows the quantitative results of our model, trained on the existing public matting datasets and our HHM50K. We build the synthetic dataset by selecting all human images (465) from AIM and D646 datasets as the foreground samples, then compositing them onto 100 different non-human background images from COCO [12] dataset. Our models trained on the composited dataset and HHM50K are evaluated on two datasets, HHM2K and AIM-Human.

Train	Number	Average Resolution
AIM [20]	431	1049 × 1257
D646 [14]	596	1762 × 1581
DPM [17]	2,000	800 × 600
AISegment [1]	34,425	800 × 600
HHM50K	50,000	3100 × 3260
Test	**Number**	**Average Resolution**
AIM [20]	50	1381 × 1656
D646 [14]	50	1362 × 1477
PhotoMatte85 [10]	85	3456 × 2304
PPM100 [9]	100	2875 × 2997
RWP636 [22]	636	1327 × 1038
HHM2K	2,000	4040 × 3570

Table 1. Comparisons with existing matting datasets.

Method	HHM2K			
	MAD	MSE	Grad	Conn
Ours ‡	62.42	52.09	4.38	61.62
Ours	**7.90**	**4.29**	**1.96**	**7.19**
Method	**AIM-Human**			
	MAD	MSE	Grad	Conn
Ours ‡	31.94	23.64	11.81	31.30
Ours	**16.47**	**9.18**	**9.85**	**15.38**

Table 2. Quantitative comparisons on different training and testing datasets. "Ours" and "Ours‡" indicate models trained on HHM50K and the composited training datasets respectively.

From Table 2, we observe that the model trained on the existing composited dataset shows poor performance. Composited datasets cannot provide rich natural scenarios for the model to simulate the distribution of real-world images. This leads to a performance bottleneck of human matting methods which usually require a large-scale dataset for higher generalization ability and stability. Our new dataset, which encompasses various human poses, provides strong support for training a well-generalized model.

---

[1] For internet images, refer to the Fair Use Act which "allows limited use of copyrighted material without requiring permission from the rights holders, such as for commentary, criticism, news reporting, research, teaching or scholarship."

[2] Pexels endorses free usage of all images.

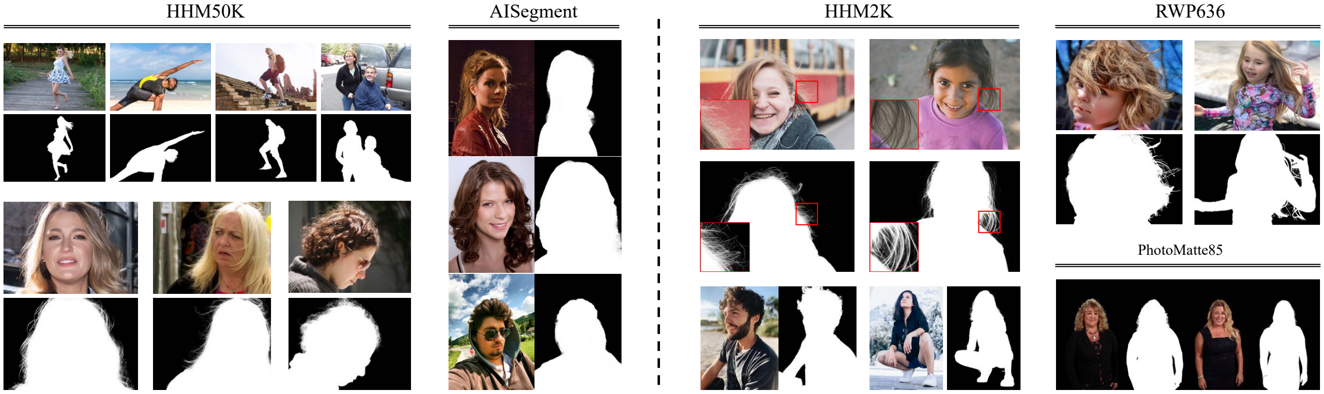

Figure 1. **Left:** visual comparisons of our HHM50K with existing human matting training dataset, AIsegment [1]. **Right:** visual comparisons of our HHM2K with existing human matting testing datasets, RWP636 [22] and PhotoMatte85 [10].

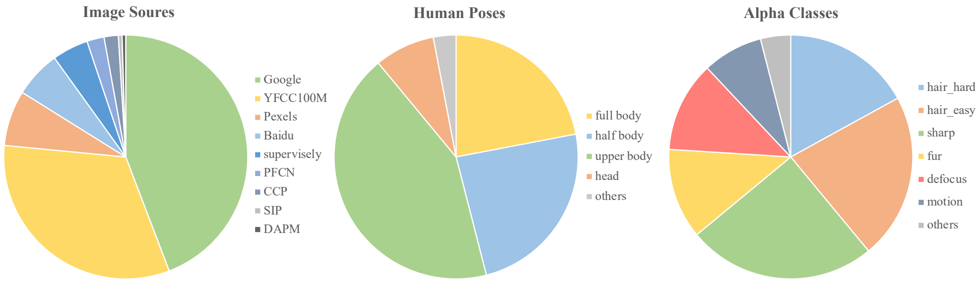

Figure 2. Distribution of image sources (left), human poses (middle), and alpha classes [18] (right) in HHM50K.

2. Low-Resolution Prior Network

In our framework, the low-resolution prior network is applied to generate spatial sparsity map. Three low-resolution prior networks are employed in the experiments, i.e., MODNet [9], RVM [11], and our self-trained lightweight human matting model, which is named LPN for simplicity. We provide its structure and optimization objectives in the following.

Structure. Given a RGB image, LPN first resizes it to 512 × 512 and extracts representative features at different levels using a backbone network (MobileNetV2 [15] is adopted). While it is straightforward to directly regress alpha mattes through a decoder with skip connections, this is not advisable as inspired by previous works [9, 13]: without prior knowledge, human matting can be divided into two sub-tasks, pixel-wise human localization (i.e., segmentation) and estimation of alpha values along boundary (i.e., matting). The goals of the two tasks are however not completely consistent with each other. Specifically, human segmentation needs more global information and thus uses large receptive field to distinguish regional pixels of human from the background. On the other hand, human matting pays more attention to local context and exploits neighboring information to disentangle foreground color from background color. Thus we focus the different levels in the decoder on different tasks, which is illustrated in Figure 3.

The decoder in the low-resolution branch uses six levels for feature reconstruction in steps. During training, we enforce the first three low-resolution levels $P_i, i \in \{1,2,3\}$ to predict multi-scale human segmentation and the last three levels $P_i, i \in \{4,5,6\}$ to predict boundary matte for auxiliary supervision. Then, the final alpha a_l is fused from human segmentation P_3 and matte P_6.

Loss Functions. For LPN, we respectively apply L2 and L1 loss for human segmentations and mattes. L1 loss and Pyramid Laplacian loss [8] are jointly utilized to optimize α_l. For simplicity, the formulation of the aforementioned losses are denoted as L_s, L_m, L_α, with i indexing level and j indexing pixel; x and $\hat{x}$ respectively denote the prediction and label. The total loss for LPN L_l consists of:

$$L_s = \sum_{i=1}^{3}\sum_{j\in I}\frac{1}{2}||P_{i,j} - \hat{P}_{i,j}||_2, i \in \{1,2,3\} \tag{1}$$

$$L_m = \sum_{i=4}^{6}\sum_{j\in U}||P_{i,j} - \hat{P}_{i,j}||_1, i \in \{4,5,6\} \tag{2}$$

$$L_\alpha = \sum_{j\in I} L_{lap}(\alpha_{l,j}, \hat{\alpha_{l,j}}) + ||\alpha_{l,j} - \hat{\alpha}_{l,j}||_1 \tag{3}$$

3. Evaluation

We conduct evaluation on multiple benchmarks, including the natural testing dataset, HHM2K, as well as composited testing datasets, AIM [20] and VM [10] in the main paper. Evaluation results on another two natural datasets,

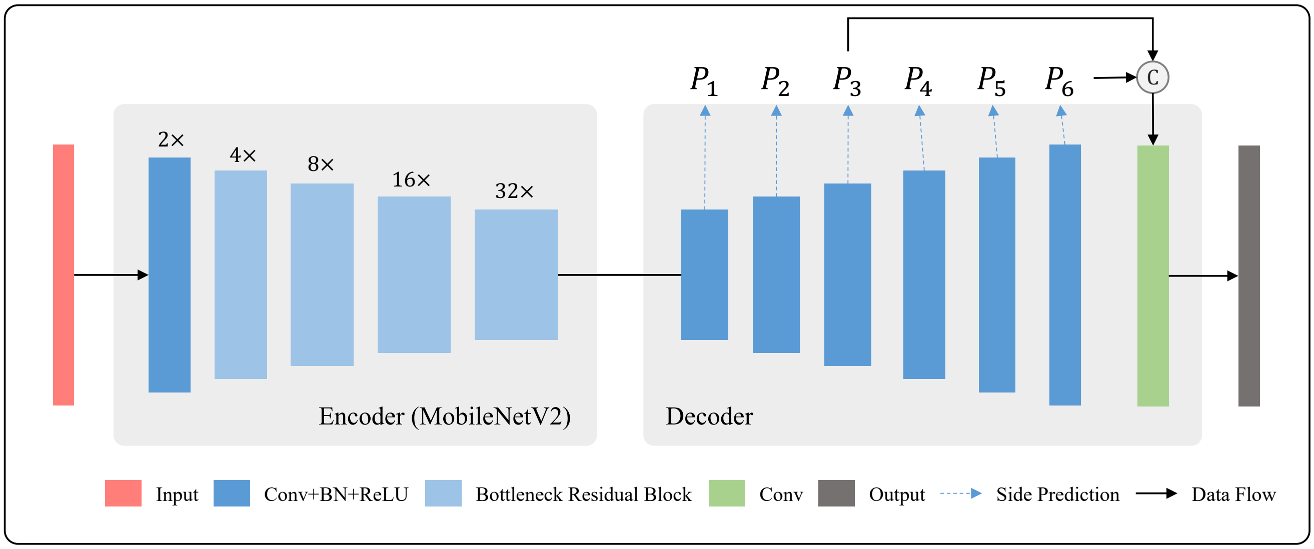

Figure 3. The structure of our self-trained LPN. Skip connections are ignored.

Method	RWP636 [22]				PPM100 [9]			
	MAD	MSE	Grad	Conn	MAD	MSE	Grad	Conn
MODNet [9]	32.88	20.48	49.39	30.56	12.44	7.40	8.11	11.00
MODNet [9] + SHM	**31.06**	**19.32**	**47.27**	**28.30**	**11.13**	**6.68**	**6.97**	**10.01**
RVM [11]	33.96	21.94	56.33	34.12	15.12	8.62	8.68	14.88
RVM [11] + SHM	**31.77**	**20.45**	**52.04**	**32.30**	**13.81**	**7.78**	**7.23**	**12.10**
LPN	29.12	18.33	48.67	29.54	11.00	6.12	7.26	9.70
LPN + SHM	**28.09**	**17.09**	**46.03**	**27.49**	**9.23**	**5.08**	**5.68**	**8.49**

Table 3. Quantitative comparisons of our SparseMat with different low-resolution networks on RWP636 [22] and PPM100 [9].

RWP636 [22] and PPM100 [9] are tabulated in Table 3.

More qualitative comparisons on RWP636, PPM100 and HHM2K are shown in Figure 4–11. Specifically, Figure 4–6 show the qualitative results on RWP636 dataset. Figure 7–8 show the qualitative results on PPM100 dataset and Figure 9–11 show the qualitative results on HHM2K dataset. We zoom in some patches of the results for better view.

Results on Videos. We compare our method with two trimap-free methods [9] and [11] on high resolution videos. Qualitative results can be found in the provided video file. Both [9] and [11] take downsampled image (512) as input, and produce low-resolution alpha matte which cannot meet the requirement of ultra high-definition displays. [11] applies deep guided filter to super-resolve the low-resolution alpha matte and achieves stable results within solid foreground and background region, but fails on the boundary region containing thin and complex hairy structures. On the contrary, our SparseMat directly processes the whole hairy region in original resolution and generates more precise alpha matte along the hairy region.

4. Limitation

The sparsity of images in daily life is usually very high, but in some special scenarios the pixels to be processed are not as sparse as expected. In this case, our method may not gain much efficiency through the sparse high-resolution module. However, most foreground objects of interest in matting in daily-life images consist of portrait, animal or other objects with small transitional regions. Sparsity inherent in these images is usually higher than 90% as mentioned in the paper. On the other hand, for highly transparent objects, they may not occupy the entire images. To provide some quantitative references, the average sparsity of the mentioned transparent objects in SIMD is higher than 60%; the images with sparsity lower than 60% of SIMD only make up 7.6%. Thus our method can practically boost performance for most cases, and rare images with super low sparsity will not benefit much from our proposed method.

References

[1] aisegment. `https://github.com/aisegmentcn/matting_human_datasets`. 1, 2

[2] baidu. `http://www.cbsr.ia.ac.cn/users/ynyu/dataset`. 1

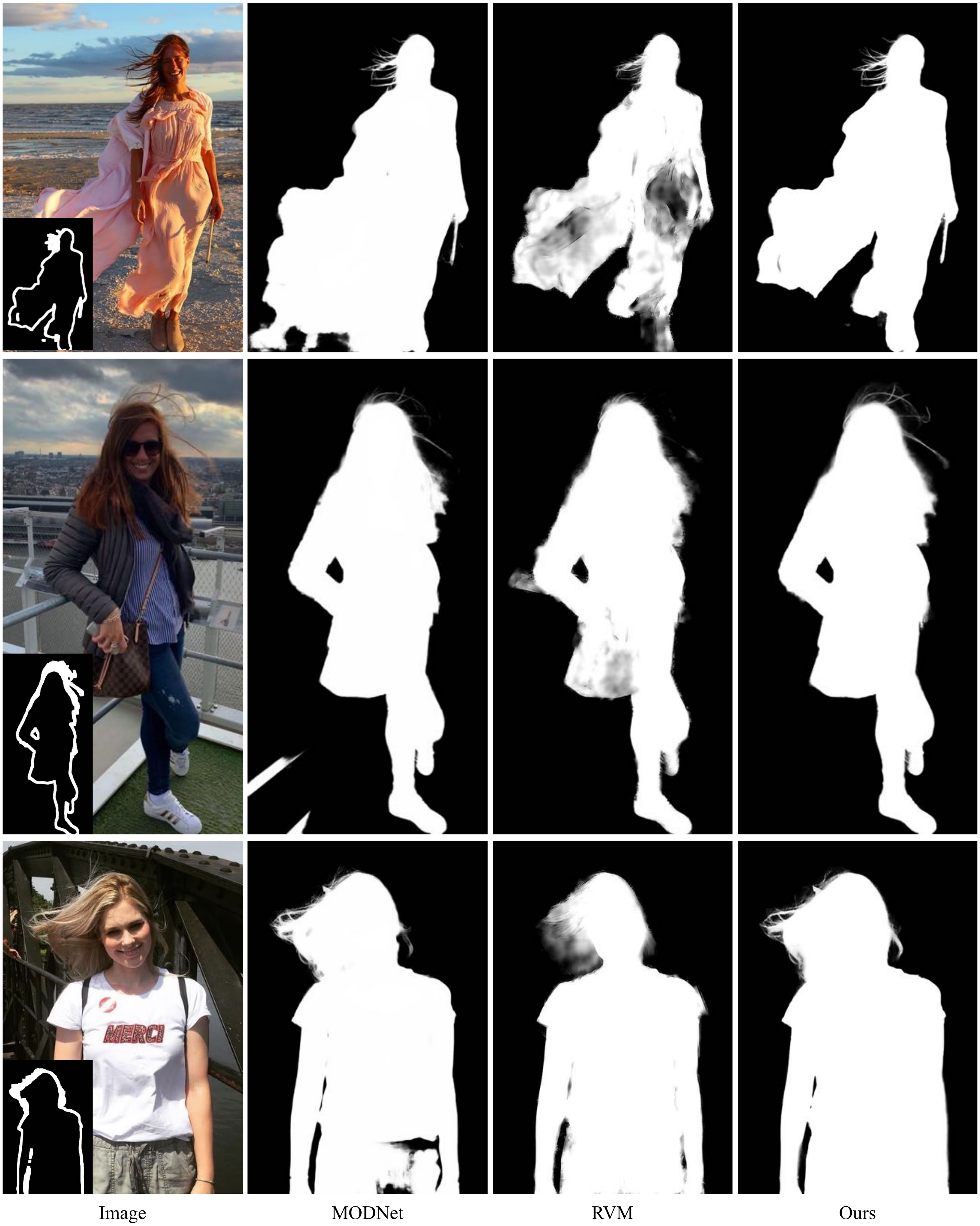

Image MODNet RVM Ours

Figure 4. Qualitative results on RWP636 [22].

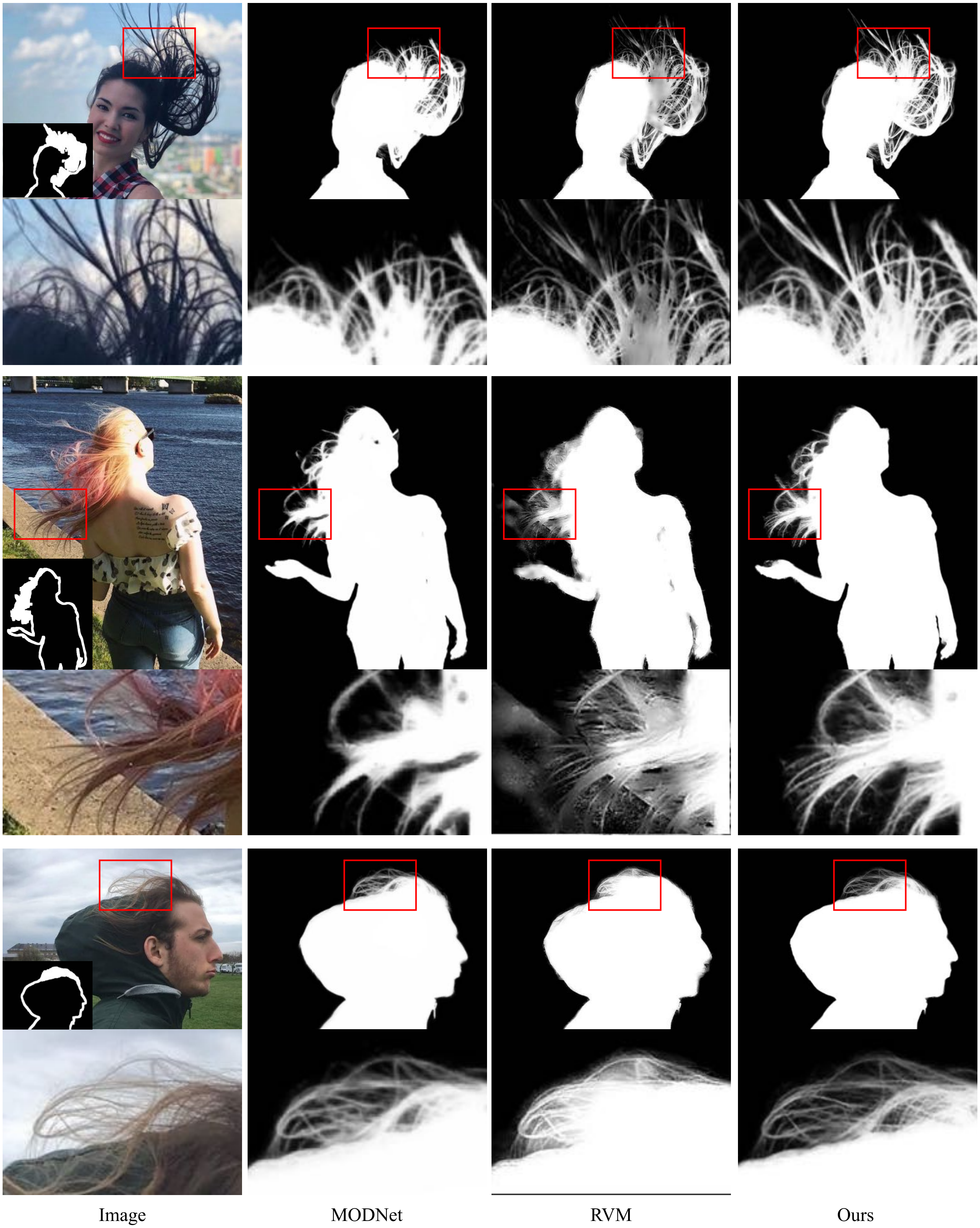

Figure 5. Qualitative results on RWP636 [22].

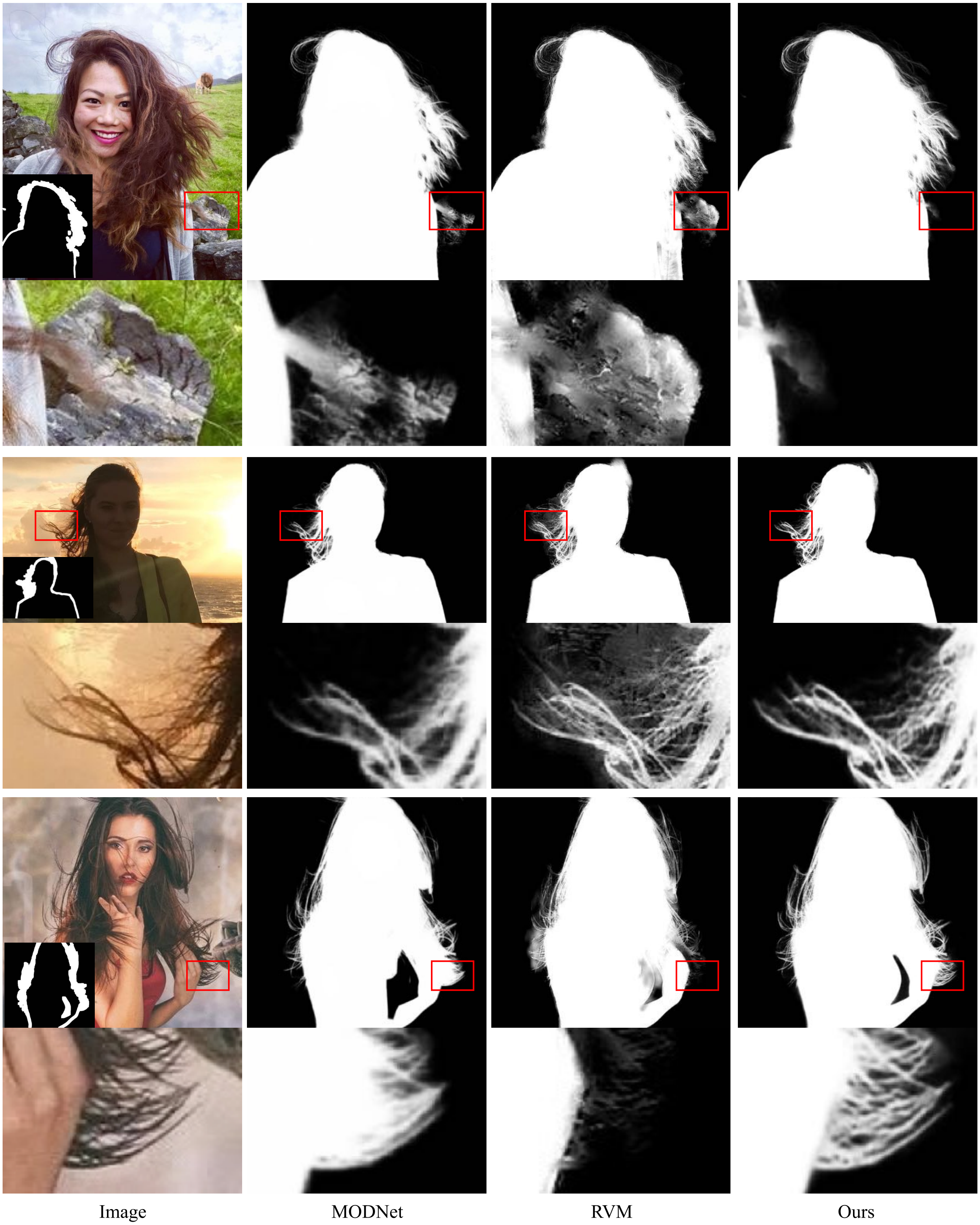

Figure 6. Qualitative results on RWP636 [22].

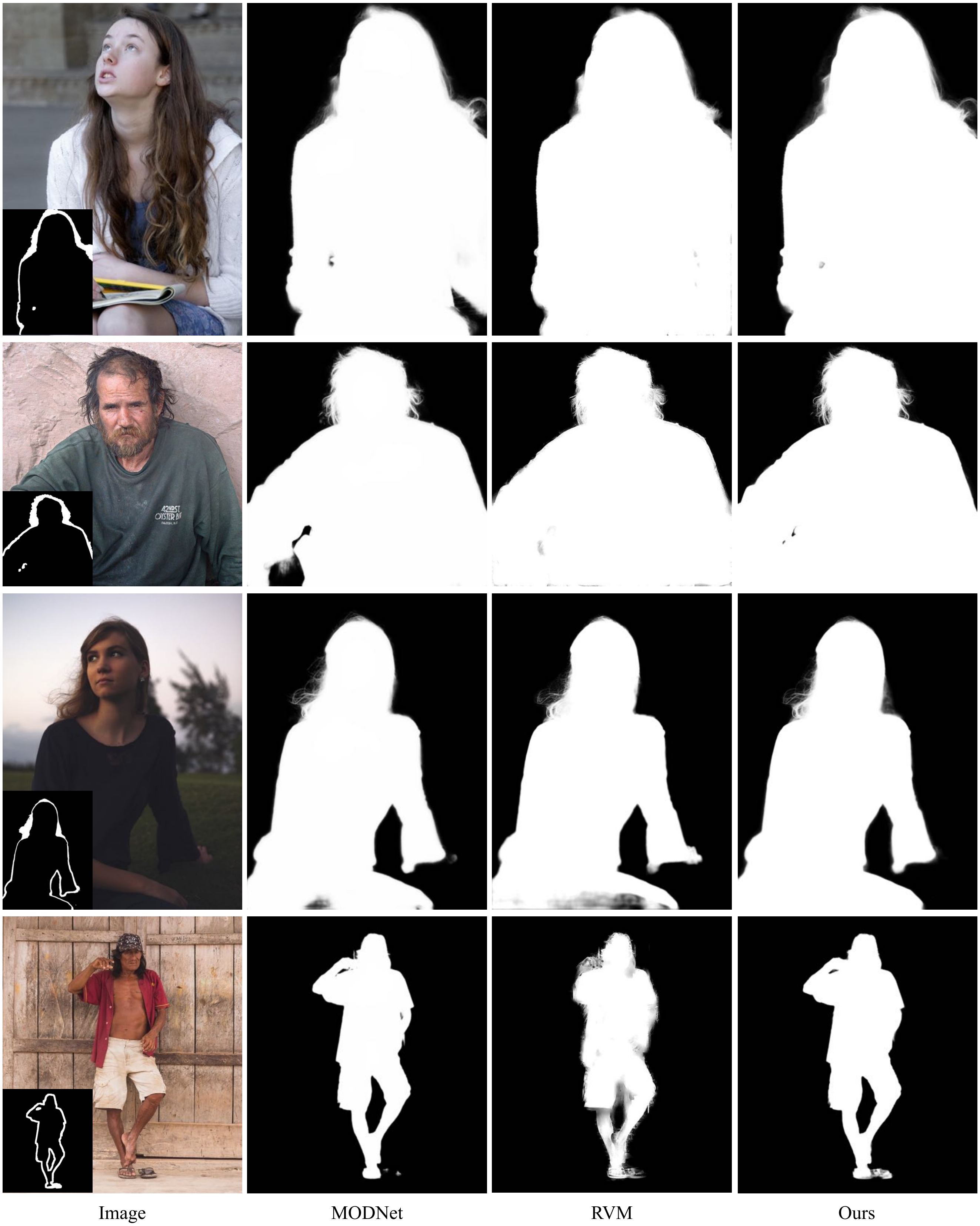

Figure 7. Qualitative results on PPM100 [9].

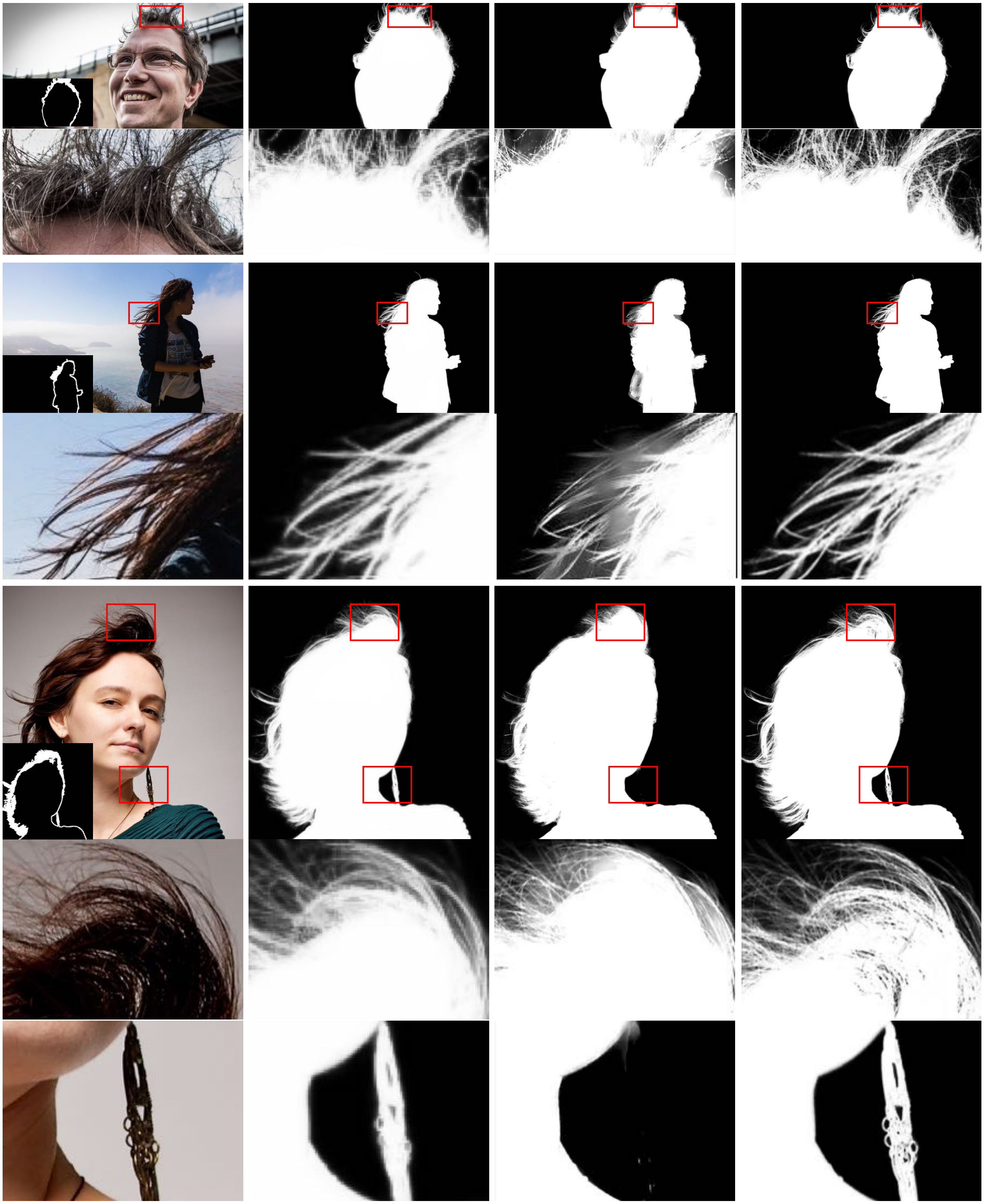

Figure 8. Qualitative results on PPM100 [9].

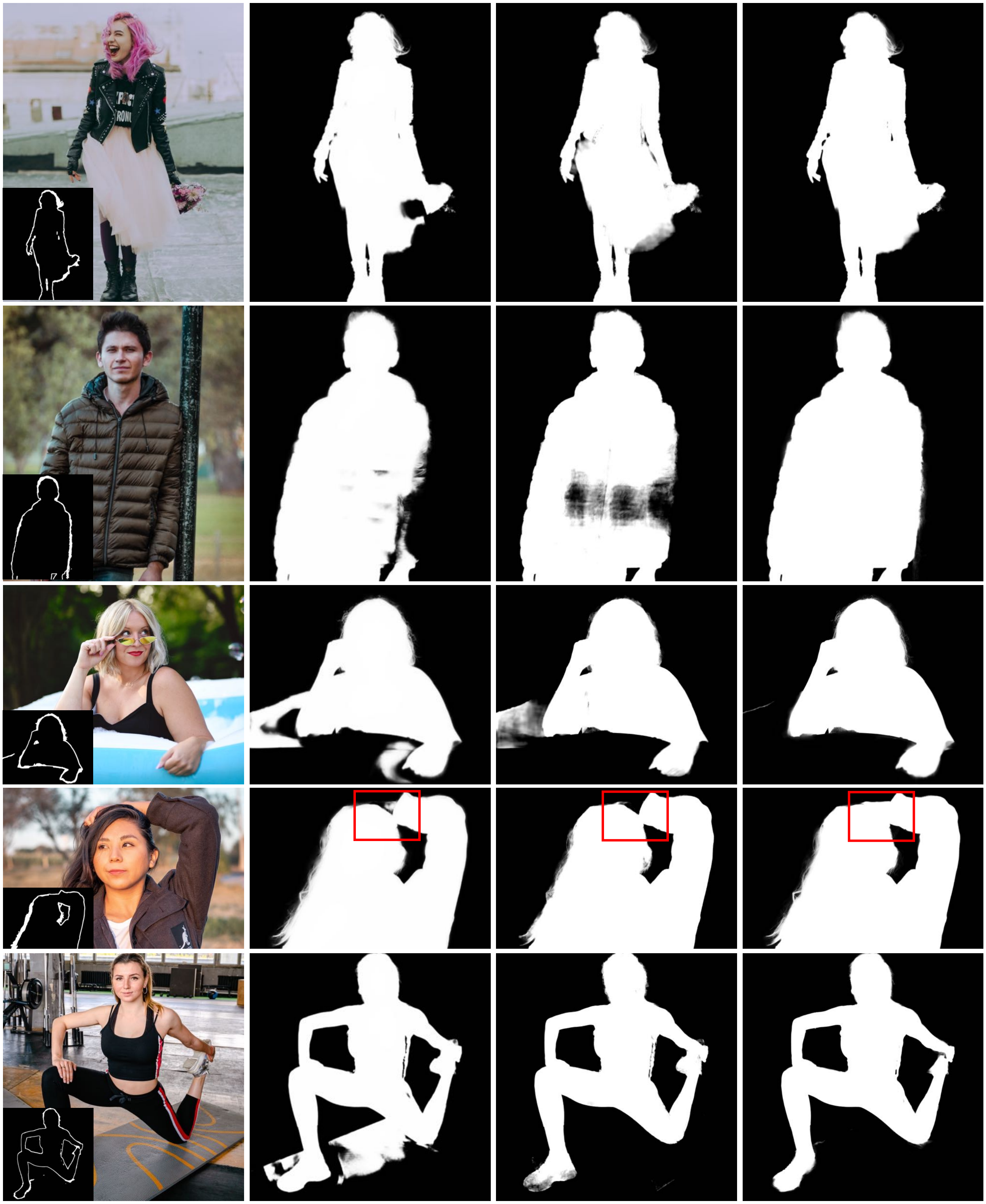

Figure 9. Qualitative results on HHM2K.

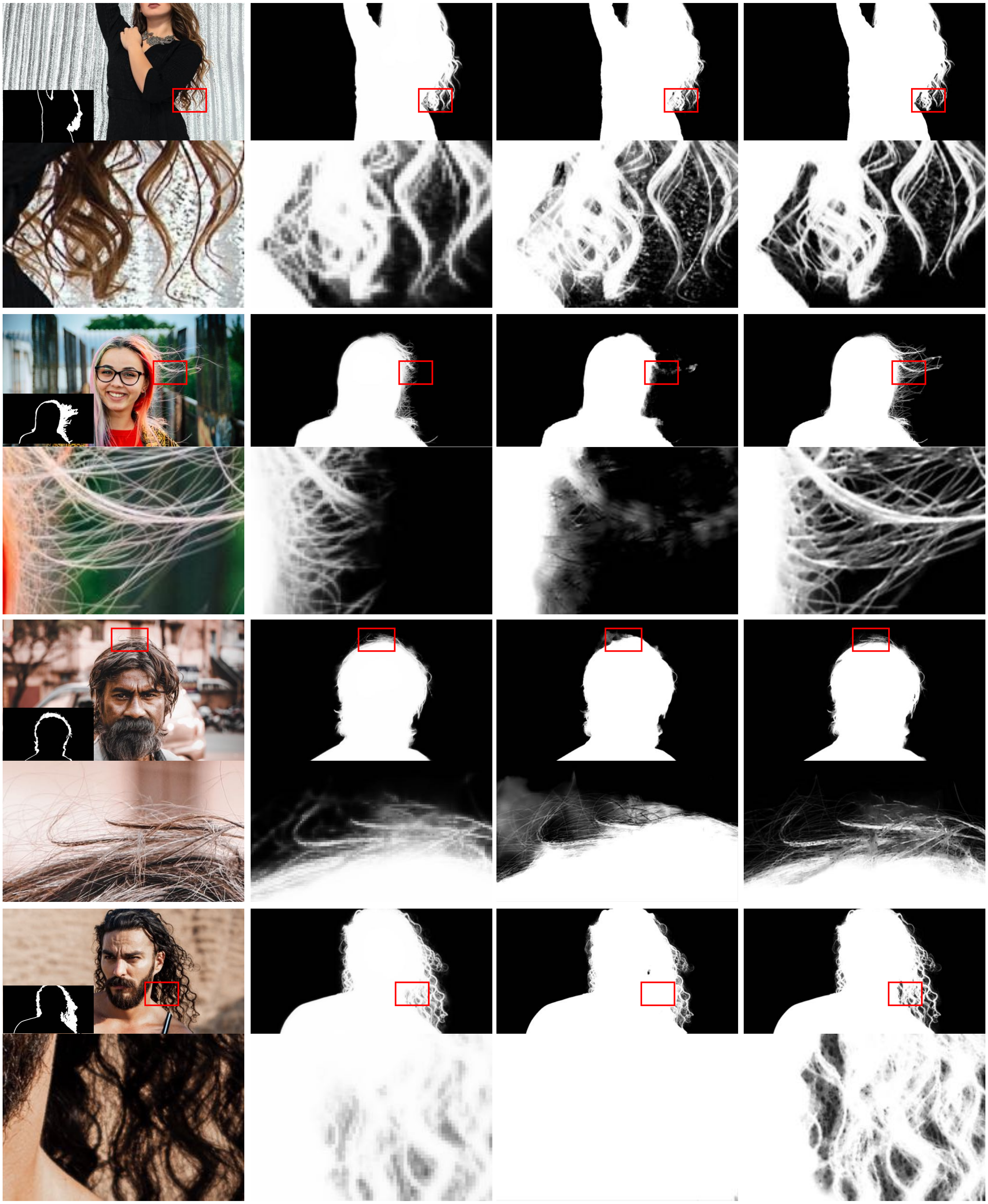

Figure 10. Qualitative results on HHM2K.

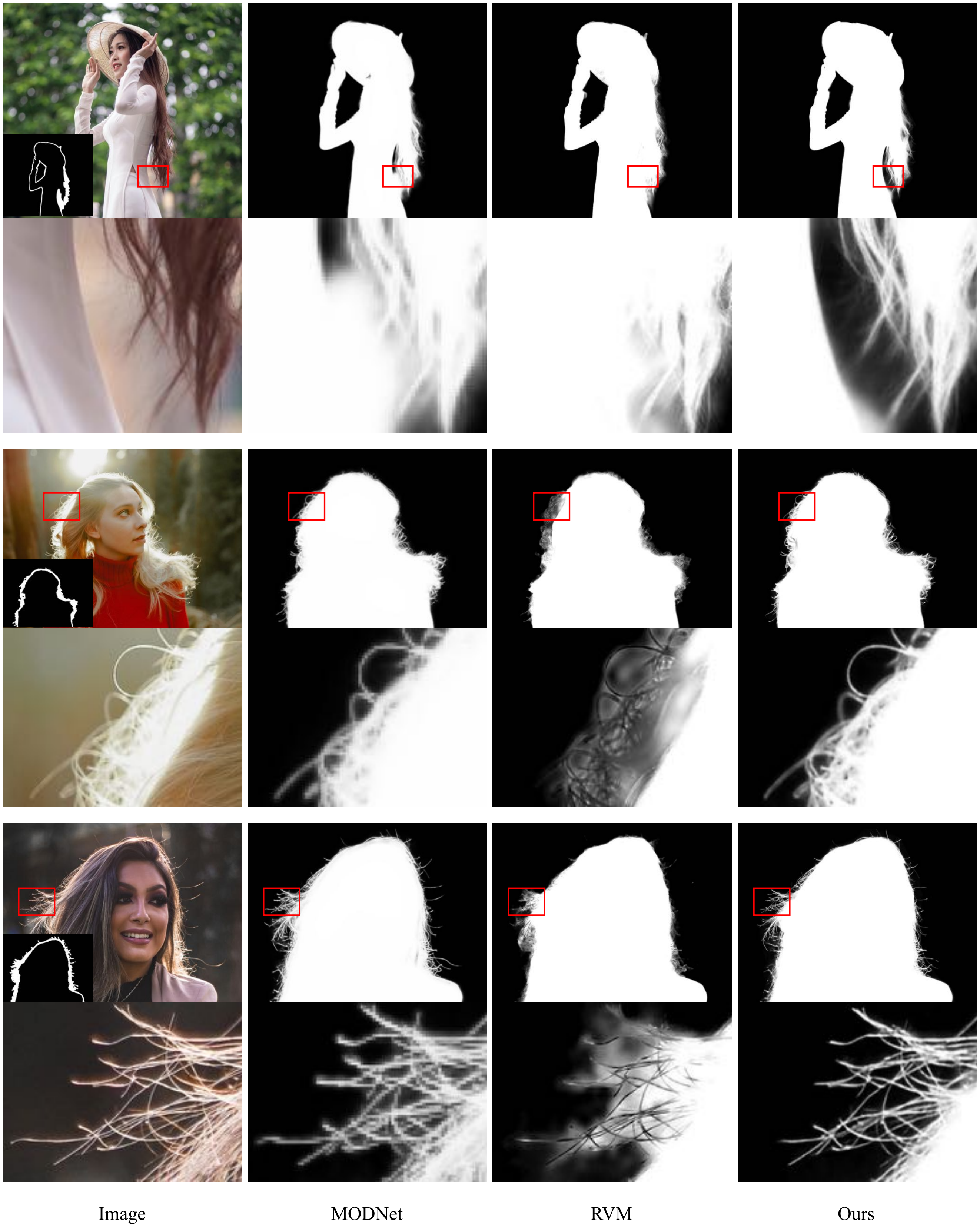

Figure 11. Qualitative results on HHM2K.

[3] google. `https://www.google.com/imghp?hl=en`. 1

[4] pexels. `https://www.pexels.com/`. 1

[5] supervisely. `https://supervise.ly`. 1

[6] Deng-Ping Fan, Zheng Lin, Jiaxing Zhao, Yun Liu, Zhao Zhang, Qibin Hou, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *CoRR*, abs/1907.06781, 2019. 1

[7] Marco Forte and François Pitié. F, b, alpha matting. *Arxiv Preprint*, 2020. 1

[8] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *International Conference on Computer Vision*, 2019. 2

[9] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson W. H. Lau. Is a green screen really necessary for real-time portrait matting? *Arxiv Preprint*, 2020. 1, 2, 3, 7, 8

[10] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[11] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. *CoRR*, abs/2108.11515, 2021. 2, 3

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, 2014. 1

[13] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-Sheng Hua. Boosting semantic human matting with coarse annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[14] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[15] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[16] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian L. Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. *Comput. Graph. Forum*, 35(2):93–102, 2016. 1

[17] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, 2016. 1

[18] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[19] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m. *Communications of the ACM*, 59(2):64–73, Jan 2016. 1

[20] Ning Xu, Brian L. Price, Scott Cohen, and Thomas S. Huang. Deep image matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[21] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[22] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan L. Yuille. Mask guided matting via progressive refinement network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4, 5, 6