Lightweight Portrait Matting via Regional Attention and Refinement Supplementary Material

1. Training Losses

The low resolution network is trained with binary cross entropy for the coarse alpha $\hat{\alpha}_c$ and focal loss [7] for the trimap $\hat{\tau}_c$ (one-hot encoded):

$$\mathcal{L}_{\hat{\alpha}_c} = -\alpha_c \cdot log(\hat{\alpha}_c) - (1-\alpha_c) \cdot log(1-\hat{\alpha}_c), \quad (1)$$

$$\mathcal{L}_{\hat{\tau}_c} = -\sum_{i=1}^{3} w_i \cdot (1-p_i)^2 \cdot \tau_{c,i} \cdot log(\hat{\tau}_{c,i}), \quad (2)$$

$$p_i = \tau_{c,i} \cdot \hat{\tau}_{c,i} + (1-\tau_{c,i}) \cdot (1-\hat{\tau}_{c,i}), \quad (3)$$

where α_c and τ_c are the downsampled ground truth alpha and trimap; i is the trimap class index; p_i measures how close the prediction $\hat{\tau}_c$ is to the ground truth τ_c and $(1-p_i)^2$ in the focal loss is designed to weigh down the well predicted samples; w_i is a weight to handle class imbalance. Empirically we set $w_i = \frac{n_i^{-0.5}}{\sum_i n_i^{-0.5}}$ for class i with n_i pixels.

In addition to computing $\mathcal{L}_{\hat{\alpha}_c}$ and $\mathcal{L}_{\hat{\tau}_c}$ at $\mathfrak{R}_8$ (the output low resolution), we also add two output heads at $\mathfrak{R}_{16}$ and $\mathfrak{R}_{32}$ and compute the losses accordingly. The two added output heads are discarded during inference.

The full resolution alpha $\hat{\alpha}$ from the refinement network is trained with alpha loss and Laplacian loss [4, 6]:

$$\mathcal{L}_{\hat{\alpha}} = \sqrt{(\alpha - \hat{\alpha})^2 + \epsilon^2}, \quad (4)$$

$$\mathcal{L}_{lap} = \sum_r \sqrt{(\mathcal{P}_r(\alpha) - \mathcal{P}_r(\hat{\alpha}))^2 + \epsilon^2}, \quad (5)$$

where $\mathcal{P}_r(\cdot)$ creates a Laplacian pyramid at resolution r. We use a 5-level pyramid with resolutions from $\mathfrak{R}_1$ to $\mathfrak{R}_{32}$.

Follwoing [4, 6, 9], We also adopt a composition loss by computing the difference between the images composed using the ground truth α and the predicted $\hat{\alpha}$:

$$\mathcal{L}_{comp} = \sqrt{(\mathcal{C}(\alpha) - \mathcal{C}(\hat{\alpha}))^2 + \epsilon^2}, \quad (6)$$

where $\mathcal{C}(\alpha) = \alpha F + (1-\alpha)B$ composes a new image with the source foreground F, a new background B and the alpha matte α. The foreground F is estimated offline using [3] given the generated pseudo ground truth trimap. The final loss is the sum of all the losses from above: $\mathcal{L} = \mathcal{L}_{\hat{\alpha}_c} + \mathcal{L}_{\hat{\tau}_c} + \mathcal{L}_{\hat{\alpha}} + \mathcal{L}_{lap} + \mathcal{L}_{comp}$.

2. Additional Baselines

Due to space limit, we show only the most representative baselines in the main paper. There are other popular baseline methods that we do not include in the main paper but are often used by prior works. We summarize the results in Tab. 1. Tab. 1 extends Tab. 2 (in the main paper) with additional baselines such as LF [10], HATT [8], SHM [1], and AIM [5]. As one as see from the table, our model outperforms all the methods in the table. The added baselines do not change our conclusion since have included the best performing baselines (e.g. DIM and P3M-Net) in the main paper. We list them here for completeness and a more comprehensive comparison.

3. Results on Real Videos

Please check out the attached demo videos in the supplementary material.

Table 1. Quantitative results on the P3M-500 tet data with additional baselines, which are listed in the upper part of the table. The lower part is copied from Table 2 of the main paper. † indicates that a trimap is used.

Method		GFLOPS	P3M-500-NP				P3M-500-P			
			SAD	SAD-T	Grad	Conn	SAD	SAD-T	Grad	Conn
LF [10]		7190.0	32.59	14.53	31.93	19.50	42.95	12.43	42.19	18.80
HATT [8]		4264.3	30.53	13.48	19.88	27.42	25.99	11.03	14.91	25.29
SHM [1]		1943.3	20.77	9.14	20.30	17.09	21.56	9.14	21.24	17.53
AIM [5]		487.4	15.50	10.16	14.82	18.03	13.20	8.84	12.58	17.75
DIM† [9]		791.6	5.32	5.32	4.70	7.70	4.89	4.89	4.48	9.68
P3M-Net [4]		364.9	11.23	7.65	10.35	12.51	8.73	6.89	8.22	13.88
MODNet [2]	512x512 input	15.7	20.20	12.48	16.83	18.41	30.08	12.22	19.73	28.61
	fullres input	103.2	63.74	13.56	25.75	62.69	95.47	13.70	37.28	94.86
BGMv2 [6]	Resnet-50	26.5	16.72	7.55	13.00	15.39	15.70	7.23	15.54	14.71
	Resnet-101	33.9	15.66	7.72	12.42	14.65	13.90	7.23	14.69	13.13
Ours		19.0	10.60	6.83	10.78	9.77	10.04	6.44	12.65	9.41

References

[1] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 1, 2

[2] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1140–1147, 2022. 2

[3] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007. 1

[4] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3501–3509, 2021. 1, 2

[5] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 1, 2

[6] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1, 2

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[8] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 1, 2

[9] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 1, 2

[10] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019. 1, 2