IEICE TRANS. INF. & SYST., VOL.E107–D, NO.6 JUNE 2024

LETTER

Real-Time Video Matting Based on RVM and Mobile ViT*

Chengyu WU[†a)], **Jiangshan QIN**[†], **Xiangyang LI**[††], **Ao ZHAN**[†], ***Nonmembers***, ***and*** **Zhengqiang WANG**[†††], ***Member***

SUMMARY Real-time matting is a challenging research in deep learning. Conventional CNN (Convolutional Neural Networks) approaches are easy to misjudge the foreground and background semantic and have blurry matting edges, which result from CNN's limited concentration on global context due to receptive field. We propose a real-time matting approach called RMViT (Real-time matting with Vision Transformer) with Transformer structure, attention and content-aware guidance to solve issues above. The semantic accuracy improves a lot due to the establishment of global context and long-range pixel information. The experiments show our approach exceeds a 30% reduction in error metrics compared with existing real-time matting approaches.

key words: *mobile ViT, video matting, deep learning, attention mechanism*

1. Introduction

Matting is a popular technology in the field of computer vision. It can effectively separate the foreground objects that people are interested in from pictures or videos. High-resolution real-time video matting have great commercial value in industries such as live streaming, but it is also challenging in deep learning research.

At present, there are three representative studies on CNN-based real-time video matting, i.e., Background Video Matting V2 (BGMv2) [1], Robust Video Matting (RVM) [2], and Matting Objective Decomposition Network (MODNet) [3]. BGMv2 has constructed a high-precision model that requires users to input a static background as a constraint to achieve real-time video matting. RVM and MODNet only need the original video frames to achieve video matting of human figures. However, BGMv2 requires user guidance and a stable environment; The ability of RVM and MODNet to capture global relationships in images is insufficient. These approaches use conventional CNN structures, which have low image accuracy in dynamic and complex backgrounds, making it easy to misjudge some background objects as foreground. Moreover, the matting result is prone to generating hollow areas, making it difficult to achieve the theoretical expectations in practical use.

Video Matting with Transformer (VMFormer) [4] adopts ViT (Vision Transformer) for matting task. VMFormer outperforms MODNet, BGMv2 and RVM in terms of accuracy. However, both its encoder and decoder use Transformer, which results in the model parameters being about twice as large as RVM model. Experiment shows that VMFormer only has 3 FPS (Frames Per Second) of inference speed processing on Nvidia Geforce RTX4060 with 1080p resolution. At present, approaches like VMFormer are difficult to be applied to fields requiring real-time performance, such as live broadcast.

Existing real-time approaches are not sensitive to long-range pixels. They frequently misjudge pixels' semantic and are not robust enough with complex background. To solve current issues of deep video matting, we design RMViT model based on the idea of RVM. Separable self-attention mechanism is introduced into the matting task to capture global information of the image. An encoder with a hybrid structure of Mobile ViT V3 [5] and inversed residual block [6] is established. The hybrid structure retains the characteristics of CNN inductive bias, and give full play to the respective advantages of CNN and Transformer. We also design an improved recurrent decoder module based on attention and content-aware guidance. The decoder is joined with CBAM [7] and CARAFE [8] operators, which has a significant improvement in upsampling process.

2. Issues and Challenges

Matting is a technology that separates foreground objects from a picture. The mathematical model is shown in Eq. (1):

$$\boldsymbol{I} = \boldsymbol{\alpha F} + (1 - \boldsymbol{\alpha})\boldsymbol{B} \tag{1}$$

Where $\boldsymbol{I}$ is the given picture, $\boldsymbol{F}$ is the foreground image, $\boldsymbol{B}$ is the background image, and $\boldsymbol{\alpha}$ is the opacity of the foreground image. It's under-constrain because there are 3 unknown factors with only 1 equation. Most approaches add constrains manually to solve this issue, e.g., BGMv2 requires a static background image $\boldsymbol{B}$ for input, and RVM

Manuscript received October 19, 2023.
Manuscript revised December 21, 2023.
Manuscript publicized January 29, 2024.

[†]The authors are with School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou, 310018, P.R.China.
[††]The author is with Southwest China Institute of Electronic Technology, Chengdu, 610036, P.R.China.
[†††]The author is with School of Communication and Information Engineering, Chongqing University of Posts and Telecommunication, Chongqing, 400065, P.R.China.

*This work was supported by the First Batch of "Pioneer" R&D Programs of Zhejiang Province in 2023 under grant 2023C01041, the Open Fund of Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications under grant BDIC-2023-B-002.

a) E-mail: jerry916@zstu.edu.cn

DOI: 10.1587/transinf.2023EDL8071

uses prior knowledge of portrait semantic. Nevertheless, added constrains result in accuracy loss or reusing cost. Existing user-guidance-free real-time matting approaches are easy to generate blur matting edges or semantic misjudgements, mainly because they are not sensitive enough to the context of images.

3. Proposed Real-Time Matting Approach

To solve current matting issues, we propose RMViT model that includes a feature extraction encoder with a hybrid structure of Mobile ViT and MobileNet V3 [6], bottleneck block, and a recurrent decoder with attention and content-aware mechanisms. The model accepts video frames or images as input, and outputs alpha matte as the result.

According to the model structure in Fig. 1, the original image is downsampled with factor k after input. In addition, to restore high-resolution details even after downsampling the original image, we adopt Fast Guided Filter (FGF) [9], which refines the low-resolution alpha image output to reconstruct the original resolution alpha matte.

3.1 Feature Extraction Encoder Based on Hybrid Structure

We propose a hybrid encoder using Inversed Residual (IR) block of MobileNet V3 and Mobile ViT V3 Block. The encoder accepts the initial downsampled image as input and outputs the processed feature map to the bottleneck block. We do not use pure Transformer structure due to following issues:

a) Mobile ViT V3 is slower than inverted residual block, which will affect the real-time performance of the model;

b) Pure ViT model lacks inductive bias characteristics and is sensitive to capacity and augmentation of the dataset.

The encoder's parameters in this structure are as shown in Table 1. Where "IR", "MViT" refer to inverted residual block and Mobile ViT V3 block respectively, L refers to the amount of Transformers in corresponding Mobile ViT block. "in", "out", "ker" and "exp" refer to input channel, output channel, convolution kernel size and expand channel size of IR blocks respectively. "se" refers to whether corresponding IR block uses short cut. "s" and "d" refer to convolution stride and dilation respectively. "act" refers to activation for IR blocks. IR blocks use hard-swish and ReLU6 as activation functions, which are represented by "HS" and "RE" respectively.

The hybrid encoder adopts separable self-attention mechanism by introducing Mobile ViT V3. Compared to most classic ViT model, the proposed structure is more lightweighted. It can significantly improve its sensitivity to global context without bringing in too much computation cost.

3.2 Decoder with Attention and Content-Aware Guidance

On the basis of RVM recurrent decoder, we propose a recurrent decoder based on attention and content-aware mechanisms. The ablation experiment shows that the accuracy of the model is significantly improved after using this decoder. The decoder block is shown in Fig. 2.

There are three decoder blocks in the model as shown in Fig. 1. The main input of each decoder block comes from the previous decoder block or bottleneck block; It also accepts skip connections (SC) feature map processed by CBAM as

Table 1 Structure of proposed encoder

Block	in	out	ker	exp	se	act	s	d	L
IR1	16	16	3	16		RE	1	1	
IR2	16	24	3	64		RE	2	1	
IR3	24	24	3	72		RE	1	1	
MViT1	24	24							2
IR3	24	40	5	72	✓	RE	2	1	
IR4	40	40	5	120	✓	RE	1	1	
IR5	40	40	5	120	✓	RE	1	1	
MViT2	40	40							4
IR6	40	80	3	240		HS	2	1	
IR7	80	80	3	200		HS	1	1	
IR8	80	80	3	184		HS	1	1	
IR9	80	80	3	184		HS	1	1	
IR10	80	112	3	480	✓	HS	1	1	
IR11	112	112	3	672	✓	HS	1	1	
IR12	112	160	5	672	✓	HS	2	2	
IR13	160	160	5	960	✓	HS	1	2	
IR14	160	160	5	960	✓	HS	1	2	
MViT3	160	160							3

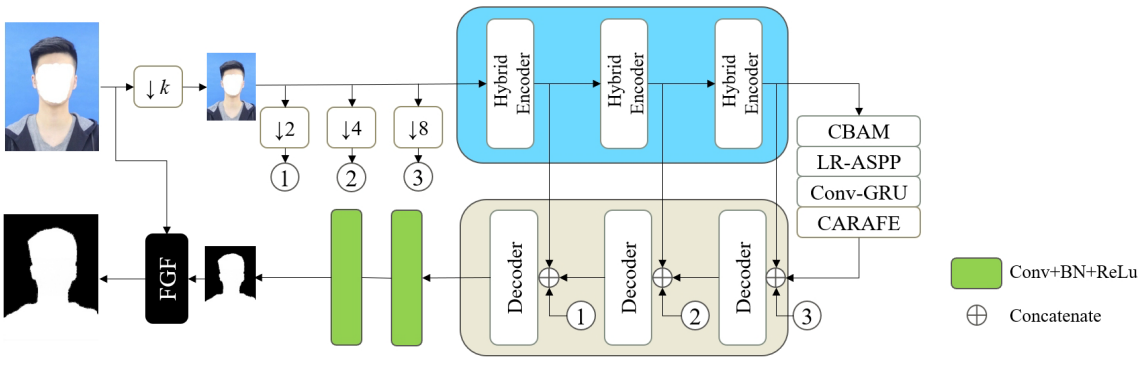


Fig. 1 Model structure

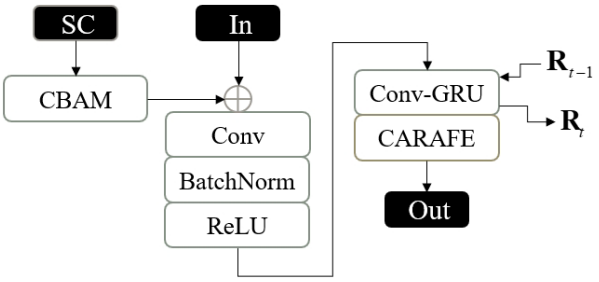


Fig. 2 Decoder with attention and content-aware mechanism

auxiliary information. In addition, to make video frames more temporally stable, ConvGRU [10] is used, which additionally accepts a recurrent feature map $\mathbf{R}_{t-1}$ from previous frame, and generates a recurrent feature map $\mathbf{R}_t$ to next frame. At the end of the module, CARAFE is used to extract key information from feature map and perform upsampling with content-aware guidance to improve accuracy.

3.3 Refine Module Based on Fast Guided Filter

Considering the demand of real-time performance on high-resolution video frames, we adopt Fast Guided Filter (FGF) as a refine module. As shown in Fig. 1, the original input frame is first downsampled by factor k and then processed by the encoder-decoder base network. The base network outputs a low-resolution alpha matte and send it to FGF together with original input frame. FGF module then generates refine alpha matte on original resolution. The value of k can be adjusted to accommodate different input resolutions. Note that the encoder-decoder base network can process frames standalone in case of low-resolution or non-real-time operation.

3.4 Training

To achieve better performance, we used specific training methods and multiple datasets. We apply AMP (Automatically Mixed Precision) and Adam optimizer to speed up training and accelerate convergence.

The datasets we used in training process are as follows:

1. Video foreground dataset: Video Matting 240K [1];
2. Video background dataset: Deep Video Matting (DVM) [11];
3. Portrait segmentation datasets: COCO [12], Supervisely Person Dataset [13], YoutubeVIS 2021 [14];
4. High-resolution foreground image datasets: PPM-100 [3], P3M-10K [15], AIM-500 [16], Adobe Matting Dataset [17], Distinctions 646 [18];
5. Image background dataset: Indoor CVPR 09 [19].

We divide the matting training process into three parts with total 35 epochs aiming at different circumstances, as follows:

- **Part 1:** In part 1 we train the model without FGF on Video Matting 240K and DVM for 20 epochs, and only use low-resolution (512×512) video sequences with a total length of 20 frames.
- **Part 2:** In part 2 we train the model still without FGF on hybrid resolution video sequences for 3 epochs. The hybrid video sequences are from the same datasets as part 1, containing low-resolution frames (512×512) with a length of 10 and high-resolution frames (2048×2048) with a length of 3.
- **Part 3:** In part 3 we train the model for high-resolution image matting task for 12 epochs. The model is trained on high-resolution foreground image datasets (P3M-10K, Distinctions 646, etc.) and image background dataset (Indoor CVPR 09). We add FGF, setting the initial downsample factor k to 0.25 during the former 10 epochs and 1.0 during the latter 2 epochs.

Portrait segmentation training is interspersed throughout the entire training process. we insert one segmentation training step after every 2 matting training steps to ensure the model's sensitivity to human figures.

To ensure segmentation and matting performance, we apply different losses and weight them as a total loss L_t. Considering details at the edge of foreground images, we apply pyramid Laplacian loss L_{lap} [20] as well as L1 loss L_1. Moreover we apply a temporal coherence loss L_c to reduce flickers of generated frames. The losses are as follows:

$$L_1 = ||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}||_1 \tag{2}$$

$$L_{lap} = \sum_{s=1}^{5} 2^{s-1} ||\mathcal{L}_s(\hat{\boldsymbol{\alpha}}) - \mathcal{L}_s(\boldsymbol{\alpha})||_1 \tag{3}$$

$$L_c = ||\frac{d\hat{\boldsymbol{\alpha}}}{dt} - \frac{d\boldsymbol{\alpha}}{dt}||_2 \tag{4}$$

$$L_t = L_1 + \frac{1}{5} L_{lap} + 5L_c \tag{5}$$

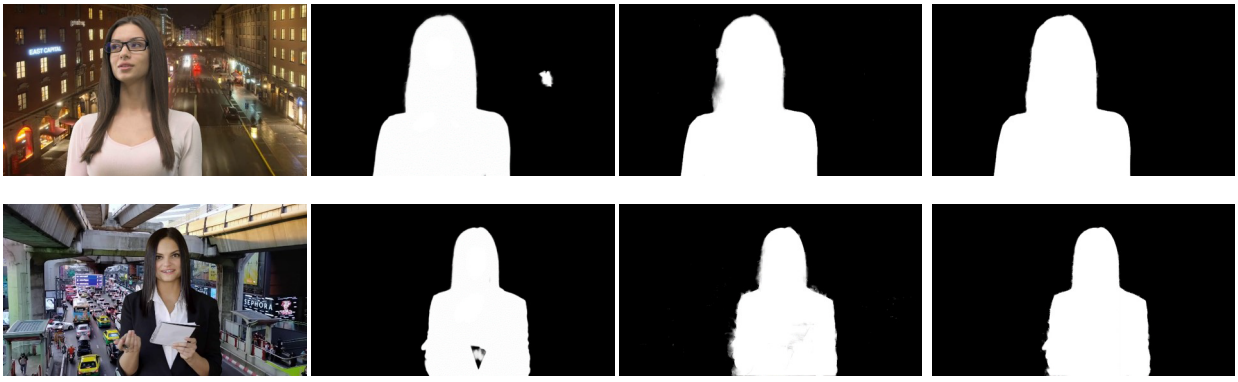
Where, $\hat{\boldsymbol{\alpha}}$ represents predicted alpha matte, $\boldsymbol{\alpha}$ represents ground truth, and $\mathcal{L}_s(\boldsymbol{\alpha})$ represents the computed result of the s-th layer of Laplacian pyramid based on $\boldsymbol{\alpha}$.

4. Experiments

The training and evaluating process uses Nvidia Geforce RTX 4060 and RTX 3060 for multi-card training, with AMD Ryzen 9 5950X CPU and mixed precision throughout the entire process. During the evaluating process, RTX 2070 laptop and Intel Core i7-9750H are also used for speed test.

4.1 Comparison Experiments

We compare the accuracy of the existing model and the proposed model on four indicators, i.e., MAD, MSE, Connectivity Error (Conn), and Gradient Error (Grad). The evaluation dataset used in this experiment is Video Matte 240K HD, with a total of 50 video clips. For the convenience of data display, we magnify MAD and MSE results by 1000 times, while the Conn and Grad values are reduced to one thousandth of the original values. The experimental results of numerical evaluation are shown in Table 2. It can be seen that RMViT has better results in image retrieval



(a) Input (b) MODNet (c) RVM (d) Ours

Fig. 3 Visualized comparison of matting result

Table 2 Comparison of matting evaluation scores

Approaches	MAD	MSE	Grad	Conn
BGMv2	23.45	18.74	10.66	41.96
MODNet	9.73	4.60	**9.43**	11.96
RVM	10.80	3.72	18.26	15.63
RMViT(Ours)	**7.12**	**2.32**	11.79	**8.05**

Table 3 Comparison of model size

Model	Size(MB)	Params(millions)
BGMv2	19.3	5.04
MODNet	25.0	8.78
RVM	14.5	3.77
RMViT(Ours)	14.9	3.82

Table 4 Comparison of speed

Model	Speed(FPS)	GPU Usage(%)
BGMv2	33	86
MODNet	28	75
RVM	35	36
RMViT(Ours)	30	37

accuracy compared to CNN-based models. BGMv2 uses a static background as the constraint, it is unstable in dynamic backgrounds and has poor anti-interference ability. Similarly, MODNet outperforms RMViT in the Grad metric, but all other error metrics are higher than RMViT. RVM has good stability for background changes, but its lack of global information perception and the lack of attention mechanism in the decoder make it comprehensively lag behind RMViT in evaluation experiments.

Figure 3 shows the visualization results of video matting, where two video frames under dynamic backgrounds are selected from Video Matte 240K HD [1]. From Fig. 3, it can be observed that RMViT can relatively clearly recognize the edges of the foreground in dynamic backgrounds. This means that it can more accurately extract foreground targets and has good semantic recognition ability when dealing with dynamic scenes. MODNet performs well in image edge detection, but there are semantic recognition errors that mistakenly recognizes background pixels as foreground or loses foreground pixels. RVM results in blurry edges and inability to clearly identify foreground and background pixels, and lacks sufficient stability and accuracy when dealing with dynamic backgrounds.

According to the results above, RMViT shows its accuracy in complex dynamic background matting. RMViT has fewer semantic errors and sharper edges in matting results. It also can be seen that RMViT has reduced Grad by 35% and Conn by 48% respectively, compared to RVM, meaning that our approach is more robust under such circumstance.

4.2 Size and Speed

We compared existing approaches to ours on size and speed evaluation. Table 3 shows that our model is lighter and has fewer parameters compared to BGM and MODNet. Compared with RVM, our model has only increased the number of parameters by 1.3% and size by 2.8%, but it has exceeded a 30% reduction in error metrics, indicating the effectiveness of our approach.

To verify the real-time performance in actual use, we test different approaches in a practical environment which includes video capturing, video codecs, frame preprocessing, data parallel and rendering. The models are tested on a laptop equipped with Intel Core i7-9750H CPU and RTX 2070 laptop GPU. The TDP (Thermal Design Power) of CPU and GPU are set to 80W and 115W respectively, so as to demonstrate whether the approaches can maintain their real-time performance on mid-end or low-end devices. All models are tested on a 1080p video sequence.

As shown in Table 4, FPS and GPU core usage are measured. In the experiment we set downsample factor $k = 0.18$. The result shows that RMViT has analogous speed compared to existing real-time approaches. BGMv2 and MODNet exceed 75% of GPU usage while RMViT and RVM are within 40%. RMViT achieves 30 FPS on 1080p resolution as well as relatively low GPU usage cost in practical environment, showing that our method is considered real-time for conven-

Table 5 Ablation experiment

Model	MAD	MSE	Grad	Conn
Ablation model 1	8.01	2.37	14.24	9.62
Ablation model 2	10.87	3.58	18.32	15.44
FGF-free model	6.71	2.05	7.22	6.98
Original model	6.14	1.44	10.76	5.97

tional video applications.

4.3 Ablation Experiment

To investigate the impact of Mobile ViT, CBAM, CARFAFE and FGF modules on the overall accuracy of the model, ablation experiment is conducted, as follows:

a) Ablation model 1: Retain Mobile ViT V3 module in the encoder, and eliminate the optimization of attention and content-aware mechanisms in the decoder;

b) Ablation model 2: Simultaneously eliminate Mobile ViT V3 modules in the encoder and the attention and content-aware mechanisms in the decoder;

c) FGF-free model: Only eliminate FGF, and directly process high-resolution videos instead of downsampling by factor k initially;

d) Original model: Do not eliminate any structure and includes all modules.

In this experiment, former 10 video clips from Video Matte 240K HD test set are selected. Default downsample factor k is set to 0.25.

The experiment results are shown in Table 5. It can be seen that self-attention mechanism brought about by the ViT structure in the encoder, as well as the attention and content-aware mechanisms in the decoder, have a significant improvement effect on model accuracy. Result of FGF-free model shows that there is a reduction on Grad, while other three error metrics slightly increase. FGF and initial downsample process do not have significant negative impact on accuracy. It is considered feasible to speed up inference with FGF refine module.

5. Conclusion

We introduce a matting approach named RMViT due to the low accuracy and semantic misjudgement. We proposed a hybrid matting model under circumstance of real-time matting task. The proposed approach adds separable self-attention mechanism in hybrid encoder, and designs decoder modules joined with attention and content-aware guidance, which make the model establish enough global context information so that the approach makes fewer semantic mistakes and sharper edges. The experiments show that our approach is better than MODNet, BGMv2 and RVM while ensuring real-time performance.

References

[1] S. Lin, A. Ryabtsev, S. Sengupta, B. Curless, S. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.8758–8767, June 2021.

[2] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp.3132–3141, Jan. 2022.

[3] Z. Ke, J. Sun, K. Li, Q. Yan, and R.W.H. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," Proc. 36th AAAI Conference on Artificial Intelligence, vol.36, no.1, pp.1140–1147, Feb. 2022.

[4] J. Li, V. Goel, M. Ohanyan, S. Navasardyan, Y. Wei, and H. Shi, "Vmformer: End-to-end video matting with transformer," arXiv preprint arXiv:2208.12801, 2022.

[5] S.N. Wadekar and A. Chaurasia, "Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," ArXiv, vol.abs/2209.15159, 2022.

[6] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp.1314–1324, Oct. 2019.

[7] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "Cbam: Convolutional block attention module," Proc. European Conference on Computer Vision (ECCV), Munich, Germany, pp.3–19, Sept. 2018.

[8] J. Wang, K. Chen, R. Xu, Z. Liu, C.C. Loy, and D. Lin, "Carafe: Content-aware reassembly of features," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp.3007–3016, Oct. 2019.

[9] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.1838–1847, June 2018.

[10] N. Ballas, L. Yao, C. Pal, and A.C. Courville, "Delving deeper into convolutional networks for learning video representations," 4th International Conference on Learning Representations (ICLR), Puerto Rico, USA, Feb. 2016.

[11] Y. Sun, G. Wang, Q. Gu, C.-K. Tang, and Y.-W. Tai, "Deep video matting via spatio-temporal alignment and aggregation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.6971–6980, June 2021.

[12] Microsoft, "Coco," https://cocodataset.org, 2020.

[13] DeepSystemsAI, "Supervisely person dataset," https://github.com/supervisely/supervisely, 2014.

[14] Youtube, "Video instance segmentation," https://youtube-vos.org/dataset/vis/, 2021.

[15] J. Li, S. Ma, J. Zhang, and D. Tao, "Privacy-preserving portrait matting," Proc. 29th ACM International Conference on Multimedia, MM '21, New York, NY, USA, p.3501–3509, Association for Computing Machinery, 2021.

[16] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," Proc. Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, ed. Z.H. Zhou, pp.800–806, International Joint Conferences on Artificial Intelligence Organization, Oct. 2021. Main Track.

[17] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.311–320, 2017.

[18] Y. Qiao, Y. Liu, X. Yang, D. Zhou, M. Xu, Q. Zhang, and X. Wei, "Attention-guided hierarchical structure aggregation for image matting," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, pp.13673–13682, June 2020.

[19] A. Quattoni and A. Torralba, "Recognizing indoor scenes," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp.413–420, June 2009.

[20] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1701–1710, June 2018.