Sampling Propagation Attention with Trimap Generation Network for Natural Image Matting

Yuhongze Zhou, Liguang Zhou, *Student Member, IEEE,* Tin Lun Lam, *Senior Member, IEEE,* and Yangsheng Xu, *Fellow, IEEE*

***Abstract*—Natural image matting aims to precisely separate foreground objects from backgrounds using alpha mattes. Fully automatic natural image matting without external annotations is challenging. Well-performed matting methods usually require accurate labor-intensive handcrafted trimap as an extra input while the performance of automatic trimap generation method, e.g., erosion/dilation manipulation on foreground segmentation, fluctuates with segmentation quality. Therefore, we argue that how to produce a high-quality trimap using coarse segmentation is a major issue in automatic matting. In this paper, we present a two-stage trimap-free natural image matting pipeline that does not need trimap and background as input. Specifically, guided by a coarse segmentation, Trimap Generation Network (TGN) estimates a trimap where the coarse segmentation can be produced by segmentation/salient object detection/matting approaches, which enables more flexibility for matting to adapt into different scenarios. Then, with an estimated trimap as guidance, our Sampling Propagation Attention Matting Network (SPAMattNet) estimates an alpha matte. Different from previous propagation-based matting networks, inspired by traditional sampling/propagation matting approaches, we propose Sampling Propagation Attention (SPA) for matting network to incorporate sampling and propagation procedures in deep learning based manner for network explainability and performance improvement. It explicitly investigates local spatial and global semantic relationships to reconstruct alpha features. To better harvest sampling/propagation and local/global information, a Cross-Fusion Contextual Module (CFC) is introduced to aggregate features from different sources. Extensive experiments are conducted to show that our matting approach is competitive compared to other state-of-the-art methods in both trimap-free and trimap-needed aspects on several challenging matting benchmarks.**

***Index Terms*—Image Matting, Trimap Generation Network, Sampling Propagation Attention**

I. Introduction

Image matting is a popular image editing task that attempts to extract an accurate foreground mask, i.e., alpha matte, from the background. Matting problem can be formulated in a general mathematical manner. An image I can be defined as a linear combination of alpha matte α, foreground F, and background B image as follows:

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

where RGB I is known, but F, B, and α are unknown. That is to say, matting attempts to solve seven unknown variables with only three variables provided. Therefore, most compelling matting methods usually require a handcrafted trimap for region constrain to reduce complexity and assist matte estimation, which makes fully-automatic natural image matting such an appealing task to explore.

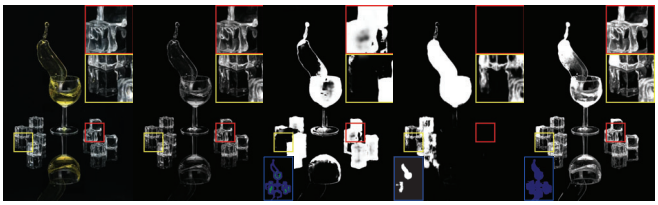

Fig. 1. A visual comparison of our trimap-free matting pipeline and other matting methods. The guidance input is located at the bottom-left of each image, where the guidance of BSHM is the output of its Quality Unification Network. See more examples in Fig. 15.

Let us first recap recent learning-based image matting approaches and their pros and cons. Learning-based image matting can be divided into three primary categories, including background-required [3]–[5], trimap-needed [6]–[19], and only-image [2], [20]–[23] input, i.e., trimap-free.

Background Matting [3], [4] requires a background image and soft segmentation as input but cannot resist interference of shadows or complex light condition. For trimap-needed matting, its accuracy is paramount, which gives the credit to auxiliary trimap. The trimap provides deterministic foreground, unknown, and background regions of image and narrows down matte estimation to unknown region. However, the manual creation of trimap is painstaking, which diminishes its application potential. Hence, trimap quality is one significant factor that can affect matting performance. One possible workaround is a general automatic trimap generation method, that is, target foreground can be roughly extracted by segmentation/salient object detection/matting approaches and then processed by image dilation/erosion. Regarding this way, segmentation quality has a dominant influence on the generated trimap, similar to what trimap is to matte.

From the above-mentioned problems, it is obvious that trimap-needed matting (resp. the trimap generation method)

Manuscript received Oct. 10th, 2022. This work was partly supported by the National Natural Science Foundation of China under Grant 62073274, the Shenzhen Science and Technology Program under the Grant JCYJ20220818103000001, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202101103. (Corresponding Author: Tin Lun Lam.)

Y. Zhou is with McGill University, Montréal, Canada. (email: yuhongze.zhou@mail.mcgill.ca)

L. Zhou, T. Lam, and Y. Xu are with School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China, and Shenzhen Institute of Artificial Intelligence and Robotics for Society. (email: liguangzhou@link.cuhk.edu.cn, tllam@cuhk.edu.cn, ysxu@cuhk.edu.cn)

Copyright © 2023 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

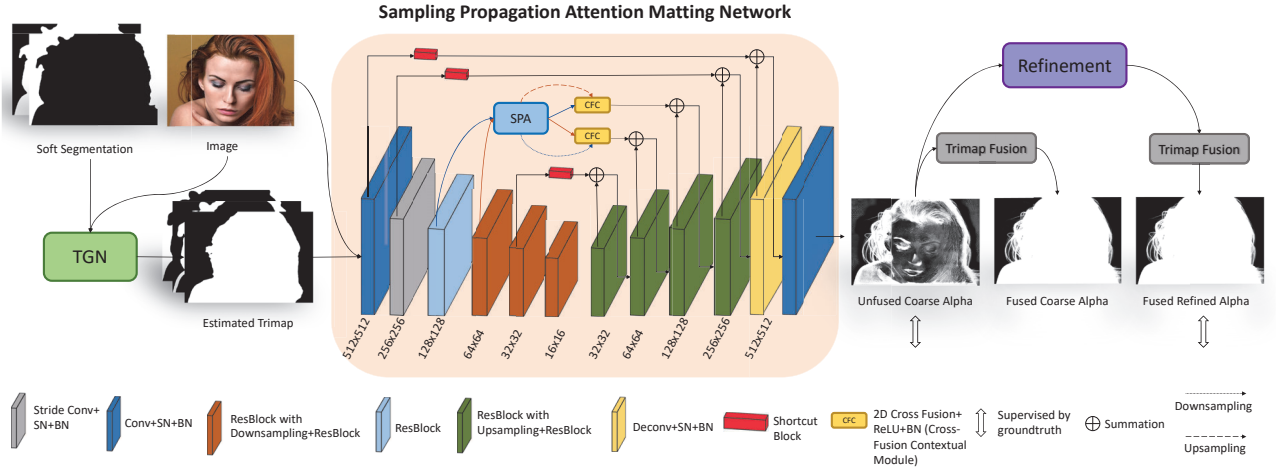

Fig. 2. The overview of our matting pipeline. With a coarse segmentation and image as input, TGN obtains an estimated trimap. With the trimap as guidance, Sampling Propagation Attention Matting Network (SPAMattNet) produces an alpha matte. The collaborative testing of TGN and SPAMattNet is denoted as Joint Inference. (SN is Spectral Normalization while BN is Batch Normalization.)

has a trimap (resp. foreground segmentation) quandary. To solve these puzzles, several trimap-free matting approaches have been proposed. Chen et al. [1] and Zhang et al. [20] estimate trimap constraints directly from images to assist human matting. Liu et al. [2] refine coarse masks sequentially by coupling coarse annotated data with fined one to promote human matting. Li et al. [23], [24] decompose matting into semantic segmentation and details matting by using a shared encoder and two separate decoders for collaborative learning. However, they mainly focus on single-category matting and usually require salient solid objects. Therefore, Yu et al. [25] use general coarse masks as guidance for matting but could suffer from performance degradation due to inaccurate masks. Li et al. [22] propose AIM-Net by investigating the possibility of extending GFM [24] to images with salient transparent/meticulous and non-salient objects with unified semantic representation. However, the shared encoder of AIM-Net might be unable to fully represent semantic/matting information, resulting in artifacts of "hard fusion" between results from semantic and matting decoders. And since their proposed Automatic Image Matting-500 benchmark mostly contains clean/blurred backgrounds without complicated multiply objects that conforms with real-world photography but simplifies the problem, their approach could degrade when encountering images with clear/complicated backgrounds.

Considering forementioned issues, as shown in Fig. 1, we argue that producing a high-quality trimap using coarse foreground segmentation is a critical cornerstone to automatic natural image matting. See detailed analyses in the supplementary material. The coarse segmentation can be produced by other segmentation/salient object detection/matting approaches [26], [27], which enables more flexibility for matting to adapt into different scenarios. Therefore, we disentangle matting into trimap and matte estimation subtasks as a workaround. Different from previous attempts, we aim to properly generalize trimap-free matting to comprehensive data. We propose a two-stage trimap-free matting approach, which consists of Trimap Generation Network (TGN) and Sampling Propagation Attention Matting Network (SPAMattNet). Since coarse foreground segmentation can provide additional semantic information and help the network capture rough locations and shapes of target objects, TGN employs a coarse foreground segmentation as guidance to estimate a proper trimap. Then, this estimated trimap serves as guidance for SPAMattNet and buffer for negative trimap quality chain reaction.

Different from previous propagation-based matting networks [11], [14], inspired by traditional sampling/propagation approaches [28], we propose a novel Sampling Propagation Attention for matting, denoted as SPAMattNet. Given an RGB image and its trimap, SPAMattNet simulates traditional sampling and propagation matting procedures in deep learning manner to enhance network explainability and performance. Specifically, SPAMattNet contains Local Sampling Propagation Attention (LSPA) and Global Propagation Attention (GPA) to investigate local spatial and global semantic information. In particular, we design distance maps between each unknown pixel and foreground/background pixels to reweight the affinity map for local sampling attention to conduct local image pixel sampling. We also introduce both local and global correlation computing for alpha feature propagation from foreground/background regions to unknown regions. Then, to dynamically aggregate sampling/propagation and local/global features, a Cross-Fusion Contextual Module (CFC) is introduced for more representative feature learning. Finally, we incorporate TGN and SPAMattNet to produce high-quality alpha matte without trimap and background as input.

We conduct experiments on synthetic matting datasets and show that our approach is competitive compared to other state-of-the-art methods on the Adobe Image Matting, alphamatting.com, and Distinctions-646 benchmarks from both trimap-needed and trimap-free perspectives. To demonstrate the real-world application capability of our approach, we conduct real data adaptation, which is evaluated by a user study, and validate our approach by extensive experiments on

challenging Real-World Portrait-636 [25], Privacy-Preserving Portrait Matting [23], and Automatic Image Matting-500 [22] benchmarks.

Our main contributions are summarized as follows:

- We propose a systematic and flexible trimap-free image matting pipeline boosted by coarse foreground segmentation that can be obtained by other segmentation/salient object detection/matting approaches, to estimate high-quality trimaps, leading to high-quality alpha mattes.
- We introduce a novel Sampling Propagation Attention into matting network, denoted as SPAMattNet, which unites sampling and propagation-based matting procedures for network explainability and performance improvement.
- We conduct extensive experiments to show that our approach is competitive compared to other state-of-the-art methods on Adobe Image Matting, alphamatting.com, Distinctions-646, Real-World Portraite-636, Privacy-Preserving Portrait Matting, and Automatic Image Matting-500 benchmarks.

II. Related Works

A. Natural Image Matting

Traditional image matting can be roughly classified into sampling-based [29]–[34] and propagation-based [28], [35]–[40] approaches, which usually require trimap as additional input. Sampling-based approaches [29]–[34] initially sample colors from foreground and background pixels for each unknown region defined by trimap and then select the best foreground-background color pair to compute alpha values according to quantitative metrics. Propagation-based methods [36]–[40] propagate alpha values from known pixels to unknown ones based on similarity measurements.

Recently, deep learning based matting approaches have shown prominent performance, which can be categorized into trimap-needed [6]–[19], [25], [41], background-required [3]–[5], and only-image [2], [20]–[23], [42] input, i.e., trimap-free. For trimap-needed methods, Xu et al. [7] propose a deep image matting solution along with a comprehensive matting dataset. Lutz et al. [8] explore matting task with a generative adversarial framework. Then, Hou et al. [41], Lu et al. [9], and Tang et al. [10] achieve appealing matting performance. Subsequently, Li et al. [11] propose a matting network with guided contextual attention. Zhou et al. [12] introduce an attention transfer network to extract objects from similar/complex backgrounds. Yu et al. [14] attempt to tackle high-resolution deep image matting. Yu et al. [25] introduce a progressive refinement network architecture with a series of guidance mask perturbation operations into matting task. Park et al. [18] introduce the first transformer based matting network that shows impressive performance. For background-required matting, Sengupta et al. [3] propose a matting approach that obtains appealing estimation but is not robust to images with shadow or under complex light conditions. Lin et al. [4] introduce a real-time high-resolution background matting approach. Xu et al. [5] propose a dataset-free unsupervised background matting approach. Besides, a few works attempt to take only image as input and produce alpha mattes [1], [2], [20]–[24], [42]. However, these trimap-free methods are not capable of producing high-quality alpha mattes unlike trimap-based matting approaches. Different from these trimap-free matting approaches, with coarse segmentation as guidance, we incorporate trimap-based matting into our two-stage trimap-free matting pipeline to enable more flexibility and produce high-quality alpha mattes.

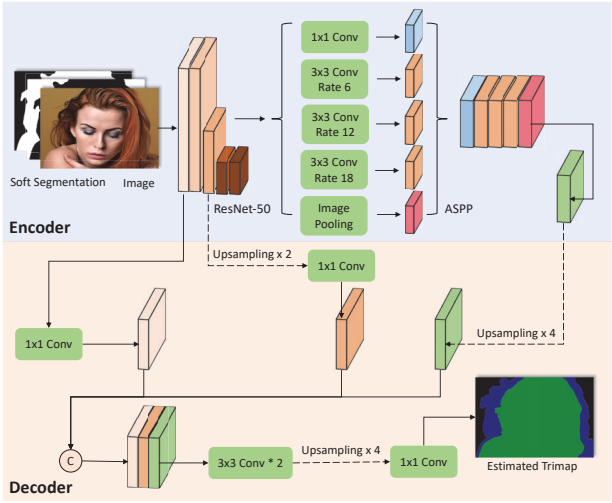

Fig. 3. Trimap Generation Network (TGN).

B. Attention Mechanism

Attention mechanism has been widely utilized in deep learning tasks like image synthesis [43] and semantic segmentation [44], [45]. The self-attention block contributes to each output position by referring to every input position [46]. Similarly, Wang et al. [47] propose non-local attention to acquire long-range contextual information and promote video classification tasks. Besides, the window-based self-attention mechanism [48], [49] has been introduced to reduce computational complexity. Recently, attention has shown its superiority in image matting [11], [14], [21] by making matting networks capture structural pixel-to-pixel dependencies to simulate the propagation-based matting procedure. Li et al. [11] simulate non-local attention by convolution and deconvolution to enhance alpha matte estimation. Qiao et al. [21] use channel/spatial-wise attention to filter out noise from hierarchical appearance cues and boost alpha mattes. Yu et al. [14] introduce three non-local attentions to propagate each trimap region of context patches to the corresponding region of query patches. Different from these attention-based matting networks, we introduce sampling propagation attention that successfully bridges traditional sampling/propagation and deep learning based matting approaches to enhance network explainability and performance.

C. Trimap Generation

Automatic trimap generation, which is popular in traditional matting [50]–[54], usually contains two steps: binary

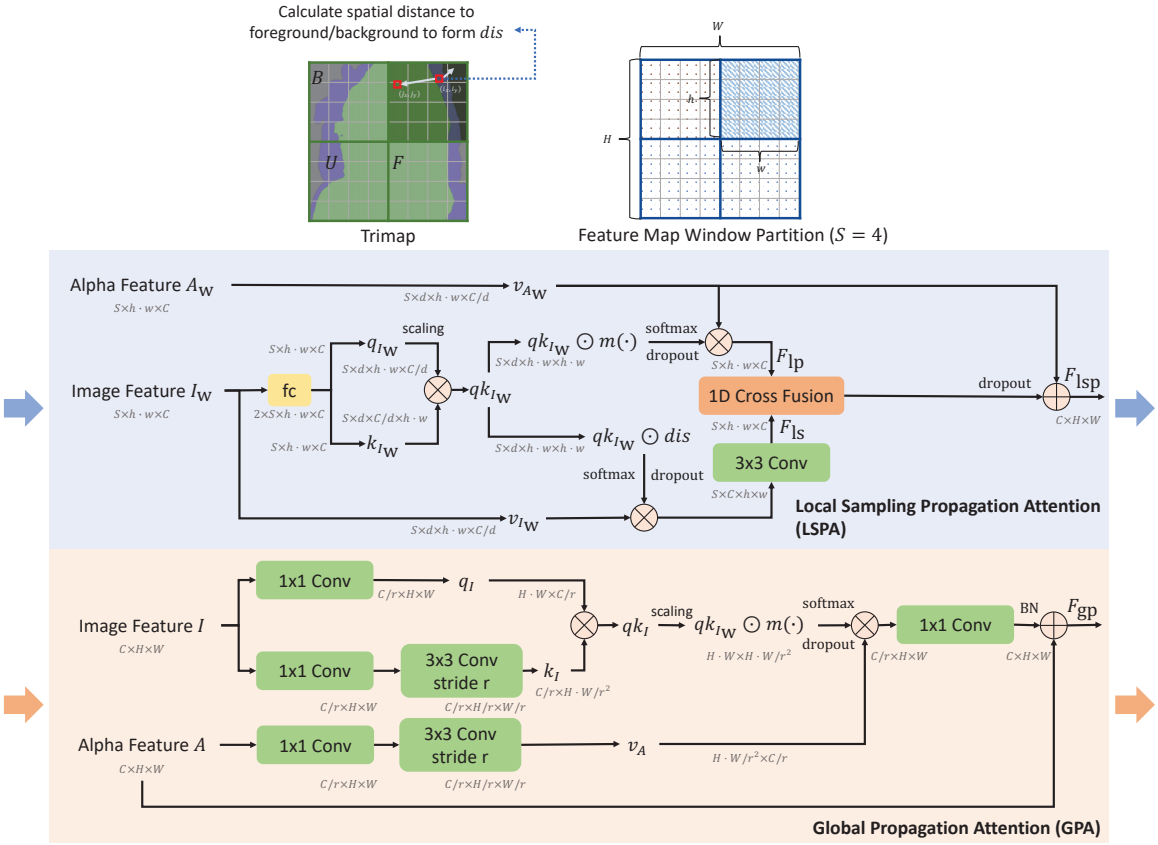

Fig. 4. The details of Sampling Propagation Attention Module (SPA). "⊕" means element-wise sum and "⊗" denotes matrix multiplication.

segmentation for foreground/background separation and image erosion/dilation. These methods mainly differentiate in how to obtain segmentation. For example, Wang et al. [50] leverage depth information to compute segmentation; Gupta et al. [53] combine salient object detection with super-pixel analysis for segmentation; Hsieh et al. [51] use graph cuts for foreground extraction; Chen et al. [54] require users to indicate foreground/background by a few clicks and apply one-shot learning for binary mask prediction. Besides, Al-Kabbany et al. [55] first introduce the Gestalt laws of grouping into matting, which assists more robust trimap generation. Cho et al. [56] integrate depth map with color distribution based processing for foreground/background separation and then introduce unknown region detection for trimap generation. Recently, neural networks have been utilized to generate implicit trimap for human matting automation [1], [57].

III. Approach

As shown in Fig. 2, we decompose our trimap-free matting approach into Trimap Generation Network (TGN) and Sampling Propagation Attention Matting Network (SPAMattNet). With coarse foreground segmentation as additional indicator, TGN attempts to understand target object shapes and relations with surroundings and perform pixel-wise classification among foreground/background/unknown regions. SPAMattNet utilizes an RGB image and TGN trimap to estimate an alpha matte.

A. Trimap Generation Network (TGN)

Trimap generation can be considered as a three-class semantic segmentation task. As shown in Fig 3, with the concatenation of an RGB image and a two-channel foreground segmentation as input, TGN outputs a three-channel feature map indicating the regional belonging possibility. We adopt Deeplabv3 [58] encoder, which consists of a ResNet-50 backbone [59] and Atrous Spatial Pyramid Pooling (ASPP). The high-level encoder feature first goes through a dropout layer by 0.5 factor and a four-time bilinear upsampling operation. Then, since directly adopting the high-level encoder feature for final classification may lose detailed spatial information, we aggregate it with low and middle-level encoder features. This aggregation can enrich the decoding process and make network emphasize on image appearance and less depend on segmentation input. And swapping the order between upsampling and the final convolution makes classification more fine-grained.

B. Sampling Propagation Attention Matting Network (SPAMattNet)

As illustrated in Fig. 2, we adopt the popular U-Net structure [11], [60] as the main matting architecture. With trimap

as prior, different from previous propagation-based attention augmented matting networks [11], [14], we augment U-Net with Sampling Propagation Attention (SPA) and Cross-Fusion Contextual Module (CFC) to fully exploit spatial/semantic relationships of image features across local/global domains and produce coarse alpha mattes. Then, we refine coarse alpha mattes with Refinement Module (RM) to estimate final alpha mattes.

1) Sampling Propagation Attention Module (SPA): In Fig. 4, we present the details of our Sampling Propagation Attention Module that contains two branches, i.e., Local Sampling Propagation Attention Module (LSPA) and Global Propagation Attention Module (GPA). We incorporate the low-level (resp. high-level) encoder feature with LSPA (resp. GPA) and bridge two different-level attention-aggregated features with 2D Cross-Fusion Contextual Module (CFC) as the decoder input.

Local Sampling Propagation Attention Module (LSPA) The detailed structure of Local Sampling Propagation Attention Module is illustrated at the top of Fig. 4. The Local Sampling Propagation Attention Module dynamically weights local sampling and propagation features to produce robust low-level features that are full of spatial information.

To be specific, we conduct LSPA locally within non-overlapping windows that partition feature maps (outlined in blue). Suppose that we partition feature maps into S evenly-arranged non-overlapping windows whose spatial dimensions are h and w. Consider an image feature map I_{w} and an alpha feature map A_{w}, where $\{I_{\text{w}}, A_{\text{w}}\} \in \mathbb{R}^{S\times h\cdot w\times C}$ and C is the number of channels. I_{w} is obtained by applying two 3×3 2-stride convolutions (with ReLU and SN/BN) on an input RGB image while A_{w} is the low-level feature map of U-Net. We first feed I_{w} into a fully connected layer to generate two feature maps $q_{I_{\text{w}}} \in \mathbb{R}^{S\times d\times h\cdot w\times C/d}$ and $k_{I_{\text{w}}} \in \mathbb{R}^{S\times d\times C/d\times h\cdot w}$, where d is the head number of multi-head self-attention operation. We reshape I_{w} to $v_{I_{\text{w}}} \in \mathbb{R}^{S\times d\times h\cdot w\times C/d}$ and A_{w} to $v_{A_{\text{w}}} \in \mathbb{R}^{S\times d\times h\cdot w\times C/d}$.

To simulate sampling procedure, the attention mechanism is applied on $q_{I_{\text{w}}}$, $k_{I_{\text{w}}}$, and $v_{I_{\text{w}}}$ to get local sampling feature F_{ls}, which is further refined by a 3×3 convolution,

$$F_{\text{ls}} = \text{softmax}\left(dis \odot \frac{q_{I_{\text{w}}}\cdot k_{I_{\text{w}}}}{\sqrt{C/d}}\right)\cdot v_{I_{\text{w}}}, \tag{2}$$

where $dis \in \mathbb{R}^{S\times d\times h\cdot w\times h\cdot w}$ is the distance map,

$$dis = D_F + D_B, \tag{3}$$

where $D_F \in \mathbb{R}^{S\times d\times h\cdot w\times h\cdot w}$ (resp. $D_B \in \mathbb{R}^{S\times d\times h\cdot w\times h\cdot w}$) is a distance map between unknown pixels and foreground (resp. background) pixels. The motivation behind D_F and D_B is as follows. The affinity operation usually emphasizes neighboring pixels since neighboring pixels usually have stronger similarities. Therefore, the local sampling branch without dis tends to collect samples near unknown pixels only. It will fail if good samples cannot be found nearby [32]. To alleviate this issue, we introduce dis to reweight the affinity map and make sampling somehow global. Therefore, for each window, the value of $D_F \in \mathbb{R}^{S\times d\times h\cdot w\times h\cdot w}$ in $(\cdot,\cdot,i,j)$ position can be formulated as

$$D_{F_{(\cdot,\cdot,i,j)}} = \begin{cases} \frac{d_{i,j}}{\min(d_{i_U,j_F})}, & (i_x,i_y)\in U, (j_x,j_y)\in F,\\ -d_{i,j}, & (i_x,i_y)\notin U, (j_x,j_y)\in F,\\ 0, & (j_x,j_y)\notin F, \end{cases}, \tag{4}$$

where $d_{i,j} = \sqrt{(i_x-j_x)^2+(i_y-j_y)^2}, i,j\in\{0,...,h\cdot w-1\}$, (i_x,i_y) (resp. (j_x,j_y)) denotes the pixel i (resp. j) position in the last two dimensions of non-reshaped $S\times C\times h\times w$ feature map that the i$^{\text{th}}$ (resp. j$^{\text{th}}$) index represents, F represents the foreground region, and U denotes the unknown region. The $\min(d_{i_U,j_F})$ is the minimum distance between the unknown pixel i to any foreground pixel j, which guarantees that dis is independent from the absolute distance. D_B also follows the similar calculation as D_F by replacing F as background B region.

Similarly, we generate local propagation feature F_{lp} by reconstructing alpha feature by image feature similarity,

$$F_{\text{lp}} = \text{softmax}\left(m \odot \frac{q_{I_{\text{w}}}\cdot k_{I_{\text{w}}}}{\sqrt{C/d}}\right)\cdot v_{A_{\text{w}}}. \tag{5}$$

The value of $m \in \mathbb{R}^{S\times d\times h\cdot w\times h\cdot w}$ in (i,j) position can be formulated as

$$m_{(i,j)} = \begin{cases} \text{clip}(\sqrt{|U|/|K|}) & (j_x,j_y)\in U,\\ \text{clip}(\sqrt{|K|/|U|}) & (j_x,j_y)\in K, \end{cases}, \tag{6}$$

$$\text{clip}(x) = \min\left(\max\left(x, 0.1\right), 10\right), \tag{7}$$

where $i,j\in\{0,...,h\cdot w-1\}$, (j_x,j_y) denotes the pixel j position in the last two dimensions of non-reshaped $S\times C\times h\times w$ feature map that the j$^{\text{th}}$ index represents, and $K=\{F,B\}$ refers to foreground F and background B regions [11].

To better fuse F_{ls} and F_{lp}, we reshape them into $\mathbb{R}^{S\times h\cdot w\times C}$. Then, it is fed into our 1D Cross-Fusion Contextual Module to encourage mutual communication between F_{ls} and F_{lp} and obtain more expressive feature representations F_{lsp}.

Global Propagation Attention Module (GPA) The bottom part of Fig. 4 shows details of our GPA. Propagation-based matting methods usually investigate color similarity between unknown and known areas and estimate opacity information as weighted combination of relevant alpha values based on similarity criteria. Since high-level features capture semantic information, to better incorporate semantic relationship information, we design Global Propagation Attention (GPA) to promote alpha feature learning.

Consider an image feature map I and an alpha feature map A, where $\{I,A\}\in\mathbb{R}^{C\times H\times W}$, I is the high-level image feature by applying a 3×3 2-stride convolution (with ReLU and SN/BN) on non-reshaped I_{w}, and A is the high-level feature map of U-Net. Due to high computational cost of affinity operation, we leverage linear transformations to transform I into low-dimensional space and obtain q_I and k_I. To be detailed, we reduce the channel dimension by a 1×1 convolution and downscale feature maps by a stride convolution, which can preserve information, speed up training, and prevent gradient explosion/vanishing. After that, q_I (resp. k_I) is reshaped into $\mathbb{R}^{H\cdot W\times C/r}$ (resp. $\mathbb{R}^{C/r\times H\cdot W/r^2}$),

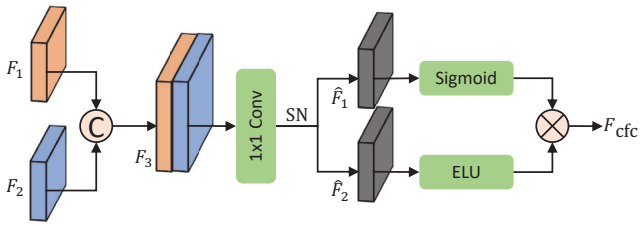

Fig. 5. Cross-Fusion Contextual Module.

where r is the downscale ratio. Similarly, A is transformed to $v_A \in \mathbb{R}^{H \cdot W/r^2 \times C/r}$. And the attention operation is applied to q_I, k_I, and v_A,

$$F_{\text{gp}} = \text{softmax}\left(m \odot \frac{q_I \cdot k_I}{\sqrt{C/r}}\right) \cdot v_A, \tag{8}$$

where $m \in \mathbb{R}^{H \cdot W \times H \cdot W/r^2}$ follows the same calculation as m in LSPA. Then, F_{gp} goes through a 1×1 convolution and residual summation to stabilize training and produce the final feature. We assume that our GPA can escort the encoder to understand semantic information of unknown areas.

2) Cross-Fusion Contextual Module (CFC): The Fig. 5 shows the 2D procedure of our CFC. Recall that features from different sources usually emphasize information in different aspects. For example, low-level features usually contain rich spatial details while high-level features are full of semantic information. Simply adding/concatenating different features with upsampling/downsampling would not fully take advantage of both. To this end, we introduce Cross-Fusion Contextual Module that dynamically fuses two different features to generate more expressive features.

Given two different-level features F_1 and F_2 that have the same size, we concatenate F_1 and F_2 along the channel (resp. last) dimension for 2D (resp. 1D) features and denote it as F_3. Then, F_3 is fed to a gating operation [61] to obtain fully aggregated features. Specifically, F_3 first goes through a linear transformation operation, i.e., a 1×1 convolution for 2D or a fully connected layer for 1D, and Spectral Normalization. Then, it is splitted evenly along the concatenated dimension to get $\hat{F}_1$ and $\hat{F}_2$. Finally, a pixel-wise attention is applied to obtain the fully aggregated feature F_{cfc},

$$F_{\text{cfc}} = \phi(\hat{F}_1) \odot \sigma(\hat{F}_2), \tag{9}$$

where ϕ is ELU activation function and σ is sigmoid function.

3) Refinement Module (RM): Refinement techniques have shown impressive performance and effective generalization in many matting-related works, including salient object detection [62] and semantic segmentation [63], [64], and matting works [2], [25]. Therefore, we use refinement module to refine the predicted coarse alpha α_{coarse} by learning the residual α_{residual} between the coarse alpha and ground truth as $\alpha_{\text{refined}} = \alpha_{\text{coarse}} + \alpha_{\text{residual}}$.

4) Loss Function: The cross-entropy loss L_{ce} [65] is leveraged in TGN. For SPAMattNet, its training loss L_α covers both coarse and refined alpha estimations,

$$L_\alpha = L_{\text{coarse}} + L_{\text{refined}}. \tag{10}$$

To obtain high-quality alpha mattes, we employ alpha prediction loss L_{alpha} and alpha hard mining loss L_{hard} for coarse and refined alpha supervisions. The alpha prediction loss [7] is defined as the absolute difference between the ground truth and predicted alpha,

$$L_{\text{alpha}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} |\hat{\alpha}_i - \alpha_i|, \tag{11}$$

where $\mathcal{M}$ refers to the unknown region and $\hat{\alpha}_i$ and α_i indicate predicted and ground-truth alpha value at position i respectively. Since the network learning ability on each alpha value varies, indiscriminately treating each alpha value might misguide the training process. To address this issue, we introduce hard mining loss [66], [67]. The hard mining loss calculates the absolute difference between the ground truth and predicted alpha, sorts all pixels, and picks the top p percent of the largest error pixels as hard samples to focus on. The hard mining loss is formulated as

$$L_{\text{hard}} = \frac{1}{|\mathcal{HM}|} \sum_{i \in \mathcal{HM}} |\hat{\alpha}_i - \alpha_i|, \tag{12}$$

where $\mathcal{HM}$ means the region that contains hard samples.

IV. Experiments

A. Datasets

We conduct experiments on three synthetic datasets, real-world human data, and three real-world matting datasets with annotations.

1) Synthetic Datasets:

- Adobe Image Matting Benchmark (AIM) [7]: It has 431 foreground images for training and 50 foreground images for testing. We follow the composition rule as Xu et al. [7] to synthesize training and test sets. The training set contains 43,100 images whose background images are from COCO dataset [68]. The test set, also known as the Composition-1k test set, contains 1,000 images whose background images are from Pascal VOC dataset [15].
- Distinctions-646 Benchmark [21]: It consists of 646 diversified foreground images [21]. Following the same composition rule as AIM, we synthesize 59,600 images for training and 1,000 images for testing.
- AlphaMatting.com Benchmark [69]: It contains eight images. Each image has three different trimaps, i.e., "small", "large" and "user".

2) Real-world Human Data: We capture 37 handheld videos, in which the subject is moving around and camera is moved randomly. The training set is the combination of our self-captured videos and 10 real-world videos from Background Matting [3], totally 22,144 real-world video frames. The test set contains 100 images by combining 10 uniformly-sampled frames of each testing video among 10 testing videos (half self-captured, half Background Matting videos). We also introduce 27 background videos, which are either self-captured or from Background Matting, to amplify data diversity.

3) Real-world Matting Datasets with Annotations:

- Real-world Portrait-636 Benchmark (RWP-636) [25] : It consists of 636 diverse and high-resolution images with matting annotations, coarse foreground segmentation masks, and labeled detail masks covering the hair region and other soft tissues.
- Privacy-Preserving Portrait Matting Benchmark (P3M-10k) [23]: It contains 10,000 anonymized high-resolution portrait images with face obfuscation and corresponding annotated alpha mattes. There are 9,421 images for training and 500 images for testing, denoted as P3M-500-P. Besides, there are another 500 public celebrity images with alpha matte annotations but without face obfuscation, denoted as P3M-500-NP, to evaluate the performance of matting models under the privacy-preserving training (PPT) setting (training on blurred images but testing on blurred/non-blurred images).
- Automatic Image Matting-500 Benchmark (AIM-500) [22]: It is a high-resolution natural image matting test set, including 424 salient opaque (SO), 43 salient transparent/meticulous (STM), and 33 non-salient (NS) images. To evaluate on it, the training set is the combination of DUTS [70] and the synthetic data by compositing foregrounds of AIM, Distinction-646, and AM-2k [24] with the high-resolution BG-20k [24] background dataset. We follow the same process as Li et al. [22] to generate the synthetic data.

B. Implementation Details

1) TGN: In training, we randomly generate coarse foreground segmentation input. Its generation process is as follows: (1) A random trimap is first produced by erosion/dilation on alpha matte with a random kernel size within $[1, 29]$; (2) The unknown and foreground areas of the random trimap is considered as the foreground of the random *initial segmentation*; (3) A random coarse segmentation is generated by erosion/dilation on the random *initial segmentation* with a random kernel size within $[1, 59]$ and followed by a random Gaussian Blur. Further, to obtain *synthesized ground-truth trimap*, we apply erosion/dilation on alpha matte with 15×15 kernel size. Please refer to the supplementary material for more details. Finally, image patches are randomly cropped from input images and then resized to 512×512. We train TGN for 129,300 iterations with 10 batch size. The learning rate is initialized to 0.001 and adjusted in every iteration. In testing, the coarse segmentation input is generated by erosion on *initial segmentation* derived from *synthesized ground-truth trimap* with 20×20 kernel size and followed by a Gaussian Blur.[1] In inference, we use original RGB images as network input.

2) SPAMattNet: We follow the similar data processing and augmentation procedure as GCA Matting [11]. SPAMattNet is trained for 200,000 iterations with 40 batch size and L_α, where

[1]Salient/segmentation models may not be suitable for coarse segmentation generation here, because (1) categories of foreground objects of synthetic matting datasets may not match segmentation datasets; (2) background objects may have the same category as foreground and can also be salient.

TABLE I
THE QUANTITATIVE COMPARISON OF TGN WITH ADAPTED DEEPLABV3 AND DEEPLABV2 ON ADOBE IMAGE MATTING AND DISTINCTIONS-646 BENCHMARKS.

AIM					
Methods	pixAcc	mIoU-Bg	mIoU-Unk	mIoU-Fg	mIoU
Deeplabv2 [71]	93.58	89.81	76.25	63.09	76.38
Deeplabv3 [58]	95.83	91.59	80.25	**70.66**	80.83
TGN	**96.41**	**93.53**	**82.78**	70.64	**82.32**
Distinctions-646					
Methods	pixAcc	mIoU-Bg	mIoU-Unk	mIoU-Fg	mIoU
Deeplabv2 [71]	92.58	94.40	72.97	53.82	73.73
Deeplabv3 [58]	95.66	95.11	79.65	65.52	80.09
TGN	**95.74**	**96.54**	**80.83**	**66.72**	**81.36**

TABLE II
COMPARISON OF DIFFERENT SEGMENTATION INPUTS FOR TGN ON ADOBE IMAGE MATTING AND DISTINCTIONS-646 BENCHMARKS.

AIM					
Methods	pixAcc	mIoU-Bg	mIoU-Unk	mIoU-Fg	mIoU
TGN-20	**96.41**	**93.53**	**82.78**	**70.64**	**82.32**
TGN-30	96.34	93.26	82.57	70.46	82.10
TGN-40	96.16	92.77	81.74	70.14	81.55
TGN-50	95.90	92.25	80.58	69.49	80.77
Distinctions-646					
Methods	pixAcc	mIoU-Bg	mIoU-Unk	mIoU-Fg	mIoU
TGN-20	**95.74**	**96.54**	**80.83**	66.72	**81.36**
TGN-30	95.66	96.20	80.22	66.82	81.08
TGN-40	95.57	96.00	79.60	**66.94**	80.84
TGN-50	95.31	95.68	78.34	66.72	80.25

$\mathcal{M}$ is the unknown region of trimap. The Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is adopted with the learning rate initialized as 4×10^{-4} plus warmup and cosine decay techniques. p is set to 50.

3) Joint Inference (JI): Joint Inference (JI) is the collaborative testing of TGN and SPAMattNet with trimap fusion techniques [1] by adopting TGN trimap as prior for SPAMattNet. From Fig. 2, we can see that raw alpha estimation of SPAMattNet is far away from impressive except unknown region. To this end, we introduce two trimap fusion methods. One is probability-based soft fusion, another is region-based hard fusion. For soft fusion, we use region probabilities estimated by TGN to reconstruct the final alpha $\widetilde{\alpha}$ from the predicted alpha $\hat{\alpha}$ [1] as

$$\widetilde{\alpha} = (1 - \hat{U})\frac{\hat{F}}{\hat{F} + \hat{B}} + \hat{U}\hat{\alpha}, \tag{13}$$

$$\hat{U} = 1 - (\hat{F} + \hat{B}), \tag{14}$$

$$\widetilde{\alpha} = \hat{F} + \hat{U}\hat{\alpha}, \tag{15}$$

where $\hat{U}$, $\hat{F}$, and $\hat{B}$ are probabilities of each pixel belonging to unknown/foreground/background regions severally. Soft fusion bridges TGN and SPAMattNet by enabling differentiability. For hard fusion, we reset trimap-predicted foreground (resp. background) to 255 (resp. 0) on the predicted alpha.

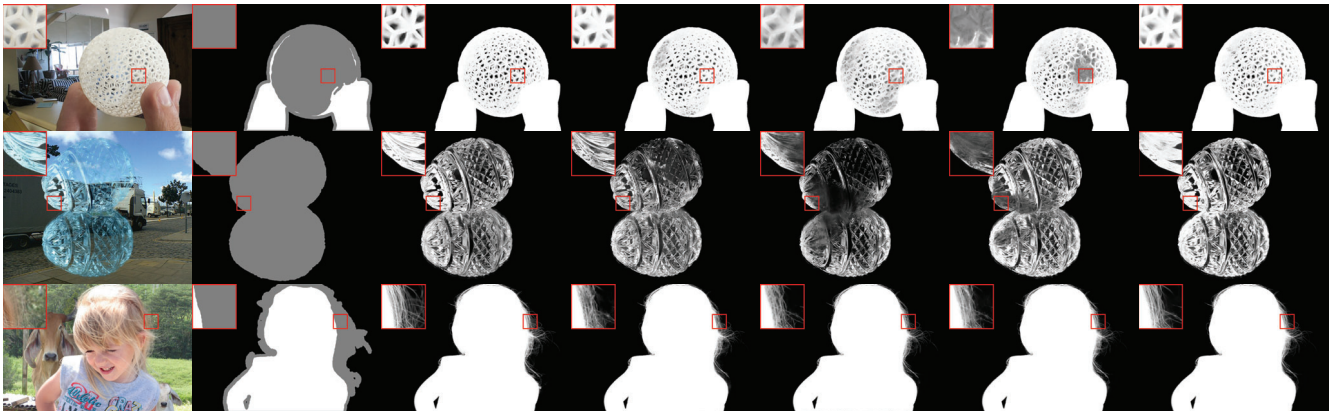

Fig. 6. Visual comparison on the Composition-1k test set.

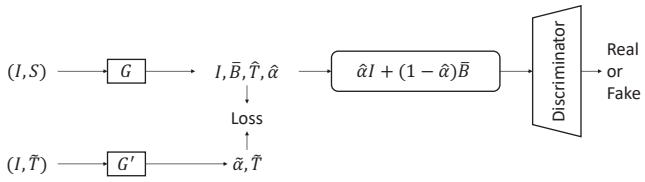

Fig. 7. Our real data adaptation architecture. Given an input image I and its coarse segmentation S, the generator G is jointly optimized by pseudo trimap $\widetilde{T}$-guided G' and the discriminator. $\widetilde{\alpha}$ is the pseudo alpha. $\hat{T}, \hat{\alpha}$ are the estimated trimap and alpha.

4) Real Data Adaptation: As shown in Fig. 7, we provide a real-world adaptation method that can be integrated with various trimap-based matting approaches to demonstrate trimap-free application capability of our approach in the real world, unlike Sengupta et al. [3] focusing on background-required approaches. Given an input image I and its coarse segmentation S derived from the official CIHP-pretrained [72] Parsing R-CNN [26] segmentation, the generator G, including TGN and SPAMattNet, produces an alpha matte $\hat{\alpha}$. To optimize the generator, we leverage a human-finetuned SPAMattNet G' and a discriminator D to guide G. For the human-finetuned SPAMattNet G', given I and a pseudo trimap $\widetilde{T}$, it produces a pseudo alpha matte $\widetilde{\alpha}$. The discriminator D attempts to distinguish if the input image is a newly composed image using $\hat{\alpha}$ or a real-world image. Specifically, the generator G minimizes

$$\begin{aligned}\min_{\theta_{\text{Real}}} E_{X,\bar{B}\sim p_{X,\bar{B}}}\left[(D(\hat{\alpha}I+(1-\hat{\alpha})\bar{B})-1)^2\right] \\ +\lambda\left[L_{\text{ce}}(\widetilde{T},\hat{T})+L_{\alpha}(\widetilde{\alpha},\hat{\alpha})\right],\end{aligned} \tag{16}$$

where $X = [I, S]$, θ_{Real} is the G parameters initialized by AIM-pretrained TGN and SPAMattNet, $\bar{B}$ is a randomly-selected background image, and λ is 0.5 and reduced by $\frac{1}{2}$ every 10,000 iterations during training. Note that L_{α} is calculated in the whole image. The discriminator D minimizes

$$\begin{aligned}\min_{\theta_{\text{Disc}}} E_{X,\bar{B}\sim p_{X,\bar{B}}}\left[(D(\hat{\alpha}I+(1-\hat{\alpha})\bar{B}))^2\right] \\ +E_{I\in p_{\text{data}}}\left[(D(I)-1)^2\right],\end{aligned} \tag{17}$$

TABLE III
TRIMAP-NEEDED EVALUATION ON THE COMPOSITION-1K TEST SET. - INDICATES NOT GIVEN IN THE ORIGINAL PAPER.

Methods	SAD ↓	MSE↓	Grad↓	Conn↓
AlphaGAN [8]	52.40	0.0300	38.00	53.00
Deep Image Matting [7]	50.40	0.0140	30.00	50.80
IndexNet Matting [73]	45.80	0.0130	25.90	43.70
AdaMatting [74]	41.70	0.0100	16.80	-
SampleNet Matting [10]	40.35	0.0099	-	-
Context-Aware Matting [41]	35.80	0.0082	17.30	33.20
GCA Matting [11]	35.28	0.0091	16.92	32.53
ATNet [12]	40.50	0.0130	21.50	39.40
HDMatt [14]	33.50	0.0073	14.50	29.90
MGMatting [25]	31.50	0.0068	13.50	27.30
SPAMattNet	**27.80**	**0.0055**	**11.57**	**23.63**

TABLE IV
THE ABLATION STUDY OF TRIMAP-NEEDED EVALUATION ON THE COMPOSITION-1K TEST SET. RM MEANS REFINEMENT MODULE.

GPA	LSPA	CFC	RM	L_{hard}	SAD ↓	MSE↓	Grad↓	Conn↓
			✓	✓	33.17	0.0076	14.90	30.23
	✓	✓	✓	✓	29.81	0.0066	12.43	26.17
✓			✓	✓	29.07	0.0059	11.73	25.19
✓	✓		✓	✓	30.13	0.0062	12.79	26.68
✓	✓	✓		✓	28.43	0.0060	**11.29**	24.34
✓	✓	✓	✓		28.42	0.0059	11.72	24.64
✓	✓	✓	✓	✓	**27.80**	**0.0055**	11.57	**23.63**

where θ_{Disc} is the D parameters. The soft fusion aggregates TGN and SPAMattNet of G into an end-to-end trimap-free matting framework. The end-to-end training and the interaction between G and D can help G jump out of the local minimum of G'. During training, the learning rate is initialized to 10^{-4} for TGN and 4×10^{-5} for SPAMattNet with 6 batch size, 100,000 iterations. Please refer to the supplementary material for more details.

5) Real-world Portrait-636 Benchmark (RWP-636): To validate the real-world generalization ability of our SPAMattNet, we directly evaluate SPAMattNet that is trained on synthetic matting datasets on RWP-636. Specifically, SPAMattNet, unlike MGMatting [25], is trained on AIM training set (including

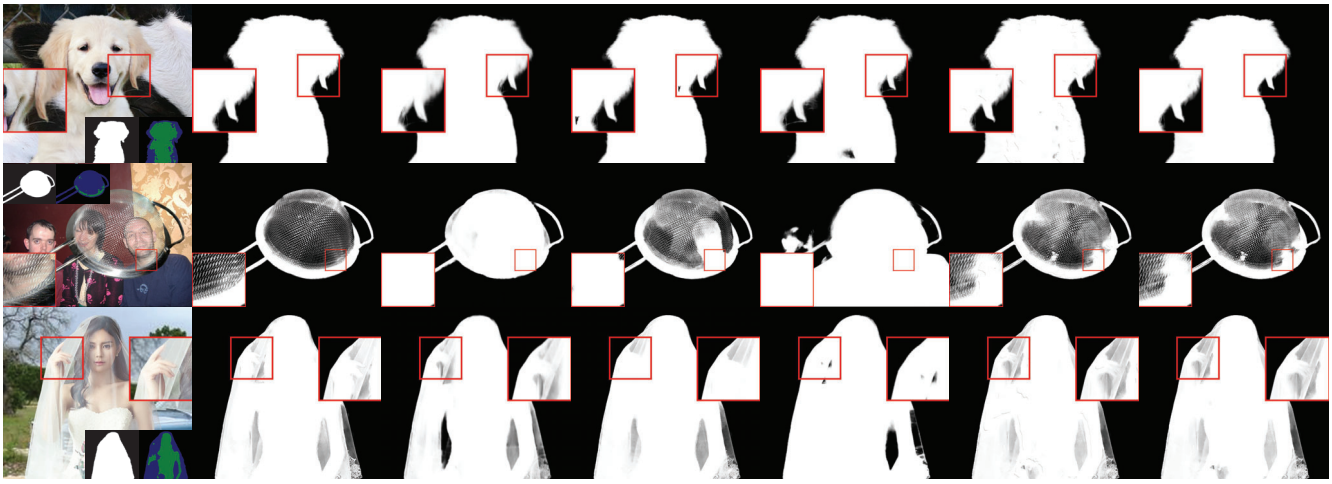

Fig. 8. Visual comparison on the Composition-1k test set in trimap-free setting. The guidance from left to right is Mask used by TGN/MGMatting and TGN Trimap.

transparent objects) without data augmentation strategies that bridges the gap between synthetic and real-world data. We generate pseudo trimaps as Yu et al. [25] do. Then, we resize images/trimaps to 512×512 for inference and then resize estimated alpha mattes back.

6) Privacy-Preserving Portrait Matting Benchmark (P3M-10k): For P3M-10k, we leverage a simple training strategy to validate our pipeline, which can be further improved. Specifically, we finetune the Supervisely Person Dataset [75]-pretrained U2Net [27] on the P3M-10k training set for 100 epoches with MSE loss, an initialized learning rate of 0.001, and 32 batch size, to obtain coarse foreground masks. We train TGN with coarse foreground masks estimated by the finetuned U2Net as additional input with 16 batch size for 88,322 iterations. We apply erosion/dilation on alpha mattes with 30×30 kernel size to generate ground-truth trimaps for TGN training. Then, we train SPAMattNet using TGN trimap as prior by calculating losses in both trimap and whole image regions with 16 batch size for 300,000 iterations. We use the same random cropping strategy as Li et al. [23].

7) Automatic Image Matting-500 Benchmark (AIM-500): For AIM-500, we follow a similar training procedure/configuration as Section IV-B6 except as noted. We finetune the DUTS-pretrained [70] U2Net [27] on the combined training set to obtain coarse foreground masks and train TGN with estimated coarse foreground masks as additional input for 120,550 iterations. Then, we train SPAMattNet for 120,550 iterations. We use the same random cropping strategy as Li et al. [22].

C. Evaluation

1) Trimap Evaluation: We use pixel classification accuracy (pixAcc) and mean IoU (mIoU) metrics of background(Bg)/unknown(Unk)/foreground(Fg)/three-region-involved to evaluate the performance of trimap estimation.

2) Alpha Matte Evaluation: We follow common evaluation metrics, including Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Mean Absolute Difference (MAD), Gradient error (Grad), and Connectivity error (Conn), to evaluate alpha mattes.

TABLE V
TRIMAP-FREE EVALUATION ON COMPOSITION-1K AND DISTINCTIONS-646 TEST SETS. JIS AND JIH MEAN JOINT INFERENCE (JI) WITH SOFT AND HARD FUSION RESPECTIVELY.

Composition-1k test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
Late Fusion [20]	58.34	0.0110	41.63	59.74
HAttMatting [21]	44.01	0.0070	29.26	46.41
MGMatting [25]	41.39	**0.0040**	19.38	36.37
MODNet [42]	290.08	0.1048	137.22	290.41
JIS	40.51	0.0046	17.44	35.88
JIH	**37.95**	0.0045	**16.82**	**34.76**
Distinctions-646 test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
Late Fusion [20]	94.23	0.0271	159.78	62.15
HAttMatting [21]	48.98	0.0090	41.57	49.93
MGMatting [25]	46.53	**0.0076**	35.68	43.72
MODNet [42]	234.92	0.0970	221.35	232.90
JIS	42.09	0.0082	37.16	39.77
JIH	**38.79**	0.0082	**36.89**	**38.64**

D. Results

1) TGN: As shown in Table I, we evaluate TGN and other popular adapted semantic segmentation methods, i.e., Deeplabv3 [58] and Deeplabv2 [71], on Adobe Image Matting and Distinctions-646 benchmarks. The results show that trimap segmentation accuracy and mIoU metrics of TGN are superior to other methods. Considering quantitative results in Table I and visual results presented in Fig. 8 and 9, our trimap estimation is competent to mentor SPAMattNet.

Ablation Study We conduct an ablation study to investigate the influence of quality of coarse segmentation on TGN. We

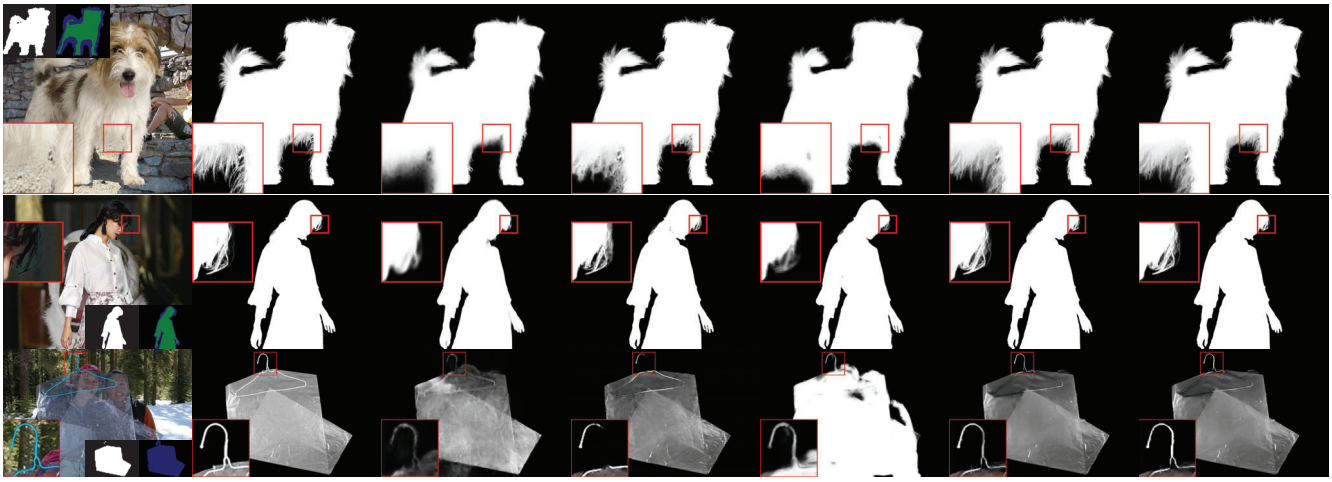

Fig. 9. Visual comparison on Distinctions-646 benchmark in trimap-free setting.

TABLE VI
THE ABLATION STUDY OF TRIMAP-FREE EVALUATION ON COMPOSITION-1K AND DISTINCTIONS-646 TEST SETS.

Composition-1k test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
JIS w/o RM	43.91	0.0047	17.98	37.67
JIH w/o RM	38.36	**0.0045**	16.96	35.12
JIS	40.51	0.0046	17.44	35.88
JIH	**37.95**	**0.0045**	**16.82**	**34.76**
Distinctions-646 test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
JIS w/o RM	45.61	**0.0082**	**35.89**	41.24
JIH w/o RM	39.68	0.0084	37.35	39.18
JIS	42.09	**0.0082**	37.16	39.77
JIH	**38.79**	**0.0082**	36.89	**38.64**

evaluate TGN with *initial segmentation* eroded with 20×20, 30×30, 40×40, and 50×50 kernel sizes and followed by a Gaussian Blur (denoted as "network-X", e.g., TGN-20, JIS-20, JIH-20, etc.). We assume that it can simulate flawed applied circumstances. The results are shown in Table II. The accuracy is all above 90% and the mean IoU of unknown region is fluctuating around 0.80, which implies that TGN is anti-jamming towards imperfect foreground segmentation and capable of offering correct trimap estimation.

2) SPAMattNet: We evaluate SPAMattNet in trimap-needed setting on Adobe Image Matting and alphamatting.com benchmarks.

Adobe Image Matting Benchmark (AIM) In Table III, we compare our approach with AlphaGAN [8], Deep Image Matting (DIM) [7], IndexNet Matting [9], AdaMatting [74], SampleNet Matting (SNMatting) [10], Context-Aware Matting (CAMatting) [41], GCA Matting [11], ATNet [12], HDMatt [14], and MGMatting [25]. With trimap provided, SPAMattNet exhibits dominating performance on all four metrics compared with other methods. In Fig. 6, we represent qualitative comparison between our approaches and other state-of-the-art methods. Our approach is capable of obtaining more fine-grained details of transparent objects than others do.

AlphaMatting.com Benchmark In Table VII, our SPAMattNet is comparable with other state-of-the-art methods on the alphamatting.com benchmark at the time of submission. The Fig. 10 represents the visual comparison between our approach and other methods on alphamatting.com benchmark.

Attention Visualization In Fig. 11, given an image query patch, we visualize affinity maps of GPA and LSPA as well as their reconstructed alpha features. The brighter the color is, the larger affinity value the pixel holds and more significant the pixel is. It is obvious that GPA can not only accurately select color-relevant pixels but also capture long-distance pixel-to-pixel relationships. From two visualized alpha features, we can see that LSPA concentrates on local detailed texture information while GPA emphasizes on more global high-level semantic information.

Ablation Study In Table IV, we present the ablation study for SPAMattNet. We can see that SPAMattNet achieves the best performance in SAD, MSE, and Conn metrics, compared to SPAMattNet without attention/CFC/RM/L_{hard}, which demonstrates the effectiveness of each part of our system. Note that, with LSPA being removed, CFC will also be removed because 2D CFC is to fuse features from LSPA and GPA. It is noted that CFC can produce better feature representation for learning than simply adding/concatenating. The results show that our SPA module is a strong performance augmenter and the vanilla U-Net architecture without SPA becomes moderate for matting task.

Stronger Decoder and Loss Functions To further demonstrate the effectiveness of SPA module, we replace SPAMattNet decoder by MGMatting decoder with its corresponding l_1 regression loss, composition loss [7], Laplacian loss [41], learning rate, and batch size, but without ASPP and RM (denoted as SPAMattNet*). In Table IX, we compare the performance, number of parameters, and FLOPs of our SPAMattNet* to MGMatting. Input image size is 512×512 to

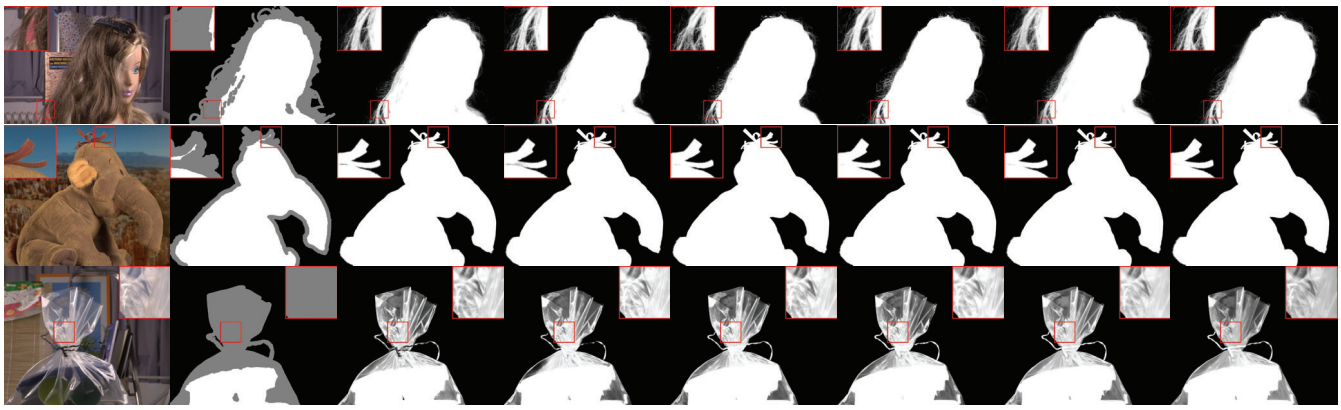

Fig. 10. Visual results on alphamatting.com benchmark.

TABLE VII
SAD RESULTS ON ALPHAMATTING.COM BENCHMARK, WHERE S, L, U REPRESENT THE TRIMAP TYPE OF SMALL, LARGE AND USER.

SAD↓	Overall rank	Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
		S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
AdaMatting [74]	15	10.2	11.1	10.8	4.9	5.4	6.6	3.6	3.4	3.4	**0.9**	**0.9**	1.8	**4.7**	6.8	9.3	2.2	2.6	3.3	19.2	19.8	18.7	**17.8**	19.1	**18.6**
SampleNet Matting [10]	15.8	9.1	9.7	9.8	**4.3**	4.8	**5.1**	3.4	3.7	3.2	**0.9**	1.1	2.0	5.1	6.8	9.7	2.5	4.0	3.7	18.6	19.3	19.1	20.0	21.6	23.2
GCA Matting [11]	17.3	8.8	9.5	11.1	4.9	4.8	5.8	3.4	3.7	3.2	1.1	1.2	**1.3**	5.7	6.9	7.6	2.8	3.1	4.5	18.3	19.2	18.5	20.8	21.7	24.7
HDMatt [14]	12.7	9.5	10.0	10.7	4.7	4.8	5.8	2.9	3.0	**2.6**	1.1	1.2	1.3	5.2	**5.9**	**6.7**	2.4	2.6	3.1	17.3	17.3	17.0	21.5	22.4	23.2
TransMatting [76]	14.3	**7.5**	**8.6**	**7.8**	4.9	5.0	5.5	3.4	3.5	3.2	1.8	1.8	2.2	5.7	6.3	8.4	2.6	2.7	**3.0**	17.6	18.0	18.9	18.3	**18.3**	21.5
Ours	**10.1**	9.0	10.6	9.6	4.5	**4.4**	5.4	**2.6**	**2.7**	2.8	**0.9**	**0.9**	1.8	5.2	6.1	9.6	**2.0**	**2.3**	3.4	**15.1**	**15.2**	**15.2**	19.6	19.7	24.5

TABLE VIII
COMPARISON OF DIFFERENT SEGMENTATION INPUTS FOR JI ON COMPOSITION-1K AND DISTINCTIONS-646 TEST SETS.

Composition-1k test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
JIS-20	40.51	0.0046	17.44	35.88
JIS-30	40.75	0.0046	17.51	36.06
JIS-40	41.15	0.0047	17.71	36.39
JIS-50	42.07	0.0050	18.22	37.27
JIH-20	**37.95**	**0.0045**	**16.82**	**34.76**
JIH-30	38.09	0.0046	16.86	34.91
JIH-40	38.28	0.0047	16.97	35.09
JIH-50	39.01	0.0049	17.46	35.81

Distinctions-646 test set				
Methods	SAD↓	MSE↓	Grad↓	Conn↓
JIS-20	42.09	0.0082	37.16	39.77
JIS-30	42.08	**0.0081**	37.15	39.73
JIS-40	42.63	**0.0081**	38.20	40.16
JIS-50	44.21	0.0086	39.75	41.66
JIH-20	38.79	0.0082	36.89	38.64
JIH-30	**38.61**	**0.0081**	**36.46**	**38.49**
JIH-40	38.89	**0.0081**	37.26	38.75
JIH-50	40.25	0.0086	39.09	40.13

TABLE IX
STRONGER DECODER AND LOSS FUNCTIONS. ⋆ DENOTES REPLACING THE DECODER OF SPAMATTNET WITH MGMATTING DECODER BUT WITHOUT ASPP AND RM.

Methods	SAD↓	MSE↓	Grad↓	Conn↓	Params	FLOPs
MGMatting [25]	31.50	0.0068	13.50	27.30	29.7M	**45.7G**
SPAMattNet⋆	**28.58**	**0.0053**	**10.60**	**24.80**	**26.0M**	49.8G

calculate FLOPs. SAD, MSE, Grad, and Conn are calculated on the Composition-1k test set in trimap-needed setting. Note that, except SPA, SPAMattNet⋆ and MGMatting share the same U-Net backbone and SPAMattNet⋆ does not use ASPP unlike MGMatting. Our SPAMattNet⋆, with a bit increase of FLOPs, has fewer parameters than MGMatting but shows better performance. Therefore, Table IX suggests that our SPA module is affordable and effectively works on convolution-based backbones.

3) Joint Inference (JI): We evaluate JI in trimap-free setting on AIM and Distinctions-646 benchmarks with coarse segmentation as additional input. For the generation of coarse segmentation, please refer to Section IV-B1. JIS and JIH mean JI with soft and hard fusion respectively.

Adobe Image Matting Benchmark (AIM) In Table V, we compare our JI with MGMatting [25] and trimap-free Late Fusion [20], HAttMatting [21], and MODNet [42] on AIM benchmark. For MGMatting and JI, we use the same coarse segmentation as additional input. The results reveal much more improvement of JI on SAD, Grad, and Conn metrics. As shown in Fig. 8, our approaches are robust to images with

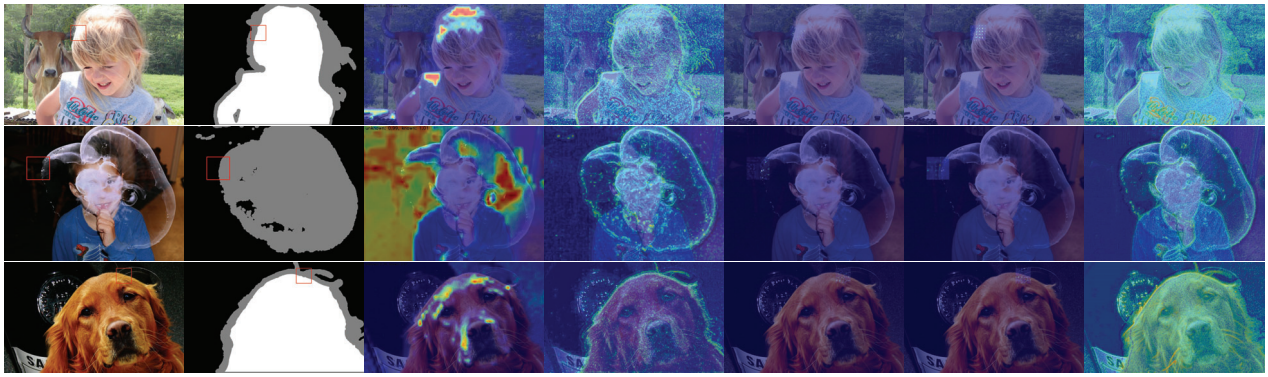

Fig. 11. The attention visualization. Given query patch marked by red box, from left to right, image, trimap, affinity map of GPA, alpha feature of GPA, affinity map of LSPA propagation branch, affinity map of LSPA sampling branch, and alpha feature of LSPA.

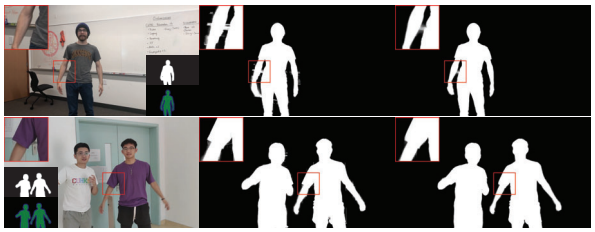

Fig. 12. Visual comparison on real-world human data between CAM and Ours. From left to right, image w/ guidance, CAM, and Ours. The guidance from top to bottom is Mask used by TGN and TGN Trimap.

TABLE X
RESULTS ON REAL-WORLD PORTRAIT-636 BENCHMARK BY TRAINING ON AIM TRAINING SET IN TRIMAP-FREE SETTING.

Methods	Whole Image		Details	
	SAD↓	MSE↓	SAD↓	MSE↓
Deep Image Matting [7]	28.5	0.0117	19.1	0.0746
GCA Matting [11]	29.2	0.0127	19.7	0.0823
IndexNet Matting [9]	28.5	0.0115	18.8	0.0727
Context-Aware Matting [41]	27.4	0.0107	18.2	0.0662
Late Fusion [20]	78.6	0.0398	24.2	0.0883
MGMatting [25]	47.9	0.0174	20.0	0.0716
MGMatting* [25]	26.8	0.0093	17.4	0.0551
SPAMattNet	27.0	0.0107	17.6	0.0654
SPAMattNet*	**26.9**	**0.0105**	**17.3**	**0.0632**

complicated backgrounds compared to other state-of-the-art methods.

Distinctions-646 Benchmark Similar to AIM benchmark, in Table V, we compares our approach with MGMatting [25] and trimap-free Late Fusion [20], HAttMatting [21], and MODNet [42]. Our JI approach exceeds Late Fusion, HAttMatting, and MODNet in all four metrics by a solid margin, especially in SAD and Conn metrics, and outperforms MGMatting in SAD, Grad, and Conn metrics. Note that MSE calculated in the whole image sometimes cannot reflect true matting performance. As shown in Fig. 9, our approach can handle both solid and transparent objects well. Further, let us consider a scenario where there is an image whose potential target foreground objects are non-salient or occluded with other equally-conspicuous objects in variegated backgrounds. Under this situation, with an RGB image as input, common trimap-free matting approaches [20], [21], [25] would struggle which object should be extracted or be adversely affected by inaccurate prior input. However, our JI strategy smartly decomposes matting into two stages and circumvents these limitations.

Ablation Study As shown in Table VI, we investigate the functionality of our Refinement Module in JIS and JIH on Composition-1k and Distinctions-646 test sets. The results show that RM can effectively improve SAD and Conn metrics. Besides, to research how robust our trimap-free JI pipeline to different segmentation inputs, we generate segmentation input by erosion on *initial segmentation* with 20×20, 30×30, 40×40, and 50×50 kernel sizes and followed by a Gaussian Blur. The corresponding evaluation results are illustrated in Table VIII. It is noted that the deviation of each metric is quite slight. Most metrics in their worst performance still set the state-of-the-art record, which reveals that our network can still remain robust when coarse segmentation is in an ill-posed condition.

Real Data Adaptation For comparison, we composite matting results to black background and compare our approach with Context-Aware Matting. For Context-Aware Matting, we use I and its corresponding $\widetilde{T}$ as input. Then, we conduct user study for evaluation. Each user was presented with one web page, showing 100 pairs of input image, ours, and CAM with random order of the last two. The participants are asked to rate the left composite relative to the right on three scales, i.e., better, similar, and worse. We survey more than 10 users and our method achieves 66.3% better, 29.7% similar but only 4.0% worse than CAM. The user study results indicate the competitive performance of our trimap-free pipeline compared to modest trimap-needed matting on real-world human images with diverse backgrounds. The Fig. 12 shows visual comparison. This experiment validates that our approach can be adapted into different domains of coarse segmentation; hence, we are confident of practical application of our pipeline. It is noted that our TGN can, to some extent, correct inaccurate coarse segmentation and our pipeline can be further improved by using stronger segmentation models or finetuning segmentation models on target domain for better coarse segmentation generation.

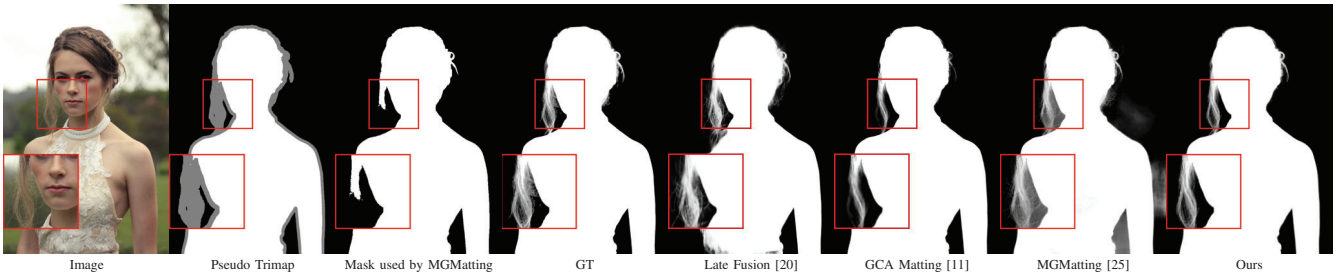

Fig. 13. Visual comparison on real-world portrait-636 benchmark.

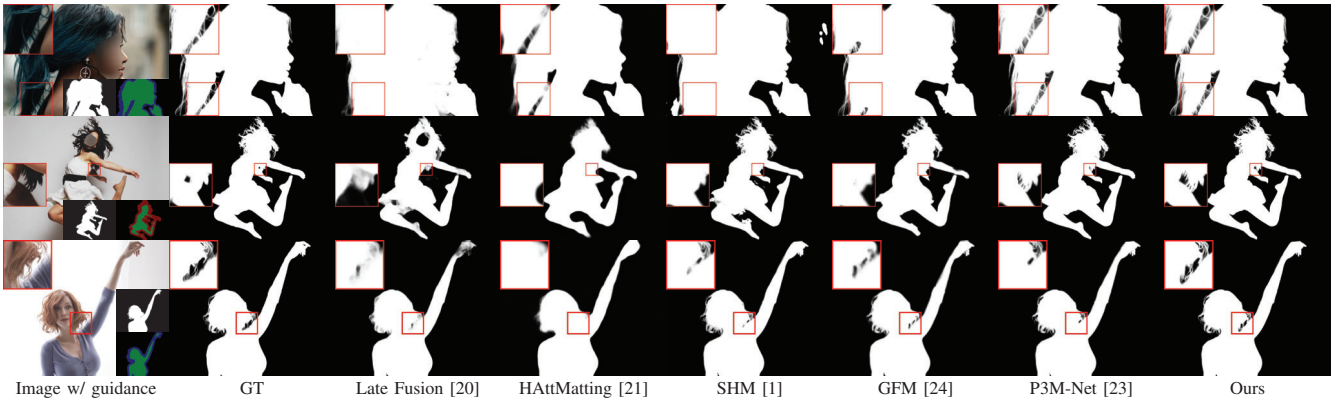

Fig. 14. Visual comparison on P3M-500-P and P3M-500-NP. The guidance from left/top to right/bottom is Mask used by TGN and TGN Trimap. Please zoom in for details.

TABLE XI
RESULTS ON P3M-500-P AND P3M-500-NP DATASETS IN TRIMAP-FREE SETTING.

Methods	P3M-500-P						P3M-500-NP					
	SAD↓	MSE↓	MAD↓	SAD-T↓	MSE-T↓	MAD-T↓	SAD↓	MSE↓	MAD↓	SAD-T↓	MSE-T↓	MAD-T↓
Late Fusion [20]	42.95	0.0191	0.0250	12.43	0.0421	0.0824	32.59	0.0131	0.0188	14.53	0.0420	0.0825
HAttMatting [21]	25.99	0.0054	0.0152	11.03	0.0377	0.0752	30.53	0.0072	0.0176	13.48	0.0403	0.0803
SHM [1]	21.56	0.0100	0.0125	9.14	0.0255	0.0545	20.77	0.0093	0.0122	9.14	0.0255	0.0545
GFM [24]	13.20	0.0050	0.0080	8.84	0.0269	0.0616	15.50	0.0056	0.0091	10.16	0.0268	0.0620
P3M-Net [23]	8.73	**0.0026**	0.0051	6.89	0.0193	0.0478	11.23	0.0035	0.0065	7.65	0.0173	0.0466
Ours	**8.42**	0.0029	**0.0049**	**6.09**	**0.0183**	**0.0414**	10.15	0.0035	0.0059	**6.75**	0.0172	**0.0407**
Ours*	8.43	0.0028	**0.0049**	6.20	**0.0183**	0.0424	**10.03**	**0.0033**	**0.0058**	6.86	**0.0166**	0.0409

TABLE XII
RESULTS ON AIM-500 BENCHMARK IN TRIMAP-FREE SETTING. RESULTS WITH * ARE FROM [22].

Methods	SAD↓	SAD-T↓	SAD-SO↓	SAD-STM↓	SAD-NS↓	SAD-Avg↓
Late Fusion* [20]	191.74	78.13	177.98	220.22	331.34	243.18
HAttMatting* [21]	479.17	114.23	509.75	338.11	270.07	372.64
SHM* [1]	170.44	69.41	154.56	204.67	329.9	229.71
SHM [1]	155.60	50.34	146.97	183.15	230.63	186.92
BSHM [2]	75.47	39.21	62.99	145.07	145.09	117.72
MGMatting [25]	60.67	30.26	49.82	101.44	146.94	99.40
GFM* [24]	52.66	37.43	35.45	123.15	181.90	113.50
AIM-Net* [22]	43.92	30.74	31.80	94.02	**134.31**	86.71
Ours	41.04	28.18	28.80	85.53	140.28	84.87
Ours*	**40.97**	**28.04**	**28.98**	**84.68**	137.98	**83.88**

Real-World Portrait-636 Benchmark (RWP-636) We compare our method[2] with trimap-needed Deep Image Matting [7], GCA Matting [11], IndexNet Matting [9], and Context-Aware Matting [41], trimap-free Late Fusion [20], and MGMatting [25]. We use pseudo trimaps as extra inputs for trimap-needed methods and our method, and foreground masks for MGMatting. We present results in Table X under two settings, Whole Image and Details. The results of compared methods are obtained through official pretrained models/inference demos. Most of them are trained on AIM training set (including transparent objects). However, Late Fusion is trained on AIM training set and an additional portrait dataset; MGMatting* is trained on the subset of AIM training set without transparent objects by applying re-JEPGing, gaussian blur, and gaussian noises to bridge the gap between synthetic and real-world images. It is noted that our method has a superior generalization ability in comparison with other methods and is on par with and sometimes even a bit better than MGMatting* in SAD of Details. We also show visual results in Fig. 13.

Privacy-Preserving Portrait Matting Benchmark (P3M-10k) In Table XI, we compare our method[2] with Late Fusion [20], HAttMatting [21], SHM [1], GFM [24], and P3M-Net [23] on P3M-500-P and P3M-500-NP under the PPT

[2]Our method with * means using SPAMattNet*. Please refer to Section IV-D2 for * meaning.

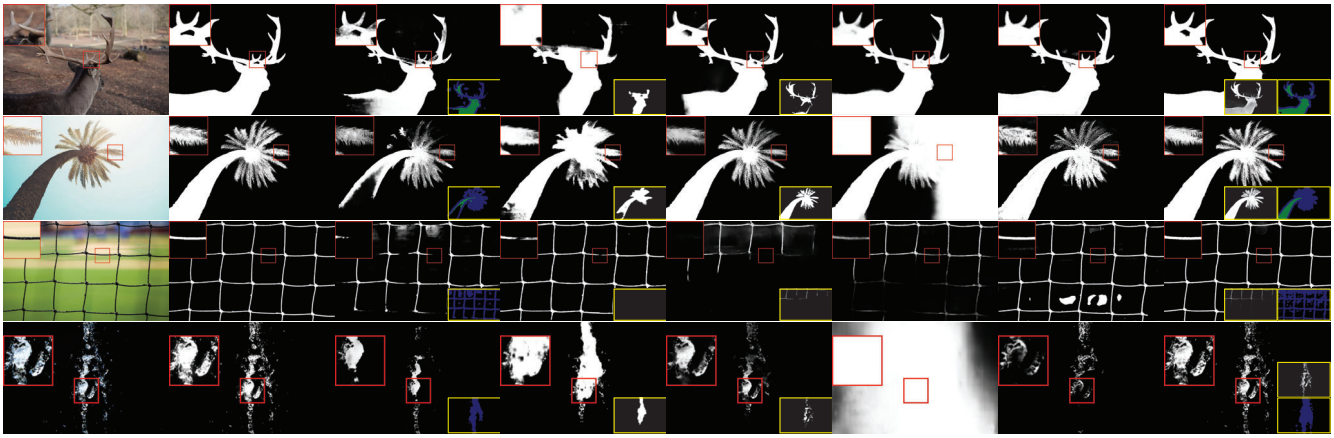

Fig. 15. Visual comparison on AIM-500 benchmark. From top to bottom, SO, STM, NS, and NS types. The guidance input (if has) is located at the bottom-right of each image, where the guidance of Ours from left/top to right/bottom is Mask used by TGN and TGN Trimap. Please zoom in for more details.

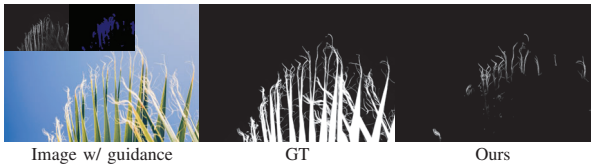

Fig. 16. A failure case of our method on NS type of AIM-500 benchmark.

setting. The results show that our method is competitive with other state-of-the-art trimap-free matting methods. Although our MSE on P3M-500-P is a bit deficient compared to P3M-Net, which is not a perfect metric for trimap-free matting, but all other metrics of ours are better than P3M-Net (SAD from 8.73 to 8.42/8.43 and SAD-T from 6.89 to 6.09/6.20). Our results on P3M-500-NP shows that our method has better generalization ability than compared methods (SAD from 11.23 to 10.15/10.03 and SAD-T from 7.65 to 6.75/6.86 compared to P3M-Net). Further, despite following the same sequential nature as SHM, our method performs much better than SHM. As shown in Fig. 14, our method can produce much sharper and clearer matting results in transition region. Interestingly, from Ours vs. Ours*, SPAMattNet* that shows much better improvement on synthetic data does not demonstrate equivalent improvement when training on real-world images. This phenomenon is also validated by trimap-based P3M-10k benchmark [23].

Automatic Image Matting Benchmark (AIM-500) In Table XII, we conduct quantitative comparison with SHM [1], Late Fusion [20], BSHM [2], HAttMatting [21], MGMatting [25], GFM [24], and AIM-Net [22]. We reimplement closed-source SHM and BSHM according to papers. The results show that our method[2] has superior performance in SAD of whole image/transition region and average SAD of three types on AIM-500 benchmark. As shown in Fig. 15, we present visual comparison where visual results of compared methods are obtained through official pretrained models/inference demos/our reimplementations. Our method can not only produce better "trimap" guidance than SHM/BSHM but also correct flaws of coarse masks to better capture target object shape unlike MGMatting, which leads to superior matting results compared to other methods. The reason why our SAD-NS is a bit inferior than AIM-Net is that there are several failure cases of our method whose distribution is the minority of the combined training set. We show one failure case in Fig. 16.

V. Limitations and Conclusion

Our main limitations can be summarized as follows: (1) a multi-stage systematic trimap-free matting approach usually hinders the real-time matting application; (2) with the progress of attention mechanism and transformer in matting task, it would be beneficial to further tackle attention efficiency issue in matting; (3) since our SPA module involves traditional sampling-based matting computation of foreground/background pixel pairs, we look forward to a more efficient deep learning representation of this in future.

In summary, we propose a novel two-stage trimap-free image matting framework. With an RGB image and its coarse foreground segmentation as input, our framework can estimate high-quality trimaps, leading to high-quality alpha mattes. Benefiting from semantic information provided by coarse foreground segmentation, our approach employs Trimap Generation Network (TGN) to roughly locate target objects and define transition regions. Then, our Sampling Propagation Attention Matting Network (SPAMattNet) focuses on transition regions under TGN trimap guidance. And our matting framework can be easily integrated with other semantic segmentation/salient object detection/matting methods to boost trimap-free matting in the real world. Comprehensive experiments indicate that our approach is comparable with or better than other state-of-the-art methods in either trimap-needed or trimap-free settings on several popular matting benchmarks, including Adobe Image Matting, alphamatting.com, Distinctions-646, Real-World Portrait-636, P3M-10k, and AIM-500 benchmarks.

VI. Acknowledgments

Thank Zhixiang Wang, Junjie Hu, and Kangfu Mei for their comments. Thank anonymous reviewers for their valuable

comments. Thank Issam Hadj Laradji for discussion.

REFERENCES

[1] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 618–626.

[2] J. Liu, Y. Yao, W. Hou, M. Cui, X. Xie, C. Zhang, and X.-s. Hua, "Boosting semantic human matting with coarse annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8563–8572.

[3] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2291–2300.

[4] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8762–8771.

[5] Y. Xu, B. Liu, Y. Quan, and H. Ji, "Unsupervised deep background matting using deep matte prior," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[6] D. Cho, Y.-W. Tai, and I. Kweon, "Natural image matting using deep convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 626–643.

[7] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2970–2979.

[8] S. Lutz, K. Amplianitis, and A. Smolic, "Alphagan: Generative adversarial networks for natural image matting," 2018.

[9] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3266–3275.

[10] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, "Learning-based sampling for natural image matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3055–3063.

[11] Y. Li and H. Lu, "Natural image matting via guided contextual attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 450–11 457.

[12] F. Zhou, Y. Tian, and Z. Qi, "Attention transfer network for nature image matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2192–2205, 2020.

[13] M. Forte and F. Pitié, "f, b, alpha matting," *arXiv preprint arXiv:2003.07711*, 2020.

[14] H. Yu, N. Xu, Z. Huang, Y. Zhou, and H. Shi, "High-resolution deep image matting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3217–3224.

[15] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 120–11 129.

[16] Y. Dai, H. Lu, and C. Shen, "Learning affinity-aware upsampling for deep image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6841–6850.

[17] Q. Liu, H. Xie, S. Zhang, B. Zhong, and R. Ji, "Long-range feature propagating for natural image matting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 526–534.

[18] G. Park, S. Son, J. Yoo, S. Kim, and N. Kwak, "Matteformer: Transformer-based image matting via prior-tokens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 696–11 706.

[19] Y. Zhou, I. H. Laradji, L. Zhou, and D. Nowrouzezahrai, "Osm: An open set matting framework with ood detection and few-shot learning," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/0092.pdf

[20] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu, "A late fusion cnn for digital matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7469–7478.

[21] Y. Qiao, Y. Liu, X. Yang, D. Zhou, M. Xu, Q. Zhang, and X. Wei, "Attention-guided hierarchical structure aggregation for image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 676–13 685.

[22] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 800–806, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/111

[23] J. Li, S. Ma, J. Zhang, and D. Tao, "Privacy-preserving portrait matting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3501–3509.

[24] J. Li, J. Zhang, S. J. Maybank, and D. Tao, "Bridging composite and real: towards end-to-end deep image matting," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 246–266, 2022.

[25] Q. Yu, J. Zhang, H. Zhang, Y. Wang, Z. Lin, N. Xu, Y. Bai, and A. Yuille, "Mask guided matting via progressive refinement network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1154–1163.

[26] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 364–373.

[27] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

[28] M. Jin, B.-K. Kim, and W.-J. Song, "Adaptive propagation-based color-sampling for alpha matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 7, pp. 1101–1110, 2014.

[29] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A bayesian approach to digital matting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2. IEEE, 2001, pp. II–II.

[30] J. Wang and M. F. Cohen, "Optimized color sampling for robust matting," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[31] E. S. Gastal and M. M. Oliveira, "Shared sampling for real-time alpha matting," in *Computer Graphics Forum*, vol. 29, no. 2. Wiley Online Library, 2010, pp. 575–584.

[32] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, "A global sampling method for alpha matting," in *CVPR 2011*. IEEE, 2011, pp. 2049–2056.

[33] E. Shahrian, D. Rajan, B. Price, and S. Cohen, "Improving image matting using comprehensive sampling sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 636–643.

[34] Y. Aksoy, T. Ozan Aydin, and M. Pollefeys, "Designing effective inter-pixel information flow for natural image matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 29–37.

[35] C. Xiao, M. Liu, D. Xiao, Z. Dong, and K.-L. Ma, "Fast closed-form matting using a hierarchical data structure," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 49–62, 2013.

[36] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.

[37] P. Lee and Y. Wu, "Nonlocal matting," in *CVPR 2011*. IEEE, 2011, pp. 2193–2200.

[38] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 228–242, 2007.

[39] A. Levin, A. Rav-Acha, and D. Lischinski, "Spectral matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 10, pp. 1699–1712, 2008.

[40] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum, "Poisson matting," in *ACM SIGGRAPH 2004 Papers*, 2004, pp. 315–321.

[41] Q. Hou and F. Liu, "Context-aware image matting for simultaneous foreground and alpha estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4130–4139.

[42] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1140–1147.

[43] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[44] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[45] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
[49] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 522–28 535, 2021.
[50] O. Wang, J. Finger, Q. Yang, J. Davis, and R. Yang, "Automatic natural video matting with depth," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*. IEEE, 2007, pp. 469–472.
[51] C.-L. Hsieh and M.-S. Lee, "Automatic trimap generation for digital image matting," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–5.
[52] S. Singh, A. S. Jalal, and C. Bhatanagar, "Automatic trimap and alpha-matte generation for digital image matting," in *2013 Sixth International Conference on Contemporary Computing (IC3)*. IEEE, 2013, pp. 202–208.
[53] V. Gupta and S. Raman, "Automatic trimap generation for image matting," in *2016 International Conference on Signal and Information Processing (IConSIP)*. IEEE, 2016, pp. 1–5.
[54] Z. Chen, Y. Zheng, X. Li, R. Luo, W. Jia, J. Lian, and C. Li, "Interactive trimap generation for digital matting based on single-sample learning," *Electronics*, vol. 9, no. 4, p. 659, 2020.
[55] A. Al-Kabbany and E. Dubois, "A novel framework for automatic trimap generation using the gestalt laws of grouping," in *Visual Information Processing and Communication VI*, vol. 9410. International Society for Optics and Photonics, 2015, p. 94100G.
[56] D. Cho, S. Kim, Y.-W. Tai, and I. S. Kweon, "Automatic trimap generation and consistent matting for light-field images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1504–1517, 2016.
[57] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *European conference on computer vision*. Springer, 2016, pp. 92–107.
[58] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[61] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4471–4480.
[62] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.
[63] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 684–690.
[64] M. Amirul Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7142–7150.
[65] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
[66] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.
[67] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3723–3732.
[68] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[69] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1826–1833.
[70] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.
[71] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[72] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 770–785.
[73] H. Lu, Y. Dai, C. Shen, and S. Xu, "Index networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 242–255, 2020.
[74] S. Cai, X. Zhang, H. Fan, H. Huang, J. Liu, J. Liu, J. Liu, J. Wang, and J. Sun, "Disentangled image matting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8819–8828.
[75] Supervisely person dataset, https://supervise.ly/explore/projects/supervisely-person-dataset-23304/datasets.
[76] H. Cai, F. Xue, L. Xu, and L. Guo, "Transmatting: Enhancing transparent objects matting with transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 253–269.

Yuhongze Zhou received the B.Eng. degree in Digital Media Technology from Zhejiang University, Hangzhou, China, in 2020, and she is pursuing the M.Sc. degree in Computer Science at McGill University, Montréal, Canada.

Her research interests include computational photography, machine learning, and computer graphics.

Liguang Zhou received the B.Eng. degree in Electrical Engineering from China Jiliang University, Hangzhou, China, in 2016, and he is pursuing the Ph.D. degree in Computer Information Engineering at The Chinese of Hong Kong, Shenzhen, China.

His research interests include robotic vision, scene understanding, and computational photography.

Tin Lun Lam received the B.Eng. and Ph.D. degrees in automation and computer-aided engineering from Chinese University of Hong Kong, Hong Kong, in 2006 and 2010, respectively.

He is an Assistant Professor at the School of Science Engineering, The Chinese University of Hong Kong. His research interests include multi-robot systems, field robotics, and human-robot interaction.

Yangsheng Xu received the B.S.E. and M.S.E. degrees from Zhejiang University, Hangzhou, China, in 1982 and 1984, respectively, and the Ph.D. degree in 1989 from University of Pennsylvania, Philadelphia, PA, USA, all in robotics.

He is currently President and Professor of The Chinese University of Hong Kong, Shenzhen, China. His research interests include robotics, intelligent systems, and electric vehicles.