Enhancing Transparent Object Matting Using Predicted Definite Foreground and Background

Yihui Liang, Qian Fu, Kun Zou, Guisong Liu, *Member, IEEE*, and Han Huang, *Senior Member, IEEE*

***Abstract*— Natural image matting is a widely used image processing technique that extracts foreground by predicting the alpha values of the unknown region based on the alpha values of the known foreground and background regions. However, existing image matting methods may not yield the most optimal results when applied to images containing transparent objects because the known foreground region is small or even absent. To address this shortcoming, in this paper, we propose a novel method named Transparent Object Matting using Predicted Definite Foreground and Background (TOM-PDFB), which can explore and utilize the definite foreground and background in the unknown region. For this purpose, a newly developed foreground-background confidence estimator is applied to predict the confidence level of the definite foreground and the definite background, thus providing the priors required for transparent object matting. Next, foreground-background guided progressive refinement network developed as a part of this work is adopted to incorporate the estimated definite foreground and background into the alpha matte refinement process. Extensive experimental results demonstrate that the TOM-PDFB outperforms state-of-the-art methods when applied to transparent objects. Project page: https://github.com/yihuiliang/TOM-PDFB.**

Manuscript received 24 August 2023; revised 4 March 2024 and 11 June 2024; accepted 22 August 2024. Date of publication 30 August 2024; date of current version 30 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62002053, Grant 62276103, and Grant 62376228; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515111082, Grant 2020A1515110504, and Grant 2023A1515010066; in part by the Natural Science Foundation of Guangdong Province under Grant 2020A1515010696; in part by the Innovation Team Project of General Colleges and Universities in Guangdong Province under Grant 2023KCXTD002; in part by the Science and Technology Foundation of Guangdong Province under Grant 2021A0101180005; in part by Zhongshan Science and Technology Research Project of Social Welfare under Grant 2021B2006 and Grant 2020B2017; and in part by the Major Science and Technology Foundation of Zhongshan City under Grant 2019B2009, Grant 2019A40027, and Grant 2021A1003. This article was recommended by Associate Editor F. M. Zhu. *(Yihui Liang and Qian Fu contributed equally to this work.) (Corresponding authors: Han Huang; Guisong Liu.)*

Yihui Liang and Kun Zou are with the School of Computer Science, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528400, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

Qian Fu is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

Guisong Liu is with the Complex Laboratory of New Finance and Economics, School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu 611130, China (e-mail: gliu@swufe.edu.cn).

Han Huang is with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China, also with the Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China, Guangzhou 510006, China, and also with the Guangdong Engineering Center for Large Model and GenAI Technology, Guangzhou 510006, China (e-mail: hhan@scut.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3452512

***Index Terms*— Image matting, definite foreground, transparent object matting.**

I. Introduction

IMAGE matting is a critical task within the field of computer vision, with a wide range of applications in image or video editing, compositing, and film post-production [1]. It involves accurately separating the foreground from the background in an image by estimating the opacity of the foreground, which is also known as alpha matte. For this purpose, a natural image is represented as a convex combination of the foreground and the background component, allowing the image matting problem to be defined mathematically as follows:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \alpha_i \in [0, 1] \tag{1}$$

where I_i, F_i, and B_i represent the color values of the input image, the foreground, and the background at pixel i, respectively, and α_i denotes the opacity of the foreground at pixel i. However, as only the value of I_i is known, the image matting is an ill-posed problem. To address this issue, trimap is provided as an additional input to impose further constraints on the image matting problem. This is achieved because a trimap divides the image into three non-overlapping regions, namely the known foreground region (where the alpha value is known to be 1), the known background region (where the alpha value is known to be 0), and the unknown region (where the alpha value is unknown and needs to be determined). The unknown region comprises a small portion of definite foreground and definite background, along with semi-transparent regions.

Predicting the alpha matte for transparent objects, such as glass, plastic bags, fog, water drops, etc. is a challenge for the standard image matting methods due to the corresponding trimaps will consist of very few or even no known foreground region [2]. As in most images with transparent elements, only a small number of pixels are associated with definite foreground, which may be scattered across the image, identifying definite foreground when generating a trimap is difficult. Moreover, as shown in Figure 1(a) and Figure 1(b), the known foreground region may not be helpful in estimating the alpha matte of transparent objects. However, transparent object matting is crucial, not only because transparent objects frequently feature in images, but also because enhancing the matting performance of transparent objects can benefit non-transparent objects as well [2].

Traditional matting methods include sampling-based methods which sample the known region and estimate each alpha value by selecting a foreground color and a background color from the sample set [3], [4], [5], and propagation-based

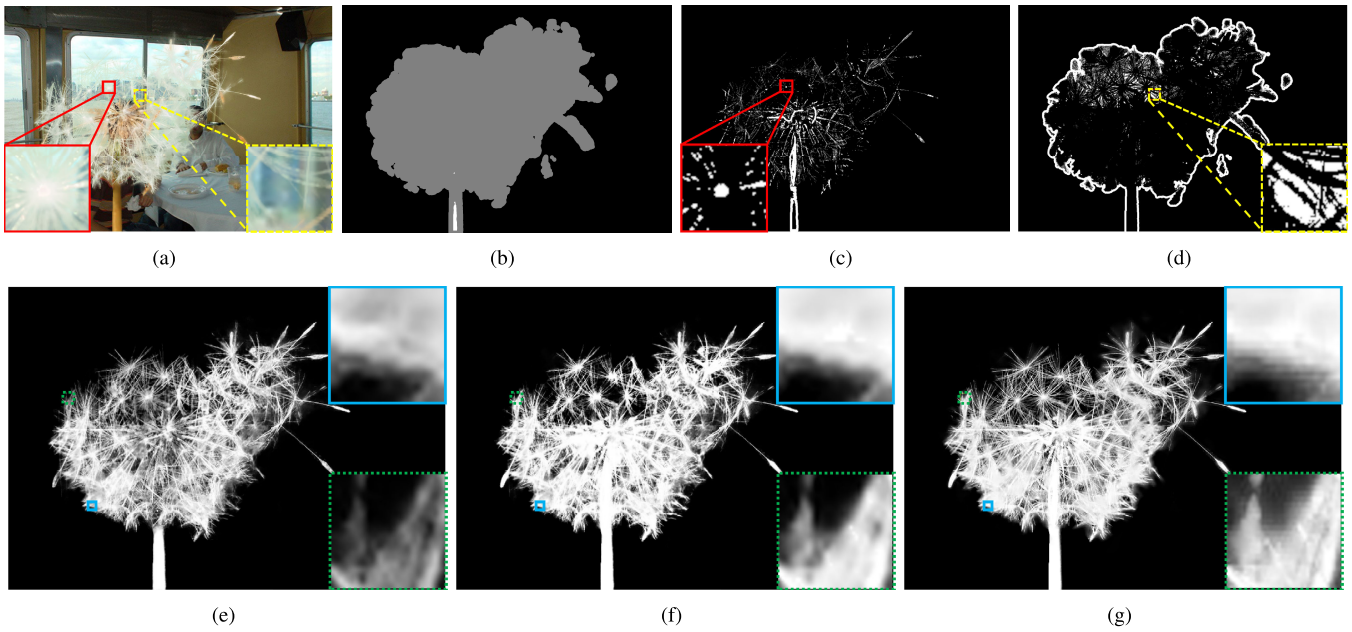


Fig. 1. An example showcasing the comparative results of transparent object matting using both the trimap and the trimap integrated with definite foreground and background from the unknown region: (a) the input image; (b) the trimap; (c) the definite foreground in the unknown region; (d) the definite background in the unknown region; (e) the resulting alpha matte by using (b); (f) the resulting alpha matte by using (b) incorporated with (c) and (d); and (g) the ground-truth alpha matte.

methods which propagate the alpha values from the known region to the unknown region [6], [7], [8], [9]. As such approaches rely on low-level features such as color and position [3], [4], [6], they are unsuitable for transparent image matting, where the background has a substantial influence on the appearance of transparent objects, leading to pronounced color similarities. To overcome this issue, deep learning matting methods are increasingly being used due to their capability to automatically discover representations from extensive datasets. The currently available deep learning matting methods can be categorized into trimap-free and trimap-based methods. As their name suggests, trimap-free methods utilize alternative information to that typically provided by a trimap to identify the foreground to be extracted [10], [11], [12], or extract specific foreground objects based on a single image [13], [14], [15], [16]. However, due to the absence of image-specific definite foreground and definite background information, the matting results are typically inadequate. In contrast, trimap-based deep learning methods utilize information from the known regions to predict the alpha values of the unknown region. Neural networks, including convolutional neural networks (CNNs), are typically used for this purpose [1], [17], [18], [19], [20], [21] even though they do not yield satisfactory results when applied to images containing transparent objects. This issue arises because a significant portion of a transparent object is usually located in the unknown region and numerous pixels comprising such objects are situated at a significant distance from the known regions. Thus, owing to the small effective reception fields of neural networks, many pixels in the unknown region cannot be correlated to the distant known features [2]. Recently, given the remarkable success of the transformer model in natural language processing (NLP), there has been a surge of efforts to incorporate Transformers into visual tasks [22], [23], [24]. Vision transformers exhibit distinct advantages over CNNs in capturing global image information through self-attention mechanisms. This has led to endeavors to integrate vision transformers into image matting tasks [2], [25]. These transformer-based methods have been successfully used to guide the model in evaluating the similarity between known and unknown regions to generate self-attention, thus aiding the alpha matte prediction. In this context, MatteFormer [25] is particularly noteworthy, as it incorporates the mean features from all three trimap regions (foreground, background, and unknown) as new queries in local windows to evaluate similarity. On the other hand, in TransMatte [2] the trimap is redesigned as a tri-token map, aiding the model in distinguishing the known and unknown regions during the assessment of similarity and relatedness. In other words, unlike the CNN-based methods, these transformer-based approaches enable unknown-region pixels to capture similar long-range features from known regions, enhancing the transparent objects matting performance. Nevertheless, the limited known foreground and subtle similarity between regions pose a challenge, limiting the full potential of self-attention mechanisms. This obstacle persists as a challenge for these transformer-based methods in matting transparent objects.

The definite foreground and background in the unknown region play an important role in transparent object matting. Figure 1 compares the alpha mattes obtained by popular image matting methods on the original trimap and on the trimap that includes the definite foreground and background in the unknown region. The improvement of the alpha matte quality demonstrates that the definite foreground and background in the unknown region significantly contribute to transparent object matting.

This observation has motivated the present study, as a part of which a novel method named Transparent Object Matting Using Predicted Definite Foreground and Background (TOM-PDFB) was developed to utilize the definite foreground and background within the unknown region. TOM-PDFB introduces two novel modules: the Foreground-Background Confidence Estimator (FBCE) and the Foreground-Background Guided Progressive Refinement Network (FB-PRN). The FBCE predicts the foreground and background confidence maps which provide the confidence levels that the pixel in the unknown region is the definite foreground pixel and definite background pixel, respectively. FB-PRN predicts foreground and background at multiple resolutions according to the information of the definite foreground and definite background estimated in the previous decoder layers, aiming to progressively refine the alpha matte. Thus, the capabilities of TOM-PDFB result in a high-quality alpha matte for transparent objects.

Accordingly, the contributions of this work can be summarized as follows:

- We show that the definite foreground and definite background in the unknown region play an important role in transparent object matting.
- We present a novel method named TOM-PDFB that predicts and exploits the definite foreground and background within the unknown region. TOM-PDFB calculates the confidence level of a pixel in the unknown region belonging to the definite foreground or background and use it to estimate and refine the alpha mattes.
- The experimental results on the Composition-1k and Transparent-460 datasets demonstrate that TOM-PDFB achieves excellent performance in the transparent object matting task and outperforms the state-of-the-art image matting methods.

The paper is structured as follows: In Section II, we review related work, followed by the description of our Methodology in Section III. We then present and discuss the experimental results in Section IV and finally, conclude the paper in Section V.

II. Related Work

A. Traditional Matting

Traditional matting methods, based on the linear combination equation shown in Eq. 1, can be classified into sampling-based and propagation-based methods. The sampling-based methods sample pixels from the known foreground and known background regions to find candidate foreground and background color pairs for each pixel in the unknown region. In addition, they use a metric to determine the best foreground and background combination to estimate the alpha value [3], [4], [5]. On the other hand, the propagation-based methods, also known as affinity-based methods, estimate the alpha matte by propagating alpha values from the known foreground and background regions to the unknown region based on the affinities or similarities between pixels [6], [7], [8], [9].

As these methods heavily rely on the known foreground and background low-level features such as color and position, they cannot accurately matte transparent objects characterized by limited known definite foreground regions. Moreover, as transparent objects are typically semi-transparent or highly transparent, the background has a significant impact on their appearance, resulting in high color similarity. This similarity in low-level features makes it difficult for traditional methods to effectively distinguish transparent objects from the background.

B. Deep Learning Matting

In recent years, deep learning matting methods have made remarkable progress by effectively leveraging semantic information contained in the image, which enables a better understanding of transparent object representation compared to that attained via traditional methods.

Convolutional neural networks (CNNs) play a crucial role in this context, as they utilize semantic features from the known regions to predict the alpha values of the unknown regions. For instance, [1] released the Composition-1K dataset and introduced the DIM, a two-stage architecture for directly predicting alpha mattes. On the other hand, SampleNet [26] uses the known foreground and background information to supervise the network and improve prediction accuracy. IndexNet [19] leverages index-guided methods for up-sampling and down-sampling to enhance the prediction details. In contrast, GCA-Matting [17] incorporates a Guided Contextual Attention module that uses the trimap as a guide to propagate high-level opacity information using learned low-level affinity. TIMI-Net [18] fuses global information from the RGB image and trimap to improve the accuracy of alpha mattes. In this context, it is also worth noting methods that aim to improve their generalizability to coarse trimaps. For example, the strategies proposed in [27] and [28] incorporate trimap adaptation as an auxiliary task coupled to the matting network, with the goal of enhancing the semantic extraction of the coarse trimap by the matting network. Recent approaches, on the other hand, incorporate additional supervisory information. For example, ContextNet [29] employs two encoder networks to extract local features and global context information. It simultaneously estimates both the foreground and alpha matte. TOM-Net [30], [31] is designed for environmental matting task that aims to matting specific transparent objects with refractive and reflective properties. It utilizes the environmental matting model to estimate the object mask, the attenuation mask, and the refractive flow field for an input image [32]. Due to the difference between the environmental matting and the natural image matting tasks, the output produced by TOM-Net is significantly different from natural image matting methods. Therefore, TOM-Net is not suitable for natural image matting. FBAMatting [20] predicted the foreground, the background, and the alpha matte simultaneously, and uses a foreground and background fusion loss that acts as a constraint on the prediction results. The experimental results reported in the literature demonstrate its ability to extract transparent objects. Some methods aim to estimate the alpha matte of various foreground objects, including transparent objects, SIM [21] clustered 20 different matting classes

and introduced a semantic trimap that consists of confidence maps for each matting class. However, the literature [33], [34], [35], [36] indicates that CNNs analyze salient regions within an image on an individual basis, which overlooks the relationships between these regions and can compromise the integrity of the object as a whole. These methods do not utilize the image features outside their reception fields due to which, accordingly [37], more than 50% of pixels in the unknown regions cannot be correlated to pixels in the known regions in the range of the effective reception fields of neural networks. This limitation makes accurately predicting the alpha matte of transparent objects difficult, especially in cases where unknown regions comprise a significant portion of the trimap.

To mitigate these limitations, building upon the success of ViT [22] and related works [23], [24], [38], [39], researchers have started applying transformers to typical vision tasks. Vision transformers are characterized by powerful self-attention mechanisms that are particularly valuable for capturing global attention across the entire image, thus offering advantages over CNNs. Transformer-based matting methods have been shown capable of learning global semantic features and capturing distant known information to guide the model in evaluating the similarity between known and unknown regions to generate attention for alpha matte prediction [2], [25]. For instance, Park [25] proposed a Prior-Attentive Swin-Transformer block that incorporates global mean features which from the known foreground, known background, and unknown regions as additional queries into the self-attention mechanism in order to generate attention by evaluating similarity. Hu [40] established the correlation between unknown and known regions by providing global context features to the local window self-attention. Cai [2] introduced the Transparent-460 dataset, a high-resolution matting dataset focused on transparent objects, and presented TransMatte, a transformer-based approach. These researchers redesigned the trimap as a tri-token map, effectively assisting the model in distinguishing features from known and unknown regions during the assessment of similarity by the proposed self-attention mechanism. The results obtained by TransMatte [2] demonstrate the superiority of transformer-based methods over CNN-based alternatives in transparent object matting tasks due to the ability of pixels in the unknown region to capture similar long-range features from the known regions. However, as most transparent objects consist of a limited known foreground and exhibit subtle similarity between known and unknown regions, this limitation presents challenges for these transformer-based methods.

Further advances have also been made with the aim of reducing the resource burden associated with drawing trimaps, giving rise to trimap-free approaches. Trimap-free methods can predict alpha mattes without using a trimap, instead relying on a background image [10], [11], scribble information [12], or coarse mask [41], [42]. Some of these methods only require a single RGB image as input [13], [14], [15], [16], [43]. However, their applicability is limited to specific types of images with opaque foregrounds [43]. Moreover, the accuracy of these trimap-free methods is still inferior to that of trimap-based methods [12], [21], [25], suggesting that the definite foreground and background information from the trimap is beneficial.

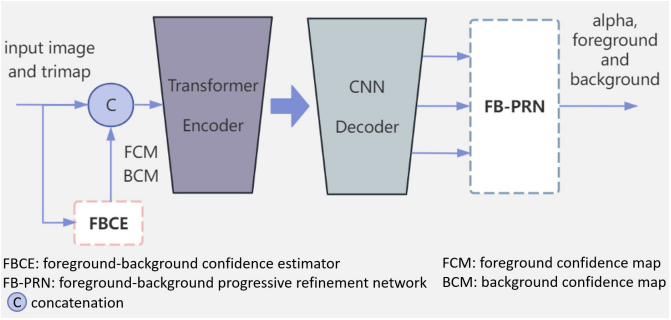


Fig. 2. Overall pipeline of transparent object matting using predicted definite foreground and background (TOM-PDFB).

III. Methodology

The core concept of TOM-PDFB lies in leveraging the definite foreground and background within the unknown region to improve the matting of transparent objects. As illustrated in Figure 2, TOM-PDFB consists of four main modules: Foreground-Background Confidence Estimator (FBCE), the transformer encoder, the CNN decoder, and foreground-Background Progressive Refinement Network(FB-PRN). FBCE estimates the confidence levels for definite foreground and background, producing a foreground confidence map (FCM) and a background confidence map (BCM), thus providing additional information of the unknown region for alpha estimations. The transformer encoder extracts features from the image and the confidence levels for definite foreground and background. The decoder generates alpha mattes, foreground image, and background image according to the features obtained by the transformer encoder. FB-PRN progressively refines the alpha mattes generated at different feature levels from the decoder to produce the final prediction.

The details of FBCE and FB-PRN are presented in subsection III-A and III-B, respectively. The transformer encoder leverages the Swin-Transformer architecture [23], which has been shown to yield promising results when applied in mainstream vision tasks with a hierarchical architecture design and the shifted window scheme [25]. The variant-ResNet decoder introduced in [41] is employed here, which involves convolution layers and upsample layers.

A. Foreground-Background Confidence Estimator

The Foreground-Background Confidence Estimator(FBCE) is designed to estimate confidence levels of the definite foreground and background in the unknown region of the trimap, which provides additional information on the unknown region, i.e. FCM and BCM, for the subsequent image matting modules. An example of FCM and BCM is given in Figure 4. FCM and BCM indicates the confidence levels at which pixels were determined to belong to the definite foreground or definite background, respectively. In 4, whiter colors indicate higher confidence levels.

The network architecture of FBCE is a encoder-decoder structure and the VGG-16 [44] is utilized as encoder due to

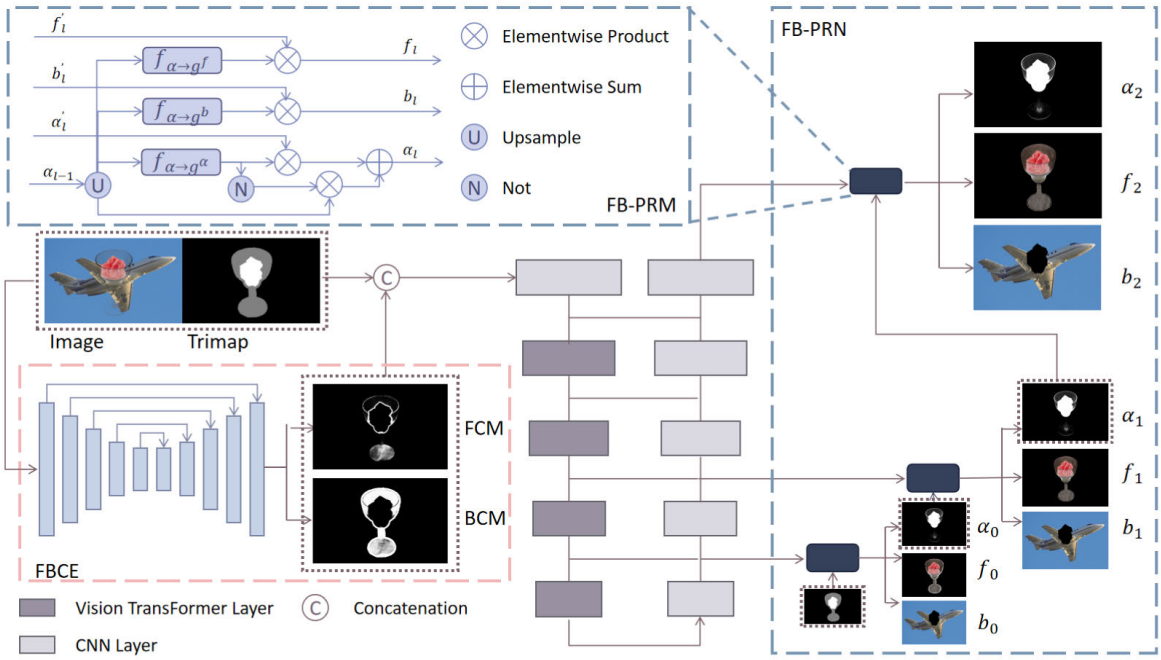


Fig. 3. Overall framework of our proposed TOM-PDFB. The foreground-background confidence estimator(FBCE) predicts the confidence level of definite foreground and definite background. The foreground-background progressive refinement network (FB-PRN) predicts the alpha matte, as well as the foreground and background, at multiple resolutions. The alpha matte obtained from the lower-resolution prediction is used as guidance for the subsequent prediction.

Fig. 4. An example of the foreground confidence map (FCM) and the background confidence map (BCM) predicted by the foreground-background confidence estimator (FBCE).

its simplicity and high effectiveness. The decoder of FBCE network consists of five pooling layers and convolutional layer groups. A foreground-background confidence loss function is designed to trained the network.

The foreground-background confidence loss function involves the foreground confidence loss term $\mathcal{L}_F$ and the background confidence loss term $\mathcal{L}_B$, which can be formulated as follows:

$$\mathcal{L}_{FBCE} = \mathcal{L}_F + \mathcal{L}_B, \tag{2}$$

where $\mathcal{L}_F$ and $\mathcal{L}_B$ measure the error of foreground confidence estimation and that of background confidence estimation, respectively.

Equation 3 provides the definition of $\mathcal{L}_F$.

$$\mathcal{L}_F = -\sum_i \mu_i c_i^f (\hat{x}_i \log p_i^f + (1-\hat{x}_i)\log(1-p_i^f)) \tag{3}$$

where p_i^f is the predicted confidence level of the definite foreground at pixel i. $\hat{x}_i$ is set to 1 for the pixel which alpha value is greater than 0 in unknown region of the trimap and is set to 0 for the pixel which alpha value is 0 in the unknown region of the trimap. c_i^f is a penalty factor that adaptively adjusts the degree of penalty according to the pixel type, which can be formulated as:

$$c_i^f = \begin{cases} \hat{\alpha}_i & 0 < \hat{\alpha}_i < 1 \\ 1 & otherwise \end{cases} \tag{4}$$

where $\hat{\alpha}_i$ denotes the ground-truth alpha value of pixel i. Considering that the goal of FBCE is to accurately predict the definite foreground and background, c_i^f is designed to severely penalize the error of definite foreground/background estimation. We note that semi-transparent pixels can also provide image matting information. When their alpha value is closer to 1, their features are similar to that of the definite foreground. Therefore, c_i^f also penalize the error in the semi-transparent pixel according to the value of $\hat{\alpha}_i$. When the pixel is a definite foreground/background pixel, the value of c_i^f is set to 1. When the pixel is a semi-transparent pixel, the value of c_i^f decreases with the value of the $\hat{\alpha}_i$, reducing the value of the foreground confidence loss term.

The μ_i in Equation 3 is the weight assigned in order to adapt the change of the number of different types of pixels in the unknown region of the trimap for the foreground confidence loss and can be defined as:

$$\mu_i = \begin{cases} clamp(\sqrt{\frac{|F|}{|S|}}), & i \in F \\ clamp(\sqrt{\frac{|S|}{|F|}}), & i \in S \end{cases} \tag{5}$$

$$\text{clamp}(\phi) = \min(\max(\phi, 0.1), 10) \tag{6}$$

where $|*|$ denotes the Cardinality of set $*$, F, B and S are three sets obtained by categorizing the pixels in the unknown region of the trimap into three distinct subsets. Pixels with an alpha value of 1 constitute the definite foreground pixel set, represented by the set F. Pixels with an alpha value between

0 and 1 constitute the semi-transparent pixel set, denoted as S. Pixels with an alpha value of 0 constitute the definite background pixel set, denoted as B.

When $|S|$ significantly surpasses $|F|$, the pixels in F, i.e. the definite foreground pixel, may provide limited local appearance information. In such instances, focusing on the pixels in the S set becomes crucial to identify pixels with appearances similar to the foreground. In this case, paying attention to the pixels in the S set may benefit model training. Conversely, when the number of pixels in F surpasses those in S, the pixel in the F set, i.e. the definite foreground pixel, can provide reliable and accurate appearance information. In this case, focusing on the pixel in the F set may allow the FBCE network to learn accurate foreground confidence features. Therefore, the value of μ_i is designed to changes according to the number of different types of pixels in the unknown region.

Similarly, the definition of the background confidence loss term $\mathcal{L}_B$ can be given by Eq. (7).

$$\mathcal{L}_B = -\sum_i \nu_i c_i^b (\hat{y}_i \log p_i^b + (1-\hat{y}_i)\log(1-p_i^b)), \quad (7)$$

where p_i^b is the predicted confidence level of the definite background, and $\hat{y}_i$ is set to 1 for the pixel which alpha value is less than 1 in unknown region of the trimap and is set to 0 for the pixel which alpha value is 1 in the unknown region of the trimap. c_i^b is a penalty factor that adaptively adjusts the degree of penalty according to the pixel type, which can be formulated as:

$$c_i^b = \begin{cases} 1-\hat{\alpha}_i & 0 < \hat{\alpha}_i < 1 \\ 1 & otherwise \end{cases} \quad (8)$$

Here, ν_i denotes the adaptive weight for the background confidence map prediction which is similar to μ_i and is obtained using the expressions given in Eq.9 below.

$$\nu_i = \begin{cases} clamp(\sqrt{\frac{|B|}{|S|}}), & i \in B \\ clamp(\sqrt{\frac{|S|}{|B|}}), & i \in S \end{cases} \quad (9)$$

B. Foreground-Background Progressive Refinement Network

The Foreground-Background Progressive Refinement Network (FB-PRN) is designed to leverage the definite foreground and background information from the image matting features provided by the CNN decoder. Its design is inspired by the Progressive Refinement Network (PRN) introduced by [41]. In contrast to PRN, FB-PRN specifically focuses on utilizing the definite foreground and background information to target the semi-transparent regions for improving the overall quality of the alpha matte for transparent objects.

As illustrated in Figure 3, The FB-PRN consists of multiple Foreground-Background Progressive Refinement Modules (FB-PRMs). The FB-PRMs are set to work on the decoder layers with output resolutions of 1/8, 1/4, and 1/1 of the matting network input resolution $H \times W$ [41], denoted as 1/8, 1/4, and 1/1 decoder layers, respectively. These decoder layers are referred to as the 0th, 1st, and 2nd layers of the FB-PRN, respectively.

Algorithm 1 Algorithm of Foreground-Background Progressive Refinement Network (FB-PRN)

Input: the height and width of the image: H, W, the foreground, background and alpha matte estimation outputs of the CNN decoder layers with output resolutions of 1/8, 1/4, and 1/1 of the input resolution: f_0', b_0', α_0', f_1', b_1', α_1', f_2', b_2', α_2', the ground-truth foreground: $\hat{f}$, the ground-truth background: $\hat{b}$, the ground-truth alpha matte: $\hat{\alpha}$.

Output: The predicted alpha matte:α_2, the loss function of the matting network:$\mathcal{L}_{total}$

$\mathcal{L}_{FB} \leftarrow 0$;
$\mathcal{L}_{\alpha} \leftarrow 0$;
$s_0 \leftarrow 8$, $s_1 \leftarrow 4$, $s_2 \leftarrow 1$; // initialize the scale factor s_l of the lth layer
for $l = 0$ *to* 2 **do**
  // Step 1: predict the alpha matte at the lth layer of FB-PRN
  if $l = 0$ **then**
    Initialize α_l^{base} ;
  else
    $\alpha_l^{base} \leftarrow \alpha_{l-1}$;
  $\alpha_l^{'} \leftarrow \mathrm{Up}(\alpha_l^{'}, (H, W))$;
  $g_l^{\alpha} \leftarrow f_{\alpha_l^{base} \rightarrow g_l^{\alpha}}$;
  $\alpha_l \leftarrow \alpha_l^{'} \odot g_l^{\alpha} + \alpha_l^{base} \odot (1 - g_l^{\alpha})$;
  // Step 2: calculate training loss for the lth layer of FB-PRN
  $w_l \leftarrow l + 1$
  $\mathcal{L}_{\alpha} \leftarrow \mathcal{L}_{\alpha} + w_l \mathcal{L}_{\alpha}^{(l)}(\hat{\alpha} \cdot g_l^{\alpha}, \alpha_l \cdot g_l^{\alpha})$;
  $\alpha_l^{base} \leftarrow \mathrm{Down}(\alpha_l^{base}, (\lfloor \frac{H}{s_l} \rfloor, \lfloor \frac{W}{s_l} \rfloor))$;
  $\hat{f}_l \leftarrow \mathrm{Down}(\hat{f}, (\lfloor \frac{H}{s_l} \rfloor, \lfloor \frac{W}{s_l} \rfloor))$;
  $\hat{b}_l \leftarrow \mathrm{Down}(\hat{b}, (\lfloor \frac{H}{s_l} \rfloor, \lfloor \frac{W}{s_l} \rfloor))$;
  $g_l^f \leftarrow f_{\alpha_l^{base} \rightarrow g_l^{\alpha}}$;
  $g_l^b \leftarrow f_{\alpha_l^{base} \rightarrow g_l^b}$;
  $\mathcal{L}_{FB} \leftarrow \mathcal{L}_{FB} + (\mathcal{L}_{l1}(\hat{f}_l \cdot g_l^f, f_l' \cdot g_l^f) + \mathcal{L}_{l1}(\hat{b}_l \cdot g_l^b, b_l' \cdot g_l^b))$;
$\mathcal{L}_{total} = \mathcal{L}_{\alpha} + \mathcal{L}_{FB}$;
return α_2, $\mathcal{L}_{total}$;

The FB-PRN procedure is described in Algorithm 1 and the symbols used in FB-PRN are defined in Table I. As shown in Algorithm 1, the decoder layers produce the provisional alpha matte α_l', the provisional foreground image f_l' and the provisional background image b_l' for each layer in FB-PRN. FB-PRM refines the alpha matte of at each layer according to the resulting α_l', f_l' and b_l', and calculates the loss for training. At the 0th layer of the FB-PRN, α_l^{base} is initialized by assigning a value of 1 to pixels corresponding to the known foreground region in the trimap, a value of 0 to pixels corresponding to the known background region, and a value of 0.5 to pixels corresponding to the unknown region. The

TABLE I
DEFINITIONS OF SYMBOLS USED IN ALGORITHM 1

Symbol	Description
l	The lth layer of the FB-PRN
H	The input image height
W	The input image width
s_l	The scale factor s_l of the lth layer
T	The input trimap
f_l'	The foreground estimation output of the CNN decoder layers at the lth layer
b_l'	The background estimation output of the CNN decoder layers at the lth layer
α_l'	The alpha matte estimation output of the CNN decoder layers at the lth layer
α_l^{base}	The initial α at the lth layer which is used for generating g_l^f, g_l^b, g_l^α and α_l
α_l	The prediction alpha matte of FB-PRN at the lth layer
g_l^f	The foreground guide mask at the lth layer
g_l^b	The background guide mask at the lth layer
g_l^α	The alpha self-guidance mask at the lth layer
$\hat{f}$	The ground-truth foreground
$\hat{b}$	The ground-truth background
$\hat{\alpha}$	The ground-truth alpha matte
$f_{\alpha_l^{base}\to g_l^f}$	The process for calculating the foreground guide mask, which defined in Eq. 13
$f_{\alpha_l^{base}\to g_l^b}$	The process for calculating the background guide mask, which defined in Eq.14
$f_{\alpha_l^{base}\to g_l^\alpha}$	The process for calculating the alpha self-guidance mask, which defined in Eq. 10
Up(X, size)	Upsampling operation on X to the specified size
Down(X, size)	Downsampling operation on X to the specified size
$\odot$	Element-wise multiplication
$\mathcal{L}_{l1}$	The regression loss, which shown in Eq. 11
$\mathcal{L}_\alpha^{(l)}$	The alpha loss at the lth layer of FB-PRN, which is shown in Eq. 12
$\mathcal{L}_{FB}$	The total loss of all layers for the foreground and background
$\mathcal{L}_\alpha$	The total loss of all layers for the alpha
w_l	The weight ratio w_l of the lth layer for alpha loss which is defined in [41]

provisional alpha matte α_l' is up-sampled to the size of $H \times W$ and alpha self-guidance mask g_l^α is calculated as follows [41]:

$$f_{\alpha_l^{base}\to g_l^\alpha}(x, y) = \begin{cases} 1, & \text{if } 0 < \alpha_{l-1}(x, y) < 1 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Subsequently, the estimated alpha matte α_l of the lth layer is generated as described in Line 12 of Algorithm 1 by using the alpha self-guidance mask g_l^α to weight the sum of the provisional alpha matte α_l' and α_l^{base}.

The alpha loss at the lth layer $\mathcal{L}_\alpha^{(l)}$ is computed by comparing the estimated alpha matte of this layer, i.e. α_l, with the ground truth alpha matte $\hat{\alpha}$. The alpha self-guidance mask g_l^α is multiplied by α_l and $\hat{\alpha}$ respectively to strengthen the loss of key regions. The alpha loss $\mathcal{L}$ incorporates three loss functions [41]: composition loss [1] ($\mathcal{L}_{comp}$); Laplacian loss [29] ($\mathcal{L}_{lap}$) and $l1$ regression loss ($\mathcal{L}_{l1}$). The definition of $l1$ regression loss can be described mathematically as:

$$\mathcal{L}_{l1}(\hat{y}_i, y_i) = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{11}$$

where N represents the number of pixels in the input y_i, and $\hat{y}_i$ is the ground truth. Denoting the ground-truth alpha matte as $\hat{\alpha}$ and the predicted alpha matte as α, the alpha loss function is obtained using the expression below:

$$\mathcal{L}(\hat{\alpha}, \alpha) = \mathcal{L}_{l1}(\hat{\alpha}, \alpha) + \mathcal{L}_{comp}(\hat{\alpha}, \alpha) + \mathcal{L}_{lap}(\hat{\alpha}, \alpha) \tag{12}$$

The alpha loss at each layer is weighted and accumulated, as illustrated in Line 15 of Algorithm 1. The accumulation is assigned progressively increasing weights w_l to strengthen the penalty for alpha matte errors at the higher layers of FB-PRN layers, whereby the value of w_l is set in line with the approach used for [41]. In addition, the foreground-background loss is calculated to enhance the supervision of the definite foreground and background in the current layer. The α_l^{base}, f' and b' are down-sampled to the size of $\lfloor \frac{H}{s} \rfloor \times \lfloor \frac{W}{s} \rfloor$ to align their size. The foreground guide mask g_l^f and background guide mask g_l^b are generated according to α_l^{base}, as illustrated in Equation 13 and 14, respectively.

$$f_{\alpha_l^{base}\to g_l^f}(x, y) = \begin{cases} 0, & \text{if } \alpha_l^{base}(x, y) = 0 \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

$$f_{\alpha_l^{base}\to g_l^b}(x, y) = \begin{cases} 0, & \text{if } \alpha_l^{base}(x, y) = 1 \\ 1, & \text{otherwise} \end{cases} \tag{14}$$

As shown in Line 21 of Algorithm 1, the foreground and background loss $\mathcal{L}_{FB}$ at the lth layer includes the sum of the $l1$ regression loss ($\mathcal{L}_{l1}$) for both foreground and background estimation. The $l1$ regression loss ($\mathcal{L}_{l1}$) is depicted in Eq. (11) and g_l^f and g_l^b are applied to constrain the measurement of foreground and background estimation error within the non-background and non-foreground region in α_l^{base}, respectively. The total loss can be obtained by adding the $\mathcal{L}_\alpha$ with $\mathcal{L}_{FB}$

The outputs of Algorithm 1, i.e. the α_2 and $\mathcal{L}_{total}$, are used as the predicted alpha matte and the loss for training the TOM-PDFB network. respectively.

At the 1st and 2nd layer of the FB-PRN, the α_l^{base} is initialized by the alpha matte estimated at the previous layer α_{l-1}. The remaining steps are identical to those described for the 0th layer of FB-PRN.

FBCE is trained separately to provide accurate foreground and background confidence maps for FB-PRN. In the next step, the transformer encoder, the CNN decoder and FB-PRN are trained simultaneously, utilizing the foreground and background confidence maps supplied by the pre-trained FBCE as input. This training strategy facilitates the learning from foreground and background confidence maps provided by FBCE and allows the transformer encoder, the CNN decoder and FB-PRN to benefit from the additional information about the definite foreground and background.

IV. EXPERIMENTS

A. Dataset

Three image matting datasets were utilized to evaluate the performance of the proposed method, including Composition-1k [1], Distinction-646 [15] and Transparent-460 [2].

Composition-1k dataset [1] is recognized as one of the most popular image matting datasets, providing both a training set and a test set. The training set of Composition-1k dataset consists of 431 foreground object images with ground truth alpha mattes. Training images are generated using the data augmentation schemes introduced in GCA [17] in which background images for training are randomly sampled from the MS COCO dataset [45] and trimaps for training are generated based on alpha mattes. The test set comprises 50 unique foreground images, each composited with 20 background images pre-defined from the PASCAL VOC2012 dataset [46] using the composition method provided with the test set, which follows the formula shown in Eq. (1) [47], resulting in 1000 test images. The corresponding trimaps are provided in the test set of Composition-1k dataset.

Distinction-646 [15] dataset includes 1,000 test images obtained using a similar methodology as Composition-1k test set. However, this dataset was released without official trimaps or other types of guidance. Therefore, the trimap generation method introduced in [1] was employed to generate trimaps from the ground-truth alpha mattes. Specifically, morphological dilation and erosion operations were applied to the region with a value of 0 and the region with a value of 1 in ground-truth alpha mattes with the random size of structural element ranging from 1 to 29.

Transparent-460 [2] mainly feature of transparent objects as the foreground, including water drops, jellyfish, plastic bags, glass, and crystals, among others. The test set consists of 1,000 images obtained using a similar methodology to that adopted in the Composition-1k test set. The dataset exhibits an average resolution of 3820×3766. Due to the rich details brought by the high resolution, this dataset was utilized in the experiments to evaluate the image matting methods performance when applied to images containing transparent objects.

B. Evaluation Metrics

For the evaluation purposes, four metrics defined in [48] were utilized, namely Sum of Absolute Difference (SAD), Mean Squared Error (MSE), Gradient (Grad) and Connectivity (Conn).

SAD and MSE are widely used to measure the difference between the predicted alpha values and the ground-truth alpha values for each pixel. On the other hand, Grad and Conn metrics are calculated for assessing the visual quality of the alpha matte predictions. Grad measures the gradient error and Conn evaluates the connectivity error difference between the alpha matte image and its corresponding ground-truth, as shown below:

$$SAD = \sum_{i}^{n} |\alpha_i - \hat{\alpha_i}| \tag{15}$$

$$MSE = \frac{1}{n} \sum_{i}^{n} (\alpha_i - \hat{\alpha_i})^2 \tag{16}$$

$$Grad = \sum_{i} (\nabla\alpha_i - \nabla\hat{\alpha_i})^q \tag{17}$$

$$Conn = \sum_{i} (\varphi(\alpha_i, \Omega) - \varphi(\hat{\alpha_i}, \Omega))^p \tag{18}$$

In the expression given above, n represents the number of pixels in the unknown region of the trimap. For pixel i, α_i denotes the predicted alpha value, and $\hat{\alpha_i}$ represents the ground-truth alpha value. The function φ measures the connectivity of pixel i with respect to the source region Ω. The source region Ω represents the largest connected region that is completely opaque in both the predicted alpha matte and the ground-truth alpha matte. Finally, the exponents q and p are custom parameters. Note that the MSE values reported in this work are scaled down by a factor of 1e-3 to facilitate readability.

C. Implementation Details

The proposed method was implemented using PyTorch [49]. The network was initialized with the Tiny model of Swin-Transformer pretrained on ImageNet [50]. During training, the network input size was set to 512×512, batch size was set to 20, and Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ was adopted. The learning rate was initialized to 4×10^{-4}. The training consisted of 200, 000 iterations, with the first 10, 000 treated as a warm-up, and the learning process exhibited a cosine learning rate decay. The data augmentation followed a similar approach as MatteFormer [25]. The state-of-the-art approaches tested as a part of the experiments were trained according to their respective suggested methods. All experiments were run on a server with a Intel Xeon Gold 5218 CPU and two NVIDIA GeForce RTX 4090 GPUs.

D. Experiments on Transparent Objects

To verify the utility of TOM-PDFB in boosting the matting performance when applied to images containing transparent objects, we applied it to two types of objects and compared the output to that yielded by the state-of-the-art image matting methods based on the four evaluation metrics described in Section IV-B. For this purpose, the methods used in comparison were trained on the Composition-1k training set and the evaluation was performed on the Composition-1k test set.

The Composition-1k test set was categorized into Transparent Partially (TP) and Transparent Totally (TT), following the strategy adopted in TransMatte [2]. Accordingly, TT denotes foreground objects that are semi-transparent or highly transparent, with minor or non-salient foreground regions in the trimap. In contrast, TP indicates that the object has significant definite foreground areas, resulting in the presence of large known foreground regions in the trimap, which provides essential information for predicting alpha values in the unknown regions. Table II and Table V present the results obtained when TOM-PDFB and other state-of-the-art methods were applied to TT and TP images in the Composition-1k test dataset, respectively. In this section, the experimental results for TT objects are discussed, while Section IV-E is designated for those pertaining to TP objects.

As illustrated in Table II, when applied to the TT objects, TOM-PDFB outperforms all other tested methods in terms of all four metrics. Moreover, TOM-PDFB exhibits 6.35%, 12.07%, 26.66%, and 8.60% improvements in SAD, MSE,

TABLE II

QUANTITATIVE COMPARISON OF IMAGE MATTING PERFORMANCE FOR TRANSPARENT TOTALLY (TT) OBJECTS ON THE COMPOSITION-1K DATASET. TT OBJECTS ARE THOSE WHOSE ENTIRE IMAGE IS HIGHLY TRANSPARENT

Method	Year	SAD	MSE (10^{-3})	Grad	Conn
DIM [1]	2017	130.41	29.49	85.92	134.80
IndexNet [19]	2019	109.11	22.21	67.05	107.95
ContextNet [29]	2019	87.63	14.37	44.32	85.18
GCA-Matting [17]	2020	84.57	14.90	40.81	81.03
SIM [21]	2021	68.68	10.34	28.82	63.69
FBAMatting [20]	2020	61.61	9.17	25.34	52.45
TransMatte [2]	2022	59.61	7.47	23.29	51.06
MatteFormer [25]	2022	55.82	5.89	20.48	46.67
Ours	-	**48.39**	**4.53**	**16.23**	**37.81**

Grad, and Conn metrics, respectively, compared to MatteFormer [25] which ranks second in the comparison. This evaluation thus confirms the effectiveness of TOM-PDFB when applied to TT objects. The outstanding performance can be attributed to the fact that the known foreground region is small or absent in the trimap for TT objects. Image matting for TT objects faces a unique challenge: it cannot leverage the information provided by the known foreground and background regions in the trimap. TOM-PDFB addresses this issus by predicting and exploiting the known foreground and background information in the unknown region which provide crucial clues for improving the alpha matte quality of transparent objects.

Figure 5 provides a visual comparison of alpha mattes obtained by TOM-PDFB and the state-of-the-art methods used in this evaluation when applied the Composition-1K dataset [1], focusing on the edges of transparent objects. Six state-of-the-art matting methods were employed, including DIM [1], IndexNet [19], GCA-Matting [17], SIM [21], TransMatte [2] and MatteFormer [25]. DIM, IndexNet, and MatteFormer focus on matting general objects, while GCA, SIM, and TransMatte are designed for transparent object matting. The alpha mattes of two completely transparent objects from the Composition-1k test set are shown in Figure 5. In the first case, the tested state-of-the-art methods used in the comparison struggled to distinguish the foreground filament from the background clothes. This leads to alpha estimation errors. However, TOM-PDFB accurately distinguished foreground from the background. The performance advantage of TOM-PDFB can be attributed to its utilization of the definite foreground information about the filament and the definite background information about clothes in the unknown region. Consequently, TOM-PDFB is capable of obtaining distinguishable features in this challenging case. In the second case, the visually prominent background makes image matting methods difficult to clearly discern the edges of the foreground cup. Consequently, the alpha matte provided by tested state-of-the-art methods degrades. However, TOM-PDFB effectively identifies the definite foreground and background in the unknown regions. As a result, it provides a high-quality alpha matte that thoroughly removes the background, provides a clear edge at the cup foot and retains the texture details.

To evaluate the performance of the presented TOM-PDFB when applied to transparent objects and assess its

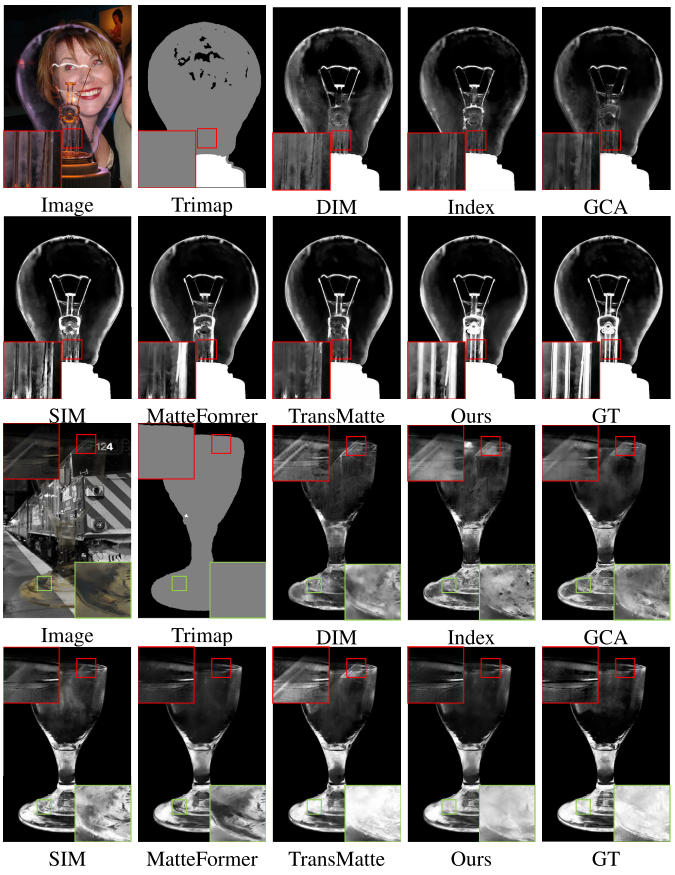


Fig. 5. The visual comparison results of the proposed method and the state-of-the-art methods when applied to the images in the Composition-1k. Best viewed by zooming in.

TABLE III

GENERALIZATION ABILITY COMPARISON OF IMAGE MATTING METHODS ON TRANSPARENT-460 TEST DATASET. ALL THE METHODS WERE TRAINED ON THE TRAINING SET OF COMPOSITION-1K DATASET

Method	Year	SAD	MSE (10^{-3})	Grad	Conn
IndexNet [19]	2019	434.14	74.73	124.98	368.48
DIM [1]	2017	351.98	53	151.37	292.04
MGMatting [41]	2021	344.65	57.25	74.54	282.79
TIMI-Net [18]	2021	328.08	44.2	142.11	289.79
FBAMatting [20]	2020	215.08	20.85	73.25	191.51
TransMatte [2]	2022	192.36	20.96	41.8	158.37
MatteFormer [25]	2022	190.42	**20.72**	36.39	158.76
Ours	-	**188.65**	21.95	**30.31**	**145.96**

generalization capabilities, Seven state-of-the-art matting models were trained on the Composition-1k training set and were tested on the Transparent-460 test set, including DIM [1], IndexNet [19], MGMatting [41], TIMI-Net [18], FBAMatting [20], TransMatte [2] and MatteFormer [25]. The Transparent-460 test set was chosen as it mainly consists of high-resolution images featuring transparent objects in the foreground. SAD, MSE, Grad and Conn metrics were employed to provide a quantitative comparison of the quality of alpha mattes obtained by the tested methods.

Table III summarizes the image matting performance of TOM-PDFB and the aforementioned state-of-the-art methods in terms of the four metrics. The presented TOM-PDFB outperforms all tested image matting methods on the SAD, Grad and Conn metrics and outperforms the CNN-based

TABLE IV
PERFORMANCE OF THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS WHEN APPLIED TO THE COMPOSITION-1K TEST SET

Method	Year	SAD	MSE (10^{-3})	Grad	Conn
DIM [1]	2017	50.4	14	31.0	50.8
IndexNet [19]	2019	45.8	13	25.9	43.7
ContextNet [29]	2019	35.8	8.2	17.3	33.2
GCA-Matting [17]	2020	35.3	9.1	16.9	32.5
MGMatting [41]	2021	31.5	6.8	13.5	27.3
TIMI-Net [18]	2021	29.1	6.0	12.9	27.3
SIM [21]	2021	27.7	5.6	10.7	24.4
FBAMatting [20]	2020	26.4	5.4	10.6	21.5
TransMatte [2]	2022	25.0	4.6	9.7	20.2
MatteFormer [25]	2022	23.8	4.0	8.7	20.5
Ours	-	**21.5**	**3.4**	**7.4**	**16.1**

TABLE V
QUANTITATIVE COMPARISON OF IMAGE MATTING PERFORMANCE FOR TRANSPARENT PARTIALLY (TP) OBJECTS ON THE COMPOSITION-1K DATASET. TP DENOTES OBJECTS WITH THE SIGNIFICANT KNOWN FOREGROUND

Method	Year	SAD	MSE (10^{-3})	Grad	Conn
DIM [1]	2017	20.19	11.89	9.86	18.68
IndexNet [19]	2019	18.61	9.28	8.22	16.09
GCA-Matting [17]	2020	14.15	6.58	6.69	11.75
ContextNet [29]	2019	13.62	5.63	5.79	10.94
FBAMatting [20]	2020	11.23	3.82	4.35	8.23
SIM [21]	2021	10.28	3.73	3.64	7.85
TransMatte [2]	2022	10.10	3.34	3.91	6.92
MatteFormer [25]	2022	10.08	3.24	3.62	6.99
Ours	-	**8.62**	**2.45**	**2.96**	**5.41**

image matting methods on the MSE metirc. It is worth noting that the presented TOM-PDFB achieved over 16.7% reduction in Grad and over 7.8% reduction in Conn relative to MatteFormer which ranks second in the comparison. The significant improvement on the Grad and Conn metrics shows that TOM-PDFB estimates alpha mattes with smooth surface changes and clear boundaries. This superior performance is attributed to the ability of FB-PRN to progressively reduce the unknown regions by leveraging definite foreground and background information across different semantic levels. These results once again confirm the effectiveness of TOM-PDFB in matting transparent objects, which is due to the exploration of definite foreground and background in the unknown region. Although TOM-PDFB fails to beat MatteFormer [25] on the MSE metric, the difference in the attained values is not significant, indicating that TOM-PDFB is still competitive on MSE metric.

E. Experiments on General Objects

The purpose of this experiment was to validate the image matting performance of TOM-PDFB when applied to general objects. Accordingly, TOM-PDFB was compared with 10 state-of-the-art image matting methods including eight CNN-based methods(DIM [1], IndexNet [19], ContextNet [29], GCA-Matting [17], MGMatting [41], TIMI-Net [18], SIM [21] and FBAMatting [20]) and two transformer-based methods(TransMatte [2] and MatteFormer [25]). All tested methods were trained on the Composition-1k dataset [1] and were evaluated on the Composition-1k test set. To provide a quantitative comparison of the quality of alpha mattes obtained by the tested methods, SAD, MSE, Grad and Conn metrics were employed. As shown in Table IV, when applied to the on the Composition-1k test set, TOM-PDFB outperforms all involved image matting methods on all four metrics, which confirms its suitability for not only matting transparent objects but also matting general objects. Moreover, TOM-PDFB demonstrated 9.67%, 15.00%, 14.94%, and 21.46% improvements on the SAD, MSE, Grad, and Conn metrics compared to the state-of-the-art MatteFormer [25] which ranks second in the comparison. It maintains a matting performance advantage over existing methods for general objects. These results indicate that predicting and utilizing definite foreground and background in the unknown region can significantly improve the matting performance for transparent objects, while also providing benefits for matting non-transparent objects.

Additional experiment was conducted to assess the generalizability of the proposed TOM-PDFB method to images with partially transparent objects. For this purpose, the matting performance of TOM-PDFB and seven state-of-the-art methods was evaluated on TP objects in test set of Composition-1K. Table V illustrates the results. Table II and V show the quantitative comparison of image matting performance for TT and TP objects on the Composition-1k dataset. TOM-PDFB outperforms all the tested methods in terms of all fore metrics, not only for TT objects but also for TP objects. These results demonstrate its generalizability, confirming that TOM-PDFB can handle both transparent and partially transparent objects. We also notice a significant discrepancy in metric values between TT and TP objects. The significant discrepancy indicates the inherent challenge of obtaining high-quality alpha mattes for transparent objects as these trimaps for transparent objects often contains a large unknown region and small known foreground and/or background regions. Image matting for TP objects can leverage a substantial amount of known information from the indicated foreground and background regions in the trimap. In contrast, image matting for transparent objects faces a scarcity of known information. Consequently, extracting and utilizing potentially definite foreground and background from the unknown regions plays an important role in transparent object matting. Experimental results have demonstrated that by effectively exploring potentially determined foregrounds and backgrounds in unknown regions, TOM-PDFB extends its suitability to partially transparent objects as well.

The purpose of this experiment was to assess the generalization capabilities of TOM-PDFB when applied to previously unseen data. TOM-PDFB was compared with five state-of-the-art methods, including four CNN-based methods (DIM [1], IndexNet [19], GCA-Matting [17], and TIMI-Net [18]), and a transformer-based method, MatteFormer [25]. All these methods were trained on the Composition-1k dataset [1] and tested on the Distinction-646 dataset [15]. As shown in Table VI, TOM-PDFB outperformed all the tested methods across all four metrics. The transformer-based methods, i.e. TOM-PDFB and MatteFormer [25], show significant advantages over the CNN-based methods in terms of all four metrics, and TOM-PDFB has small values on all four metrics indicating its ability to provide high-quality alpha mattes.

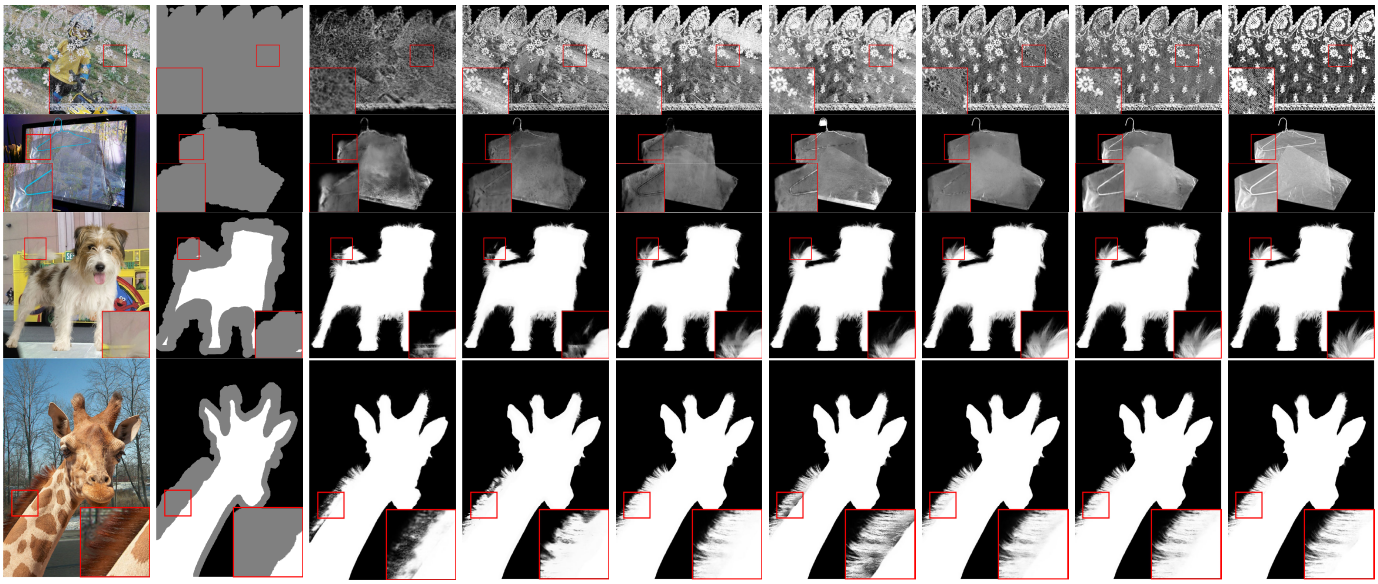


Fig. 6. Visual comparison of alpha mattes obtained by TOM-PDFB and five state-of-the-art image methods when applied to four images from the Distinction-646 test set. All tested methods were trained on the Composition-1K training set. Best viewed by zooming in.

TABLE VI

GENERALIZATION ABILITY COMPARISON OF IMAGE MATTING METHODS ON DISTINCTION-646 TEST DATASET. ALL THE METHODS WERE TRAINED ON THE TRAINING SET OF COMPOSITION-1K DATASET

Method	Year	SAD	MSE (10^{-3})	Grad	Conn
DIM [1]	2017	72.56	45.60	80.74	74.39
IndexNet [19]	2019	60.93	34.02	50.26	60.93
GCA-Matting [17]	2020	50.12	28.50	36.92	49.01
TIMINet [18]	2021	53.07	27.22	41.24	52.62
MatteFormer [25]	2022	32.18	16.49	13.96	29.44
TOM-PDFB	-	**31.86**	**15.72**	**12.49**	**28.95**

TABLE VII

ABLATION STUDIES CONDUCTED ON THE COMPOSITION-1K DATASET: EVALUATING THE ROLE OF FB-PRN (P), FBCE (C), SEGMENT MAP WITH EQUAL INTERVALS (S_{avg}), AND SEGMENT MAP WITH DEFINITE FOREGROUND, DEFINITE BACKGROUND, AND SEMI-TRANSPARENT REGIONS (S_{fb})

Method	SAD	MSE (10^{-3})	Grad	Conn
baseline	24.64	4.07	8.90	19.85
baseline + S_{avg}	24.01	3.95	8.89	19.03
baseline + S_{fb}	23.24	3.91	8.94	18.18
baseline + P	23.89	3.86	8.69	19.17
baseline + C	23.05	3.82	8.89	18.11
baseline + P + S_{fb}	22.40	3.64	8.19	17.12
baseline + P + C	21.45	3.35	7.37	16.08

Figure 6 presents a visual comparison of the alpha mattes obtained by applying the tested matting methods to the Distinctions-646 test set. As shown in the first two rows of Figure 6, the alpha matte produced by TOM-PDFB preserves the intricate structural details of transparent foreground objects. Moreover, TOM-PDFB can not only handle transparent objects, it also provides high-quality alpha mattes for opaque objects (as shown in the last two rows of Figure 6). This benefit is attained because a challenging aspect of matting opaque objects lies in their semi-transparent or transparent elements, and TOM-PDFB is equipped to handle such challenges. Additionally, both FBCE and FB-PRN presented in this work focus on determining definite foreground and background, aiding in predicting opaque objects that typically have large definite foreground and background regions. These experimental results reaffirm the matting performance superiority of TOM-PDFB and underscore its generalization capabilities.

F. Ablation Study

To evaluate the effectiveness of FBCE and FB-PRN, an ablation study was conducted on the Composition-1k dataset.

The effectiveness of the definite foreground and background in the unknown region was evaluated in the first experiment of the ablation study. Two strategies for dividing the unknown region of the trimap were compared. The first approach - denoted as baseline + S_{avg} - segregated the unknown region of the trimap into three equal intervals based on the alpha value. The second approach - denoted as baseline + S_{fb} - segregated the unknown region of the trimap into definite foreground, definite background, and semi-transparent image regions. As can be seen from Table VII, the obtained results clearly demonstrate the benefits of predicting definite foreground and definite background for the matting performance.

The effectiveness of FBCE was assessed by comparing the matting results of baseline + C (using FBCE to generate the foreground confidence map and background confidence map) with baseline + S_{fb} (using cross-entropy loss for segmentation). The ablation study results presented in Table VII, confirm that utilizing the output from FBCE provided a slight improvement over baseline + S_{fb}.

Subsequently, the FB-PRN effectiveness was evaluated. The quantitative results presented in Table VII demonstrate that, when FB-PRN was used in isolation, the performance improvement was limited. However, when FB-PRN was combined with S_{fb} or C (serving as prior information), a significant performance improvement is observed between the results of baseline + P + S_{fb} and that of baseline + P + C.

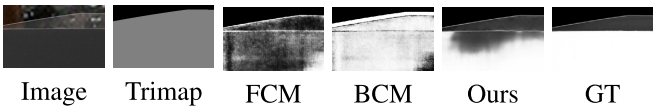


Fig. 7. An example of the failure case of TOM-PDFB.

This observation indicates that the inclusion of S_{fb} or C greatly benefits FB-PRN. Furthermore, the enhancements in the combined results compared to those achieved using only S_{fb} or C, suggesting that FB-PRN effectively utilizes the prior information provided by S_{fb} or C.

G. Limitations

While TOM-PDFB consistently delivers exceptional matting results across various datasets and object types, there are scenarios where it might fall short in producing high-quality alpha mattes. Specifically, this limitation arises when the trimap used as its input is inaccurate or when prediction errors occur during the Foreground-Background Confidence Estimator (FBCE) process.

TOM-PDFB achieves remarkable performance by predicting and leveraging the definite foreground and background in the unknown region. FBCE plays a crucial role in this process, determining the definite foreground and background according to the provided image and trimap. However, when an incorrectly labeled trimap is used as input, FBCE may predict inaccurate FCM and BCM and provide incorrect information for image matting resulting in degraded matting performance.

In cases where prediction errors arise during the Foreground-Background Confidence Estimator (FBCE) process, leading to inaccurately predicted FCM and BCM, the model can be misled. Consequently, this undermines the effectiveness of the subsequent alpha matting process. Although FB-PRN can gradually reduce the unknown region at different levels and thus partially correct the errors, there may still be significant biases that resulting in suboptimal matting results. An example is shown in Fig. 7. In this case, the background confidence map (BCM) predicted by FBCE contains error: the bottom part of the image is incorrectly predicted as the background. This discrepancy leads to errors in the estimation of alpha matte.

V. Conclusion

As a part of this work, a method named Transparent Object Matting using Predicted Definite Foreground and Background (TOM-PDFB) for precise matting of images containing transparent objects was proposed. It comprises of the Foreground-Background Confidence Estimator (FBCE) that explores potentially definite foreground and background in the unknown region to developing the matting network. In the next step, the proposed Foreground-Background Progressive Refinement Network (FB-PRN) using the definite foreground and background predicted in the previous decoder layers to progressively refine the alpha matte.

The extensive experimental results reported in the preceding sections demonstrate the effectiveness of TOM-PDFB when applied not only on images containing transparent objects but also general objects, thus confirming the importance of definite foreground and background for transparent object matting. However, erroneous FCM or BCM produced by FBCE or an inaccurate trimap have the potential to mislead the subsequent alpha matting process and lead to discrepancies in the predicted alpha matte. Feature work includes exploring methods to exclude outliers during FBCE prediction, aiming to enhance the robustness of TOM-PDFB, especially when dealing with inaccurate trimaps.

References

[1] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 311–320.

[2] H. Cai, F. Xue, L. Xu, and L. Guo, "TransMatting: Enhancing transparent objects matting with transformers," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 253–269.

[3] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A Bayesian approach to digital matting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Dec. 2001, pp. 1–13.

[4] Y. Liang, H. Huang, Z. Cai, and Z. Hao, "Multiobjective evolutionary optimization based on fuzzy multicriteria evaluation and decomposition for image matting," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 5, pp. 1100–1111, May 2019.

[5] H. Huang, Y. Liang, X. Yang, and Z. Hao, "Pixel-level discrete multiobjective sampling for image matting," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3739–3751, Aug. 2019.

[6] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, Sep. 2013.

[7] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008.

[8] X. Chen, D. Zou, S. Z. Zhou, Q. Zhao, and P. Tan, "Image matting with local and nonlocal smooth priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1902–1907.

[9] Y. Aksoy, T. O. Aydin, and M. Pollefeys, "Designing effective inter-pixel information flow for natural image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 228–236.

[10] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2288–2297.

[11] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8762–8771.

[12] X. Yang et al., "Smart scribbles for image matting," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 4, pp. 1–21, Nov. 2020.

[13] Y. Zhang et al., "A late fusion CNN for digital matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7461–7470.

[14] J. Li, J. Zhang, S. J. Maybank, and D. Tao, "Bridging composite and real: Towards end-to-end deep image matting," *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 246–266, Feb. 2022.

[15] Y. Qiao et al., "Attention-guided hierarchical structure aggregation for image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13673–13682.

[16] Q. Liu, S. Zhang, Q. Meng, B. Zhong, P. Liu, and H. Yao, "End-to-end human instance matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2633–2647, Apr. 2024.

[17] Y. Li and H. Lu, "Natural image matting via guided contextual attention," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11450–11457.

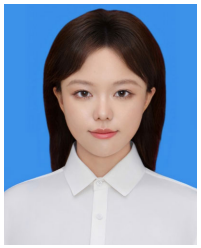
[18] Y. Liu et al., "Tripartite information mining and integration for image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7555–7564.

[19] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3265–3274.

[20] M. Forte and F. Pitié, "F, B, alpha matting," 2020, *arXiv:2003.07711*.

[21] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11115–11124.
[22] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[23] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[24] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12124–12134.
[25] G. Park, S. Son, J. Yoo, S. Kim, and N. Kwak, "MatteFormer: Transformer-based image matting via prior-tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11686–11696.
[26] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, "Learning-based sampling for natural image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3050–3058.
[27] S. Cai et al., "Disentangled image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8818–8827.
[28] A. Goel et al., "IamAlpha: Instant and adaptive mobile network for alpha matting," in *Proc. BMVC*, London, U.K., 2021, pp. 1–11.
[29] Q. Hou and F. Liu, "Context-aware image matting for simultaneous foreground and alpha estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4129–4138.
[30] G. Chen, K. Han, and K. K. Wong, "TOM-Net: Learning transparent object matting from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9233–9241.
[31] G. Chen, K. Han, and K.-Y.-K. Wong, "Learning transparent object matting," *Int. J. Comput. Vis.*, vol. 127, no. 10, pp. 1527–1544, Oct. 2019.
[32] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin, "Environment matting and compositing," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, 1999, pp. 205–214.
[33] Y. Liu, L. Zhou, G. Wu, S. Xu, and J. Han, "TCGNet: Type-correlation guidance for salient object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6633–6644, Jul. 2024.
[34] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3688–3704, Jul. 2022.
[35] Y. Liu, D. Zhang, N. Liu, S. Xu, and J. Han, "Disentangled capsule routing for fast part-object relational saliency," *IEEE Trans. Image Process.*, vol. 31, pp. 6719–6732, 2022.
[36] Y. Liu, X. Dong, D. Zhang, and S. Xu, "Deep unsupervised part-whole relational visual saliency," *Neurocomputing*, vol. 563, Jan. 2024, Art. no. 126916.
[37] Q. Liu, H. Xie, S. Zhang, B. Zhong, and R. Ji, "Long-range feature propagating for natural image matting," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 526–534.
[38] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.
[39] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[40] L. Hu, Y. Kong, J. Li, and X. Li, "Effective local–global transformer for natural image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, p. 3888–3898, 2023.
[41] Q. Yu et al., "Mask guided matting via progressive refinement network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1154–1163.
[42] Y. Zhou, L. Zhou, T. L. Lam, and Y. Xu, "Sampling propagation attention with trimap generation network for natural image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5828–5843, 2023.
[43] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 800–806.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[45] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[47] A. R. Smith and J. F. Blinn, "Blue screen matting," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 1996, pp. 259–268.
[48] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1826–1833.
[49] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.
[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

Yihui Liang received the B.S. degree in digital media technology from Xi'an University of Technology, China, in 2012, and the M.Eng. and Ph.D. degrees in software engineering from South China University of Technology, China, in 2015 and 2019, respectively. He is currently an Associate Professor with the School of Computer Science, University of Electronic Science and Technology of China, Zhongshan Institute. His current research interests include alpha matting and image processing.



Qian Fu received the M.S. degree in computer science and technology from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2024. Her research interests include transparent object matting.

Kun Zou received the B.S. degree in computer science and technology from the Huazhong University of Science and Technology, China, in 2003, and the Ph.D. degree in technology for computer applications from South China University of Technology, Guangzhou, China, in 2008. He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His current research interests include graphics and image processing.

Guisong Liu (Member, IEEE) received the B.S. degree in mechanics from Xi'an Jiaotong University, Xi'an, China, in 1995, and the M.S. degree in automatics and the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2000 and 2007, respectively. He was a Visiting Scholar with Humbolt University, Berlin, Germany, in 2015. Before 2021, he was a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. He is currently a Professor and the Dean of the School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu. He has filed over 20 patents, and published over 70 scientific conference papers and journal articles. His research interests include pattern recognition, neural networks, and machine learning.

Han Huang (Senior Member, IEEE) received the B.Man. degree in information management and information systems from the School of Mathematics, South China University of Technology (SCUT), Guangzhou, China, in 2003, and the Ph.D. degree in computer science from SCUT in 2008. He is currently a Full Professor with the School of Software Engineering, SCUT. His research interests include theoretical foundation and application of evolutionary computation and microcomputation. He is a Distinguished Member of CCF.