# Text-Guided Portrait Image Matting

Yong Xu, Xin Yao, Baoling Liu, Yuhui Quan, and Hui Ji

***Abstract*—Image matting is a technique used to separate the foreground of an image from the background, which estimates an alpha matte that indicates pixel-wise degree of transparency. To precisely extract target objects and address the ambiguity of solutions in image matting, many existing approaches employ a trimap or background image provided by the user as additional input to guide the matting process. This paper introduces a novel matting paradigm termed text-guided image matting, utilizing a textual description of the foreground object as a guiding element. In contrast to trimap or background-based methods, text-guided matting offers a user-friendly interface, providing semantic clues for the objects of interest. Moreover, it facilitates batch processing across multiple frames featuring the same objects of interest. The proposed text-guided matting approach is implemented through a deep neural network comprising three-stage cross-modal feature fusion and two-step alpha matte prediction. Experimental results on portrait matting demonstrate the competitive performance of our text-guided approach compared to existing trimap-based and background-based methods.**

***Impact Statement*—This paper proposes a new approach to image matting, termed text-guided image matting. Departing from conventional guidance-based methods, text-guided matting relies solely on concise textual descriptions of the foreground object for guidance. It provides semantic insights and facilitates efficient batch processing of multiple frames with identical objects. The deep neural network developed for this purpose shows competitive performance in portrait matting, outperforming traditional trimap-based or background-based methods. This work marks a significant step towards more intelligent image matting solutions, enhancing user-friendliness through the integration of semantically-driven artificial intelligence.**

***Index Terms*—Image Matting, Cross-modal Learning, Attention, Text Gudiance.**

## I. Introduction

Image matting is an important tool in the realm of multimedia, with a broad spectrum of applications ranging from image editing and image fusion to digital art creation and filmmaking; see *e.g.* [37], [39]. In image matting, an image $\bar{\boldsymbol{I}}$ is modeled as the composite of a foreground $\boldsymbol{F}$ and a background $\boldsymbol{B}$:

$$\bar{\boldsymbol{I}} = \boldsymbol{\alpha} \odot \boldsymbol{F} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \boldsymbol{B}, \qquad \boldsymbol{\alpha}(i) \in [0, 1], \qquad (1)$$

where $\boldsymbol{\alpha}$, the so-called *alpha matte*, represents the opacity of the foreground intensity at each pixel. The symbol $\odot$ denotes the operation of element-wise multiplication. The main task of image matting is about extracting both the foreground and the alpha matte from an input image. This extracted information is instrumental in the synthesis of new, natural images against varying backgrounds.

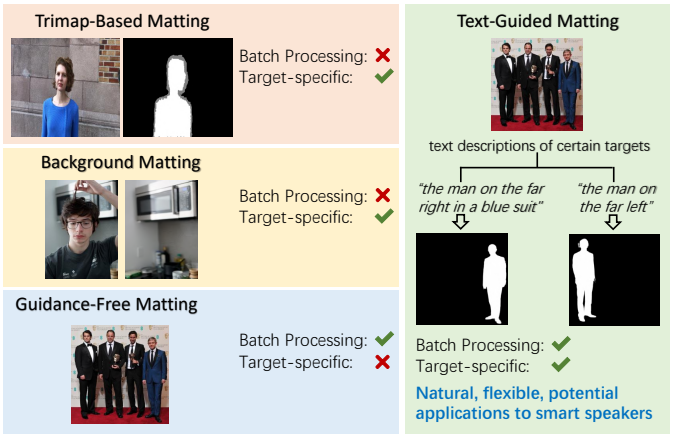



Fig. 1. Illustration of text-guided matting and other matting paradigms.

Image matting is a complex and inherently ill-posed problem, presenting significant solution ambiguity as it requires estimating three unknowns at each pixel. To address this ambiguity, many existing methods necessitate additional input for guidance. For instance, a wide range of studies employ a user-annotated trimap (*i.e.*, a binary image with three regions: foreground, background, and unknown) for guidance; see *e.g.* [2]–[4], [12], [19], [29], [31], [34], [38], [40], [41], [45], [48]–[50], [53], [54], [56], [57], [66], [67]. To lessen the burden on the user, some recent works utilize a captured background image for guidance [26], [44], [58]. However, these methods can be labor-intensive, particularly in batch processing like video matting, due to the demands for extensive annotation or data acquisition. Moreover, background-based matting often has usability limitations, requiring specific hardware or user skills for capturing consistently lit, well-aligned images. Guidance-free matting (*e.g.* [21], [64]) attempts to circumvent these challenges, where no additional guidance information is introduced. However, this guidance-free approach is substantially more difficult, and existing attempts have not yet reached satisfactory performance levels. Also it cannot specify particular objects of interest.

Addressing the limitations of existing methods, this paper develops a novel image matting paradigm, termed text-guided image matting. This method leverages concise textual descriptions of the object of interest to guide the matting process. Essentially, it automates the creation of the alpha matte based on these verbose textual descriptions; refer to Fig. 1 for an illustration comparing this new matting paradigm with others.

Yong Xu, Xin Yao, Baoling Liu and Yuhui Quan are with School of Computer Science and Engineering at South China University of Technology, Guangzhou 510000, China. (email: yxu@scut.edu.cn; xinyao240@gmail.com; csblliu@foxmail.com; csyhquan@scut.edu.cn)

Hui Ji is with Department of Mathematics at National University of Singapore 119076, Singapore. (email:matjh@nus.edu.sg)

Corresponding author: Yuhui Quan.

This work was supported in part by National Natural Science Foundation of China under Grants 62072188 and 62372186, in part by Natural Science Foundation of Guangdong Province under Grants 2023A1515012841 and 2022A1515011755, in part by Fundamental Research Funds for the Central Universities under Grant x2jsD2230220, and in part by Singapore MOE AcRF Tier 1 under Grant A-8000981-00-00.

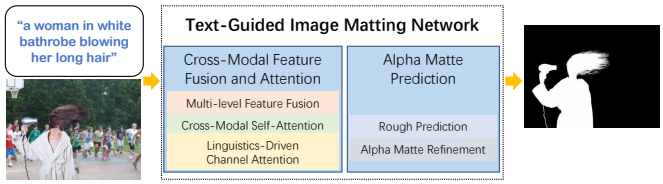



Fig. 2. Illustration of basic procedure of proposed text-guided matting.

This new paradigm achieves an ideal equilibrium between user-friendliness and performance. It exceeds trimap-based or background-based matting in ease of use, while requiring minimal additional effort compared to guidance-free matting.

The proposed text-guided matting paradigm offers several practical advantages. First, it facilitates natural and descriptive expression of objects of interest. Second, textual descriptions offer semantic matting cues that could provide more guidance than auxiliary images. For instance, the description "with long brown hair" aids in recognizing both color and shape for alpha matte prediction. Third, images on websites and social media often come with captions, which enables automatic text description generation, or at least a simplified process through advanced image captioning tools (*e.g.* [7]). Fourth, text descriptions can be reused across multiple image frames, eliminating the need for frame-by-frame annotation in batch processing. Lastly, the integration of speech recognition paves the way for hands-free operation in text-guided matting, offering potential applications in smart speakers and other intelligent devices.

However, to realize the advantages mentioned above, text-guided image matting poses greater challenges than trimap-based or background-based methods. This is because a textual description offers much weaker constraints on an object compared to a trimap or a background image. Additionally, text-guided image matting is a cross-modal problem, as textual descriptions and images are inherently different modalities. The primary challenge in this cross-modal context is to effectively fuse data from both modalities and extract sufficient information about the object's alpha matte from its textual description. In this paper, we address this challenge by designing a specific deep neural network (NN) to tackle such a challenge, as outlined in Fig. 2.

In the proposed NN, the first part focuses on extracting cross-modal information through feature fusion and attention mechanisms. This includes three key components designed to maximize the guidance from text descriptions: multi-level feature fusion, cross-modal self-attention, and linguistics-driven channel attention. The second part of the NN is dedicated to predicting the alpha matte in two stages. Initially, it employs the fused features to estimate an alpha matte that signifies the foreground referenced by the text description, providing a rough outline of alpha values. Subsequently, this alpha matte undergoes refinement to further improve its accuracy. See below for a summary of the contributions of this paper.

- We introduce a novel image matting paradigm that utilizes text-based cross-modal guidance, offering a user-friendly interface while maintaining competitive performance compared to traditional guidance-based methods.
- We develop a deep NN for text-guided image matting, comprising three-stage cross-modal feature fusion and two-step alpha matte prediction to efficiently leverage cross-modal guidance from textual descriptions.
- Building upon the foundation of established image matting datasets, we have augmented these collections with textual annotations, facilitating future research of deep learning for text-guided matting.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents the proposed text-guided image matting approach. Section IV reports the experimental results. Section V concludes this paper.

## II. Related Work

Image matting has been extensively studied in the past. Most methods introduced constraints or external guidance on objects or backgrounds. Earlier work often used a green/blue-screen environment [47] to simplify the problem by focusing on uniformly colored backgrounds. For more complex backgrounds, many studies have adopted a trimap or scribbles [20] as guidance for solving the problem. There are two main types of trimap-based methods: affinity-based and sampling-based; see [68] for a comprehensive survey. Affinity-based methods (*e.g.* [6], [20]) compute the affinity matrix from the input image and then propagate the alpha values from known areas to unknown areas in the trimap via this affinity matrix. Sampling-based methods (*e.g.* [14]–[16], [51], [53]) estimate the color of unknown areas of foregrounds and backgrounds respectively by sampling the color of known areas, and then estimate the alpha matte via (1). Li *et al.* [24] leveraged manifold learning to combine these two types of methods for further performance improvement.

The aforementioned methods are constrained by handcrafted priors on the alpha matte, limiting their versatility in practice. Recently, deep NNs have been increasingly employed for image matting, which learn matting priors from extensive datasets. Xu *et al.* [57] proposed an end-to-end NN for image matting, together with a benchmark dataset. Wang *et al.* [54] utilized deep features to enhance traditional Laplacian matting. Tang *et al.* [50] developed two NNs for background and foreground sampling, followed by matting. Lu *et al.* [34] proposed an index-guided U-shaped NN for matting. Hou *et al.* [12] employed two NN-based encoders to extract local and global information respectively. Cai *et al.* [4] separated matting into trimap adaptation and alpha estimation using two NNs. Qiao *et al.* [41] applied spatial-channel attention for edge and shape enhancement. Liu *et al.* [29] leveraged coarse annotations from segmentation datasets for weakly-supervised learning. Zhou *et al.* [67] designed an attention transfer module to minimize artifacts in the matte. Liu *et al.* [31] focused on information alignment for fine-grained image detail recovery. Zheng *et al.* [66] redefined matting as a Gaussian process, improving training efficiency with neighbor pixel pairs. Cai *et al.* [3] and Park *et al.* [38] employed transformer-based NNs, redefining trimap input as learnable tri-tokens to integrate trimap information into deep image features.

Recently, there is an increasing interest on non-trimap-based schemes for image matting [26], [44], [55], [63]. Rather than

use a trimap, the mask-guided matting proposed by Yu *et al.* [63] uses coarse binary masks for image matting, which is done by a progressive refinement NN. Ding *et al.* [11] replaced the trimap input with several user clicks, providing a versatile and user-friendly manner. The background matting proposed by Sengupta *et al.* [44] uses a captured background for guidance and trained an NN with an adversarial loss for better adaption to real data. Lin *et al.* [26] proposed a coarse-to-fine NN to accelerate background matting for real-time applications, with a small NN to predict an error map for guiding the refinement of alpha matte. Based on untrained NN priors, Xu *et al.* [58] an unsupervised learning approach to background matting, which requires training two NNs on each test image. Background matting is vulnerable to possible misalignment and illumination changes between the observed image and the background, which often occur in practice.

There are also some attempts on developing image matting methods without external guidance. Zhang *et al.* [64] introduced some form of self-guidance by training two NN-based decoders for foreground and background classification. Li *et al.* [21] proposed to automatically generate a trimap via an attentive NN. In comparison to the trimap/background-based ones, these guidance-free methods are inapplicable to target-specific matting for the image with multiple foreground objects, due to the lack of guidance.

The user interface of our proposed text-guided image matting is akin to the so-called "referring image segmentation" (RIS) introduced by Hu *et al.* [13], designed to segment objects from an image based on natural language descriptions. Ding *et al.* [8] further enhanced this by incorporating clicks, offering additional cues alongside text for more precise object localization. In [13], visual features and linguistic features are encoded via a convolutional NN (CNN) and a recurrent NN (RNN), respectively. These features are fused via $1 \times 1$ convolution and passed to another CNN for segmentation mask prediction. Liu *et al.* [28] suggested initially merging visual features with the linguistic features of each word, followed by encoding the fused word-wise features using long short-term memory (LSTM) units. Ye *et al.* [59] introduced a dual convolutional LSTM network for further performance improvement. Li *et al.* [23] diverged from treating individual word features as RNN units, instead of focusing on refining multi-level visual features recurrently. Parallel to our work, Li *et al.* [22] have also explored referring image matting, with a focus on dataset development.

Although image matting and segmentation seem closely related, they diverge notably in their characteristics and methodologies. Foreground segmentation generally involves predicting a binary mask with 0/1 values to identify foreground objects. In contrast, image matting entails estimating an alpha matte with continuous values, which captures both the object's location and its varying transparency. This added intricacy renders image matting a more complex task than segmentation, especially when dealing with semi-transparent areas. Consequently, the efficient integration of both low-level and high-level features is crucial in image matting. This integration is a central consideration in our design of the NN for text-guided image matting. Specifically, our NN incorporates a three-stage cross-modal feature fusion scheme and features a two-step process to precisely predict a continuous-valued alpha matte. The first step roughly delineates the locations and transparency of foreground objects, leveraging RIS data for augmentation. The following step is dedicated to refining the alpha matte, thereby achieving improved accuracy.

In the design of NN for cross-modal feature fusion, attention mechanisms have been extensively utilized in various cross-modal tasks including RIS (e.g., [1], [18], [30], [35], [65]). Yu *et al.* [61] and Luo *et al.* [36] adopted a two-stage method: initially detecting candidate instances in an image to simulate attention, followed by isolating the target instance as directed by the text description. Shi *et al.* [46] employed self-attention to assess the significance of each word for every pixel, fusing such information through a weighted mean. Ye *et al.* [60] implemented self-attention in their fusion process with multi-level visual features. Chen *et al.* [5] utilized spatial attention to enhance the quality of fused features. Ding et al. [9] leveraged a transformer to extract more contextual information.

## III. Methodology

Our proposed NN for text-guided image matting is illustrated in Fig. 3. It comprises two main components: (i) a three-stage cross-modal feature fusion, which integrates the text description of the object of interest into the feature representation, and (ii) a two-step alpha matte prediction process that initially predicts the alpha matte using the fused features and subsequently refines it for further accuracy improvement.

### A. Cross-Modal Feature Fusion and Attention

The NN's input includes an image of size $H \times W \times C$ and a text description consisting of $N$ words. Initially, the NN utilizes an image encoder and a text encoder to embed input elements into a latent space, enabling the fusion of visual and linguistic semantics. The image encoder, defined as a CNN featuring down-sampling layers, converts the image into a feature tensor. The text encoder applies word embedding to the text description and then channels the resulting $N$ word vectors through the gated recurrent units (GRU). Consequently, the text description is transformed into to $N$ linguistic feature vectors, denoted by $\{\boldsymbol{p}_n\}_{n=1}^{N}$, encoding the contextual information of the words. To effectively utilize the guidance from the text description, we implement a three-stage attention strategy to fuse the features, specifically designed for the matting process.

*1) Multi-level feature fusion:* Linguistic and visual features are initially fused in the first stage as follows. The NN aggregates the linguistic features $\{\boldsymbol{p}_n\}_{n=1}^{N}$ into a single feature vector $\boldsymbol{b}$ using a fully-connected (FC) layer. The vector $\boldsymbol{b}$ is then mapped to a vector $\boldsymbol{b}_0 \in \mathbb{R}^c$ via another FC layer. To align the vision and linguistics domains, the feature tensor from the image encoder is converted to $\boldsymbol{X}_0 \in \mathbb{R}^{h \times w \times c}$ via a convolutional layer and then element-wisely multiplied with $\boldsymbol{b}_0$ along the channel dimension. It is a common strategy in multi-modal fusion [33].

Considering the varying scales of information within the intermediate feature tensors produced by the image encoder,

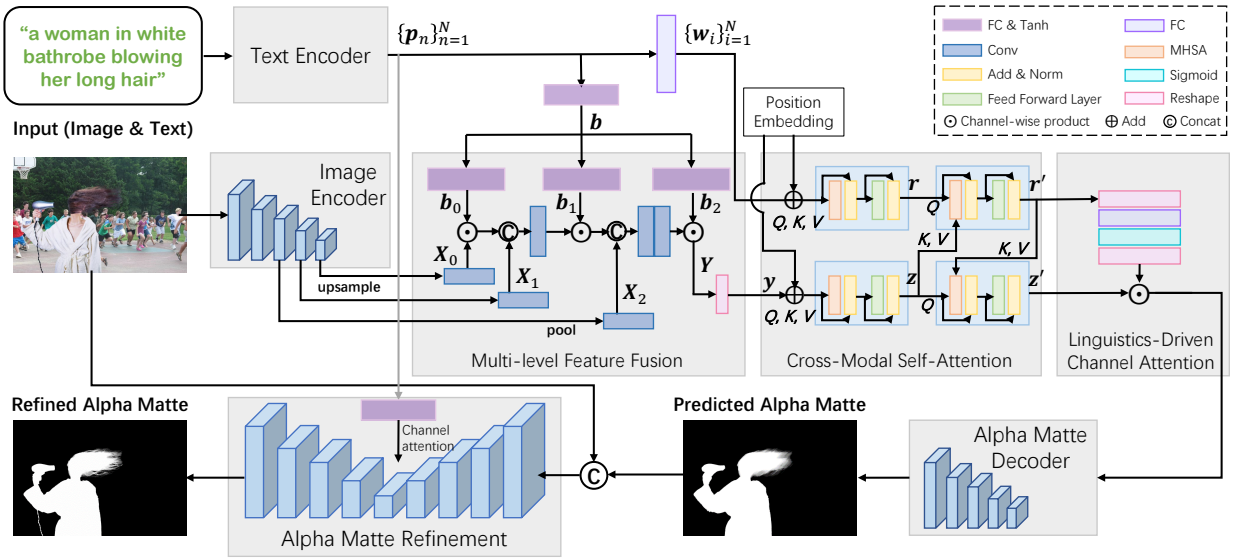



Fig. 3. Architecture of proposed NN for text-guided image matting.

we also incorporate feature tensors from the last two downsampling layers for fusion. These tensors are similarly transformed into $\boldsymbol{X}_1 \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{h\times w\times c}$ via convolutional layers. Each feature tensor $\boldsymbol{X}_j$ at the current scale is combined with the preceding scale's tensor $\boldsymbol{X}_{j-1}$ through concatenation and $1\times 1$ convolution. Concurrently, the linguistic feature $\boldsymbol{v}$ is transformed to corresponding vectors $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^c$ using FC layers for fusion. Finally, a fused feature tensor $\boldsymbol{Y} \in \mathbb{R}^{h\times w\times c}$ is generated, encapsulating context information from the text description and different levels.

*2) Cross-modality self-attention-based fusion:* This stage focuses on exploiting the spatial dependencies and interactions between visual and linguistic features. To accomplish this, multi-head self-attention (MHSA) [52] is initially applied separately to the visual features $\boldsymbol{Y}$ and the linguistic features $\{\boldsymbol{w}_i\}_{i=1}^N$. Then, these features are utilized in a cross-modal fashion to enhance the interaction between the two modalities.

Self-attention relates different positions of a sequence to create another representation of the sequence. It processes each input through three FC layers, yielding sets of keys, queries, and values stored as $\boldsymbol{K} = [\boldsymbol{k}_1, \cdots, \boldsymbol{k}_L] \in \mathbb{R}^{L\times d}$, $\boldsymbol{Q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_L] \in \mathbb{R}^{L\times d}$, and $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_L] \in \mathbb{R}^{L\times d}$ respectively, where $L$ and $d$ denote the number and dimension respectively for keys/queries/values. The output of the self-attention is then calculated as:

$$\mathrm{SA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\Big)\boldsymbol{V}. \tag{2}$$

Subsequently, MHSA is employed to obtain varied representations of $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$, computes self-attention for each, and concatenates the results. MHSA is defined as follows:

$$\begin{aligned}&\mathrm{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = [\mathrm{Head}_1, \cdots, \mathrm{Head}_M]\boldsymbol{W}^{\mathrm{O}}, \\ &\text{where } \mathrm{Head}_m = \mathrm{SA}(\boldsymbol{Q}\boldsymbol{W}_m^{\mathrm{Q}}, \boldsymbol{Q}\boldsymbol{W}_m^{\mathrm{K}}, \boldsymbol{Q}\boldsymbol{W}_m^{\mathrm{V}})\end{aligned} \tag{3}$$

with learnable weight matrices $\{\boldsymbol{W}_m^{\mathrm{Q}}\}_m$, $\{\boldsymbol{W}_m^{\mathrm{K}}\}_m$, $\{\boldsymbol{W}_m^{\mathrm{V}}\}_m$, for $1 \le m \le M$ and $\boldsymbol{W}^{\mathrm{O}}$. For further improvement on training efficiency, a typical MHSA block sequentially attaches a residual addition, a layer normalization, a feed forward layer (*i.e.*, a stack of two FC layers) and a layer normalization after the MHSA, which is also implemented in the proposed NN.

For the feature tensor $\boldsymbol{Y}$, it is flattened into a second-order tensor $\boldsymbol{y} \in \mathbb{R}^{hw\times c}$. Each row $\boldsymbol{y}(i,:) \in \mathbb{R}^c$ is treated as a sequence, and this sequence is fed to an MHSA block, yielding new features $\boldsymbol{z}_i \in \mathbb{R}^c$ for each $i$. Regarding the linguistic features $\{\boldsymbol{p}_n\}_{n=1}^N$, each $\boldsymbol{p}_n$ is first transformed to a feature vector of length $c$. This sequence is then processed by an MHSA block, producing new features $\boldsymbol{r}_n \in \mathbb{R}^c$ for each $n$. Position encoding [52] is also applied to the input of MHSA to improve their awareness to relative locations during self-attention.

To facilitate cross-modal interactions, we initially employ $\{\boldsymbol{z}_i\}_i$ to generate the pair (key, value), denoted by $(\boldsymbol{K}^{\mathrm{Image}}, \boldsymbol{V}^{\mathrm{Image}})$, and use $\{\boldsymbol{r}_n\}_n$ to generate the queries denoted by $\boldsymbol{Q}^{\mathrm{Text}}$. The cross-modal MHSA for $\{\boldsymbol{r}_n\}_n$ is then executed using a modified self-attention that follows [9], [10]:

$$\mathrm{SA}_{\mathrm{Image}}^{\mathrm{Text}} = \mathrm{SA}(\boldsymbol{Q}^{\mathrm{Text}}, \boldsymbol{K}^{\mathrm{Image}}, \boldsymbol{V}^{\mathrm{Image}}). \tag{4}$$

The outputs from this cross-modal MHSA are denoted $\{\boldsymbol{r}'_n\}_n$. Similarly, $\{\boldsymbol{r}'_n\}_n$ is used to create the pairs (key, value) denoted by $(\boldsymbol{K}^{\mathrm{Text}}, \boldsymbol{V}^{\mathrm{Text}})$, and use $\{\boldsymbol{z}_i\}_i$ to create the queries denoted by $\boldsymbol{Q}^{\mathrm{Image}}$. The cross-modal MHSA on $\{\boldsymbol{z}_i\}_i$ is then with a modified self-attention:

$$\mathrm{SA}_{\mathrm{Text}}^{\mathrm{Image}} = \mathrm{SA}(\boldsymbol{Q}^{\mathrm{Image}}, \boldsymbol{K}^{\mathrm{Text}}, \boldsymbol{V}^{\mathrm{Text}}). \tag{5}$$

The results of this cross-modal MHSA are denoted as $\{\boldsymbol{z}'_n\}_n$. In essence, visual features and linguistic features alternatively guide the self-attention mechanism for each other. This approach effectively utilizes textual guidance to have feature representation with best quality.

*3) Linguistics-driven channel attention:* In the previous stage, the fusion mainly targets spatial dimensions. For further performance improvement, we integrate guidance from linguistic features through a channel attention driven by these features. Note that deep feature maps of an image correlate

differently with the object of interest. For instance, curvy features may be more closely related to curly hair compared to straight features. Hence, linguistic features can effectively select relevant feature maps. This selection process is achieved via channel attention. Specifically, the linguistic features are fed into an FC layer with a Sigmoid activation function. The output is reshaped to generate the weights for adjusting the significance of each channel in the visual features. Afterward, these reshaped re-calibrated features are fed to the subsequent modules for predicting the alpha matte.

### B. Alpha Matte Prediction

With the fused feature tensor in hand, it is input into a decoder implemented via a CNN with up-sampling layers. This decoder generates an initial estimate of the alpha matte, denoted as $\boldsymbol{\alpha}_0$. While this estimate effectively localizes the target object, its accuracy in representing opacity values might be compromised due to potential information loss during feature fusion. To counter this, we introduce a second CNN dedicated to refining the alpha matte estimate. This refinement CNN uses a combination of the initial alpha matte and the original image as input, producing another alpha matte $\boldsymbol{\alpha}_1$ that is more aligned with the input image. The input text description is re-engaged as channel attention for the intermediate features. The original text features are converted into attention weights using a FC layer and then applied to the feature maps. Empirically, this channel attention offers a moderate improvement since the text description generally only contains limited information about detailed optical transparency.

### C. Training and Inference

Recall that our NN generates two alpha matte estimates: $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$. Given a ground-truth alpha matte $\boldsymbol{\alpha}_{\text{gt}}$, the training loss is defined for both $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ as follows:

$$\mathcal{L}_{\text{predict}} := \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_{\text{gt}}\|_1 + \lambda\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_{\text{gt}}\|_1, \quad \lambda \in \mathbb{R}^+. \tag{6}$$

Furthermore, we utilize an RIS dataset to create an auxiliary task, aiming at improving the learning of rough alpha matte prediction. This scheme mitigates possible over-fitting, especially when the number of training samples for text-driven matting is limited. Let $\boldsymbol{\beta}_{\text{gt}}$ denote the ground-truth segmentation mask corresponding to a pair of an image and a text description. We then define the auxiliary training loss using cross-entropy as follows:

$$\mathcal{L}_{\text{aux}} := -\sum_i \boldsymbol{\beta}_{\text{gt}} \log \boldsymbol{\beta}_1(i) + (1-\boldsymbol{\beta}_{\text{gt}}(i)) \log\left(1 - \boldsymbol{\beta}_1(i)\right), \tag{7}$$

where $\boldsymbol{\beta}_1$ denotes the binary version of $\boldsymbol{\alpha}_1$, obtained by applying a threshold of $0.5$. Essentially, we anticipate that the initial alpha matte prediction will accurately localize the objects of interest, aligning with the ground-truth segmentation mask. The total loss is then given by

$$\mathcal{L} := \mathcal{L}_{\text{predict}} + \gamma\mathcal{L}_{\text{aux}}, \quad \gamma \in \mathbb{R}^+. \tag{8}$$

During inference, an image and a corresponding text description of the target of interest are input into the NN, from which we derive an alpha matte. Subsequently, the foreground is extracted using a multiplication-based method, a common approach in existing literature such as [25], [34], [57].

## IV. Experiments

### A. Training Data and Implementation Details

The proposed method is evaluated on portrait matting. Since no public dataset for text-guided portrait matting exists, we constructed a training dataset by combining samples from three existing related datasets listed below. (i) SC (synthetic-composite) Adobe dataset [57]: From this popular matting dataset, we select 269 portrait foregrounds of 431 general foregrounds. As this dataset does not provide text descriptions, we generate a text description for each foreground object using an autonomous image captioning tool [7], followed by re-organization and manual corrections for low-quality captions. See Fig. 4 as well as supplemental material for some examples of these text annotations. During training, we composited input images from two different portrait foregrounds and one background, using one portrait's description to predict its alpha matte. (ii) Human2K dataset [32]: This large-scale portrait image matting dataset includes 2000 distinct foreground portraits. Similar to the procedure for processing SC Adobe dataset, we annotated text descriptions and composited training images. (iii) RefCOCO RIS dataset [62]: Used solely for auxiliary training loss $\mathcal{L}_{\text{aux}}$, it features 19,994 images with 142,209 referring expressions for 50,000 objects (humans and non-humans), including ground-truth binary segmentation masks. For each image in training, one object and its description were randomly selected as input. All the training/test data will be released together with the code for re-producible research upon the paper's acceptance.

In all the experiments, the parameters $\lambda, \gamma$ in the training loss are set to $1$ and $0.5$, respectively. We employ the DarkNet [43] as the image encoder, the Tok2Vec function provided by the spaCy library as the text encoder, and the MaskNet [63] as the NN for refining the alpha matte. The pre-trained models of DarkNet and MaskNet are utilized for initialization. The entire NN is trained using the Adam optimizer with a learning rate of $10^{-4}$ over $50$ epochs. For ease of reference, we name the proposed method TIM (Text-based Image Matting). Four quantitative metrics are employed for performance evaluation: SAD (Sum of Absolute Differences), MSE (Mean of Squared Error), Grad (Gradient) and Conn (Connectivity). These metrics are calculated by comparing the estimated alpha mattes with the ground-truth mattes at their original sizes.

### B. Evaluation on SC Adobe Dataset

We assessed the performance of our trained TIM model on the test set of the SC Adobe dataset. In line with the methodology in [57], we chose $11$ portrait foregrounds from the test set and combined them with 20 randomly selected backgrounds, yielding a total of 220 test images. The text descriptions for these images were generated in the same manner as for the training data, as detailed in Section IV-A.

Since text-guided image matting is a new topic, there are no methods of this topic for comparison. To ensure a comprehensive evaluation, we selected various methods for comparison: (i) IM [34] and GCA [25], two prominent trimap-based methods that use a pair of image and trimap as input; (ii)

(a) “a woman with glasses and flowing hair holding a cup of coffee”
(b) “a man wearing glasses with a black hair”
(c) “an old man wearing a Santa hat”
(d) “a beautiful woman with roses on her long hair”
(e) “a young woman with black hair and white shirt”
(f) “a woman with white hair wearing black shirt”
(g) “a woman in a white dress with wings”
(h) “a woman in pink with big hair seeing her cell phone”
(i) “a woman in a blue dress holding a blue dress”
(j) “a children in golden hair and blue shirt” (k) “a girl with a ponytail”
(l) “a brown-curly-haired woman in a denim jacket”

Fig. 4. Examples of our text annotations and corresponding foreground images (combined with background images to form training samples).

TABLE I
EVALUATION ON PORTRAIT IMAGES OF SC ADOBE DATASET. THE BEST (SECOND BEST) RESULTS ARE BOLDFACED (UNDERLINED).

| Method | Input | SAD | MSE | Grad | Conn |
|---|---|---|---|---|---|
| IM [34] | Trimap | 13.86 | 0.0015 | 7.91 | 12.70 |
| GCA [25] | Trimap | **12.49** | 0.0021 | 4.53 | 10.83 |
| DMP [58] | Background | 15.63 | 0.0017 | 15.05 | 13.68 |
| BGMv2 [26] | Background | 16.46 | 0.0019 | 11.86 | 14.82 |
| MGM [63] | Mask | 13.35 | **0.0007** | **4.40** | **7.74** |
| C2F [42] | Mask | 16.13 | 0.0027 | 16.73 | 19.27 |
| LF [64] | Free | 25.90 | 0.0030 | 11.50 | 19.94 |
| AIM [21] | Free | 181.66 | 0.0684 | 66.65 | 181.69 |
| MODNet [17] | Free | 79.60 | 0.0233 | 33.30 | 78.26 |
| VLT [9] | Text | 149.69 | 0.0405 | 58.41 | 147.95 |
| TIM [Ours] | Text | 14.34 | 0.0009 | 4.47 | 8.70 |

DMP [58] and BGMv2 [26], two background-based methods requiring a pair of image and background as input; (iii) MGM [63] and C2F [42], two mask-based methods that utilize the foreground regions of trimaps provided by the dataset as masks;[1] (iv) LF [64], AIM [21] and MODNet [17], three guidance-free methods that only requires an image; and (v) VLT [9], an RIS method we adapted for our task by fine-tuning its published pre-trained model with our image matting training data (including our text annotations), using the cross-entropy loss (instead of $\ell_1$ loss for optimal results). For all trimap-based and mask-based models, whenever feasible, we re-trained them using our portrait data constructed from Adobe and Human2K datasets for a fair comparison. For further performance improvement, their published pre-trained models were used for initialization.

[1]Originally designed for video matting, C2F takes a rough mask and a video frame as input. For image matting, we adapt it to process a single frame, using binarization on the ground-truth alpha matte for a suitable mask.

TABLE II
EVALUATION ON PORTRAIT IMAGES OF HUMAN2K DATASET.

| Method | Input | SAD | MSE | Grad | Conn |
|---|---|---|---|---|---|
| IM [34] | Trimap | 9.69 | 0.0006 | 10.56 | 9.04 |
| GCA [25] | Trimap | 5.55 | 0.0002 | 3.45 | 4.36 |
| TIMINet [32] | Trimap | 4.76 | 0.0002 | 2.37 | 3.24 |
| MGM [63] | Mask | 10.26 | 0.0004 | 5.37 | 5.92 |
| C2F [42] | Mask | 21.62 | 0.0013 | 24.41 | 15.12 |
| LF [64] | Free | 38.68 | 0.0041 | 33.82 | 32.62 |
| AIM [21] | Free | 193.87 | 0.0429 | 82.00 | 192.73 |
| VLT [9] | Text | 173.47 | 0.0223 | 96.02 | 168.51 |
| TIM | Text | 11.08 | 0.0006 | 6.01 | 7.02 |

The quantitative results are listed in Table I for performance comparison of different methods. Notably, TIM's performance is second-best in all metrics except SAD, closely trailing the top performer, MGM, which is a mask-guided method utilizing a well-annotated trimap's foreground area as the input mask. This outcome is understandable considering that TIM relies on a brief text description rather than a precise mask. Nevertheless, TIM outperforms another mask-based method, C2F, as well as various trimap-based, background-based, and the guidance-free ones, in three metrics. When compared to VLT, TIM demonstrates significantly better results. This is attributed to the harder challenge of text-guided image matting, which demands precise alpha matte estimation, a task more complex than RIS.

These results highlight the feasibility and superior performance of text-guided image matting. Visual examples shown in Fig. 5 also demonstrate TIM's effectiveness, not only in correctly identifying the subject in the first sample but also in capturing semi-transparent parts with rich details, as evidenced in all three samples.

### C. Evaluation on Human2K Dataset

TIM's performance is also evaluated on the Human2K test set, which contains 100 different portrait foregrounds. We utilized the 2000 evaluation images provided by [32], composed of these 100 portrait foregrounds and 20 randomly chosen backgrounds. The text descriptions were generated similarly to the training data, as outlined in Section IV-A. For comparison, we used most of the methods from the previous subsection, including IM [34], GCA [25], MGM [63], C2F [42], LF [64], AIM [21] and VLT [9]. In addition, we quote the results of TIMINet [32] for comparison. Background matting methods were excluded as the Human2K test set does not provide backgrounds. For a fair comparison, all models pre-trained on the SC Adobe training set were fine-tuned using our constructed portrait matting dataset.

See Table II for the quantitative results. The trimap-based methods perform quite well, largely due to the high accuracy of trimaps in the test set; see *e.g.* the second column in Fig. 6

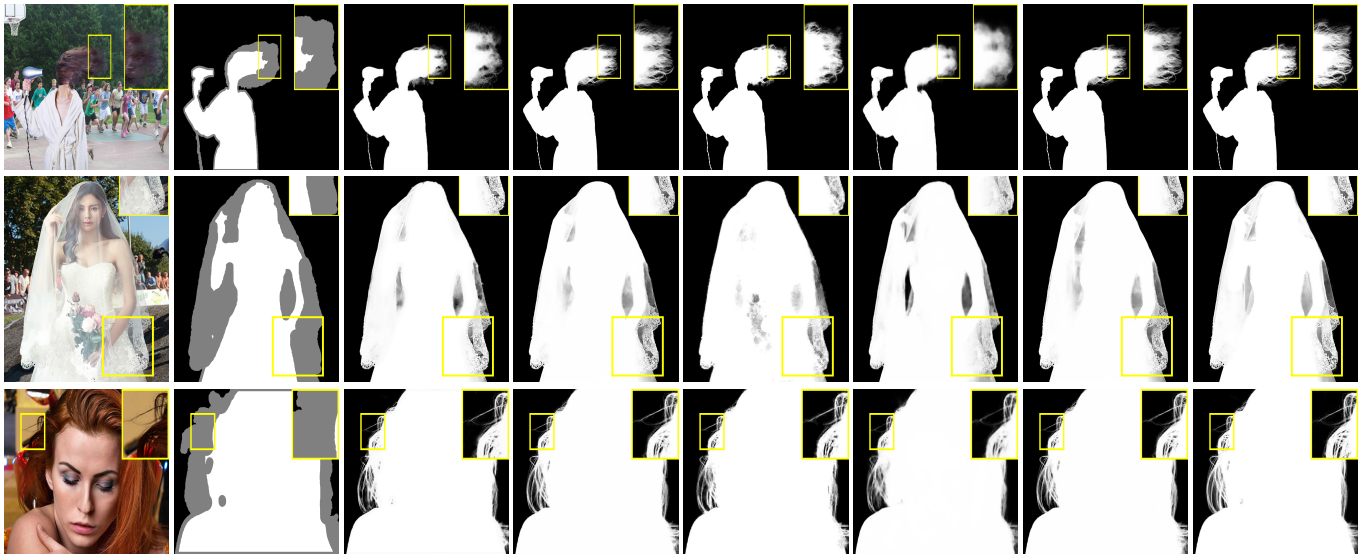



Fig. 5. Alpha mattes predicted by different methods on images from SC Adobe dataset. Descriptions for each row, from top to bottom are "a woman in white bathrobe blowing her long hair", "a woman with a wedding dress", and "a girl with red hair", respectively.

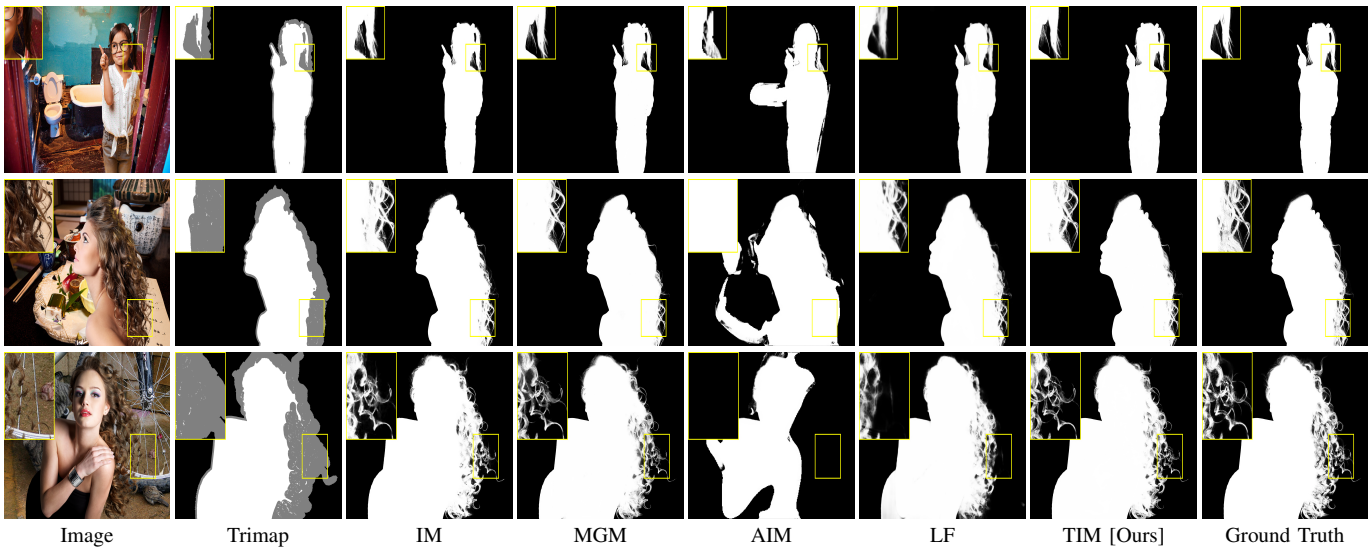



Fig. 6. Alpha mattes predicted by different methods on images from Human2K dataset. Description for each row, from top to bottom are "a cute little girl wearing glasses pointing", "a beautiful brunette woman with curly hair", and "a young woman posing with long curly hair", respectively.

versus that in Fig. 5. The trimap-based methods assume the trimaps are accurate and directly project known areas into final results, making it challenging for mask-guided methods like MGM, and TIM to surpass trimap-based methods when trimap quality is high. Although TIM's performance trails MGM, it is comparable to or even surpasses IM, a trimap-based method, and significantly outperforms C2F, another mask-guided method. This demonstrated the effectiveness of TIM in utilizing text guidance for high-quality alpha matte prediction. Compared to guidance-free methods like LF and AIM, TIM shows superior performance, balancing better results with greater user-friendliness than obtaining trimaps. TIM also notably outperforms VLT. See Fig. 6 for some visual results, where AIM includes non-portrait regions into the alpha mattes. In contrast, TIM accurately captures targets with richer details, showing competitive accuracy compared to IM and MGM.

TABLE III
RESULTS OF ABLATION STUDIES.

| TIM | SC Adobe dataset | | | | Human2K dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | SAD | MSE | Grad | Conn | SAD | MSE | Grad | Conn |
| Original | 14.34 | .0009 | 4.47 | 8.70 | 11.08 | .0006 | 6.01 | 7.02 |
| w/o MLFF | 16.52 | .0016 | 4.62 | 10.92 | 12.74 | .0008 | 6.50 | 8.24 |
| SLFF | 15.74 | .0014 | 4.56 | 9.87 | 12.19 | .0007 | 6.23 | 7.71 |
| w/o CMSA | 15.96 | .0013 | 4.81 | 10.38 | 13.01 | .0009 | 6.69 | 8.57 |
| w/o LDCA | 18.15 | .0018 | 5.15 | 12.81 | 12.71 | .0008 | 6.34 | 8.15 |
| w/o Refine | 82.48 | .0137 | 47.78 | 70.71 | 99.67 | .0115 | 82.72 | 90.77 |
| w/o $\mathcal{L}_{\text{aux}}$ | 21.45 | .0023 | 7.89 | 15.61 | 17.24 | .0016 | 9.49 | 10.22 |

### D. Evaluation on Real-World Data

The generalization performance of TIM was assessed on some real portrait images sourced from the Internet. We


Fig. 7. Comparison of matting results from different methods on real-world images (cropped and resized for clarity) featuring multiple people. Each sample uses two descriptions $\mathcal{D}1$ and $\mathcal{D}2$. First image: $\mathcal{D}1$ is "a woman on right in a wedding dress" and $\mathcal{D}2$ is "a man on the left". Second image: $\mathcal{D}1$ is "the man on the far right in a blue suit." and $\mathcal{D}2$ is "the man on the far left". Third image: $\mathcal{D}1$ is "the woman on the right" and $\mathcal{D}2$ is "the woman on the left". Trimap inputs for IM are provided in the supplemental material.

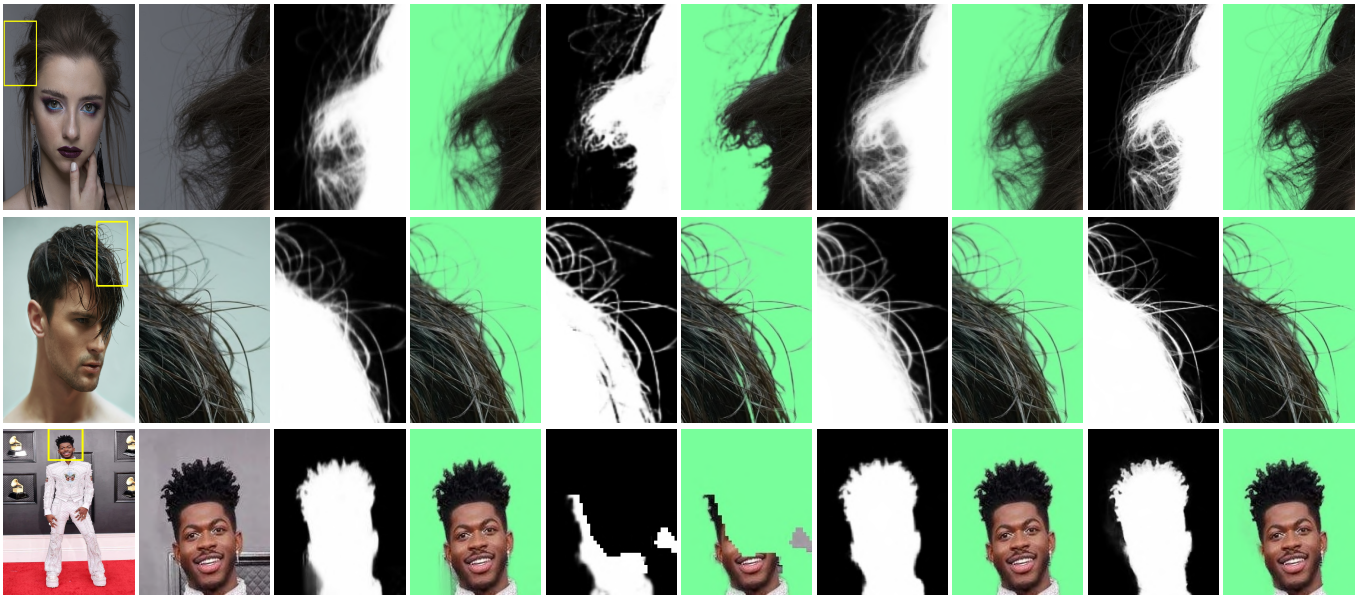



Fig. 8. Comparison of matting results of different methods on real images of single human subject. First image: "the girl with earrings". Second image: "the naked man". Third image: "a man with white clothes".

collect 20 images from Google, each of which contains one or multiple people. The text descriptions are annotated by the same way as that in Section IV-A. Fig. 7 displayed the results. For images with multiple persons, we annotated two descriptions for different individuals, as noted in the figure captions, and computed the results separately. The extracted foreground and alpha matte were then combined with a green background, offering an alternative accuracy check.

TIM consistently identifies the correct target with precise details, outperforming guidance-free methods that often fail to distinguish the intended subject. For instance, most methods wrongly include all individuals in the second image, while only MODNet responds to the first and third images. TIM, however, accurately captures the designated person in all samples, including semi-transparent elements like the headdress in the first image and the hair in the third image. Additional testing was conducted using real images with single subjects from the internet. Fig. 8 illustrates these results, where TIM shows competitive performance.

### *E. Ablation Studies*

Recall that there are three stages of the text-guided feature fusion in TIM. To evaluate the effectiveness of each stage, we established three baselines: "w/o MLFF (multi-level feature fusion)", "w/o CMSA (cross-modality self-attention-based fusion)" and "w/o LDCA (linguistics-driven channel


Fig. 9. Visual inspection of results in ablation studies.

attention)". Additionally, we created three more baselines for further analysis: "SLFF (single-level feature fusion)", "w/o $\mathcal{L}_{\text{aux}}$" and "w/o Refine". In setting up these baselines, we increased the channel number of the image encoder to ensure the baseline models are comparable in size to the original TIM for a fair comparison.

The baselines are defined as follows. (i) w/o MLFF: Multi-level feature fusion is substituted by up-sampling and convolving the image encoder's final output features (for dimension consistency) without including the linguistic features from the text encoder; (ii) SLFF: This baseline further modifies the previous one by fusing linguistic features from the text encoder using an FC+Tanh layer; (iii) w/o CMSA: Cross-modality self-attention-based fusion is omitted, and the output from the preceding stage is fed directly to the subsequent stage; (iv) w/o LDCA: Linguistics-driven channel attention is removed, and the output of the preceding stage is passed straight to the alpha matte prediction module; (v) w/o $\mathcal{L}_{\text{aux}}$: The auxiliary loss is disabled, and consequently, training data from the RefCOCO RIS dataset is not utilized; and (vi) w/o Refine: The alpha matte refinement module is excluded, and the output from the alpha matte decoder is taken as the final result.

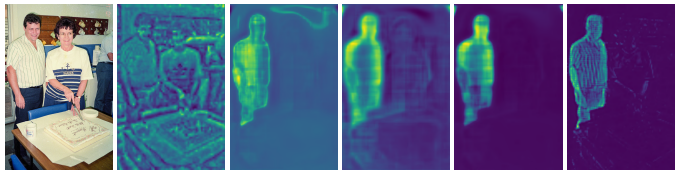



Fig. 10. Visualization of feature maps output at different modules. (a) Input image with text "man on the left"; (b) image encoder output; (c) MLFF output ; (d) CMSA output; (e) LDCA output. (f) Coarsest scale feature map from refinement step..

See Table III for the baseline results. Each of the three stages in the text-guided feature fusion significantly enhances quantitative performance across all four metrics. The auxiliary task defined by $\mathcal{L}_{\text{aux}}$ also plays a critical role. Given our limited size of text-based matting training data, the loss $\mathcal{L}_{\text{aux}}$ utilizes the RefCOCO dataset to counter possible over-fitting, thereby improving the learning of text-based object localization. As anticipated, the alpha matte refinement stage is the most impactful. Without it, TIM's performance significantly declines, leading to the loss of many details in the predicted mattes. This is primarily because the auxiliary training loss prompts TIM's first stage to concentrate on object localization with roughly estimated alpha values, making the refinement stage essential for achieving accurate and detailed results. See Fig. 9 for some visual results, which demonstrate how each module enhances qualitative performance, particularly in transparency details and object localization. Notably, CSMA and $L_{\text{aux}}$ are more crucial for accurately locating the target object. In contrast, MLFF, LDCA and the refinement step contribute more to the details in the predicted alpha matte. Overall, each part of TIM is essential to its effectiveness.

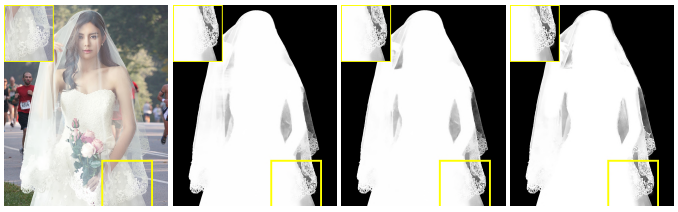



Fig. 11. Matting results of TIM using two different text descriptions: ($\mathcal{D}1$) "a woman with a wedding dress"; and ($\mathcal{D}2$) "a female with a white veil".

To more clearly illustrate the role of each module, Fig. 10 visualizes the intermediate features generated at various stages. It is evident that the modules do not distinctly exhibit separate functions such as part recognition or ordering. Instead, they collectively and progressively improve the localization of the target portrait.

### F. More Analysis

We evaluated the robustness of TIM to text descriptions. In Fig. 11, TIM processes two distinct descriptions for the same image separately. The results, interestingly, are remarkably


Fig. 12. Matting results of TIM using different text descriptions: ($\mathcal{D}1$) "a woman"; ($\mathcal{D}2$) "a woman with long hair"; ($\mathcal{D}3$) "a woman in a white bathrobe"; ($\mathcal{D}4$) "a man blowing his long hair".


Fig. 13. Matting results of TIM on RefCOCO. The text inputs are (a) "left woman"; (b) "man in suit and tie"; (c) "guy beard"; (d) "a man on the phone"; (e) "a woman wearing a blue white top"; (f) "right girl"; (g) "a cute baby". The text inputs are directly quoted from RefCOCO if available.

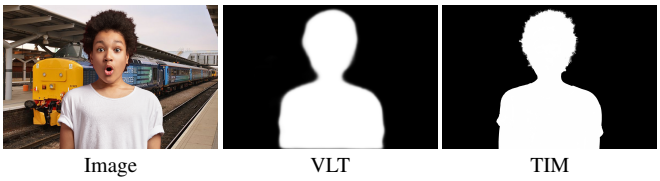



Fig. 14. Visual results of TIM and VLT. The text input is "African American boy with black curly hair".


Fig. 15. Matting results on video frames. The given text description is "the woman on left with long hair".

similar and highly accurate. In Fig. 12, we further evaluate the performance with varied and somewhat imprecise text descriptions. It shows that TIM's results are not overly sensitive to minor inaccuracies in descriptions. Additionally, Fig. 13 shows the results of TIM on the samples of the RefCOCO dataset, demonstrating its effectiveness even with cross-dataset images and text descriptions.

Text-guided matting bears resemblance to RIS. Fig. 14 shows a comparison between the results of TIM and VLT [9]. VLT was initially developed for RIS and produces coarser matting results compared to TIM. This is primarily attributed to the absence of a refinement stage and the use of a resizing strategy in VLT for computational feasibility. In comparison, TIM is capable of capturing finer details.

### *G. Batch Processing on Video Frames*

We also apply TIM to video matting. When a text description remains consistent through a video clip, matting can be done with only one description. This is often the case for an object with consistent description through a video clip. See Fig. 15 for a demonstration on the frames from a real-word video collected from Internet, compared to two recent video matting methods including MODNet [17], RHVM [27] and C2F [42]. While these three methods focus on all persons in a video frame, limiting their flexibility for targe-specific matting, TIM can identify the person of interest with varying poses in a video clip, requiring only a single text description. This demonstrates TIM's significant potential for video matting.

Although our TIM is designed for single-image input and does not yet exploit temporal cues from videos, leaving room for accuracy enhancement, TIM still delivers commendable matting results. In Fig. 16 and Fig. 17, we present TIM's results on video frames from [44], featuring single persons. As these frames include corresponding backgrounds, BGMv2 [26]

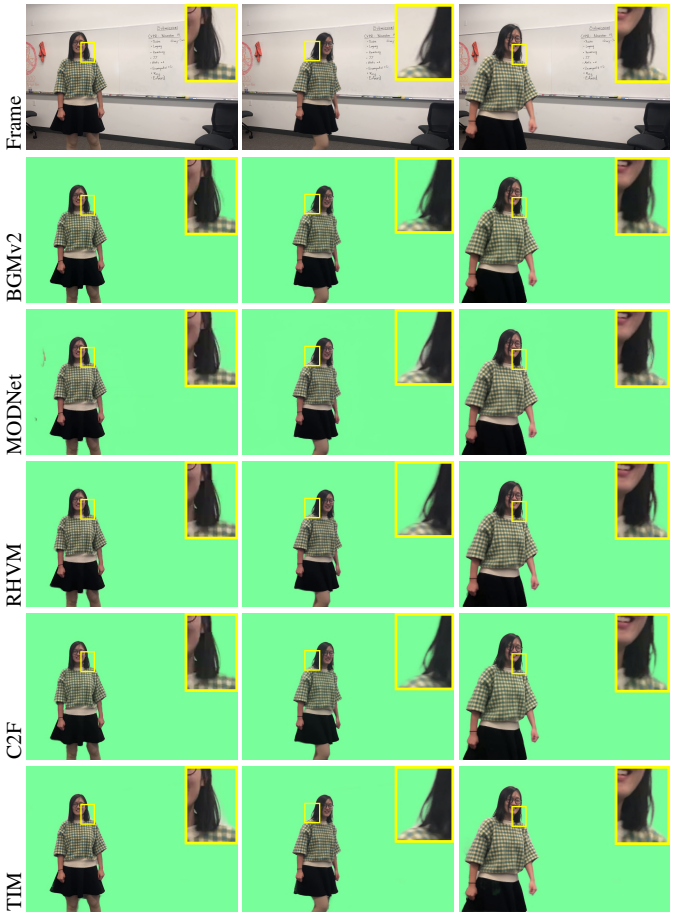



Fig. 16. Matting results on cropped video frames. The given text description is "a woman with glasses and skirt".

is also used for comparison. Even in the case of single human subjects, the results of TIM remain competitive against other methods.

## V. Conclusion

This paper introduced a text-guided image matting approach, providing a novel alternative to existing paradigms with the added benefit of semantic clues. Implemented through a deep NN featuring cross-modal fusion and a two-step prediction process, the proposed approach is more user-friendly than background-based methods and more efficient in batch processing than both trimap-based and background-based methods, while delivering competitive performance. Its potential for video matting is also notable.

There are several potential extensions of this paper that offer promising practical advantages. Firstly, the focus on portrait matting could be broadened to encompass general objects with transparency. Secondly, there is a lot of room for further improving the handling of real-world images through model adaptation. Thirdly, incorporating temporal cues to boost the accuracy of video matting represents an interesting research problem for the advancement of video matting. Fourthly, the development of more interpretable NNs could be instrumental in mitigating over-fitting. Lastly, it is very promising to merge textual semantic cues with spatial guidance from clicks,

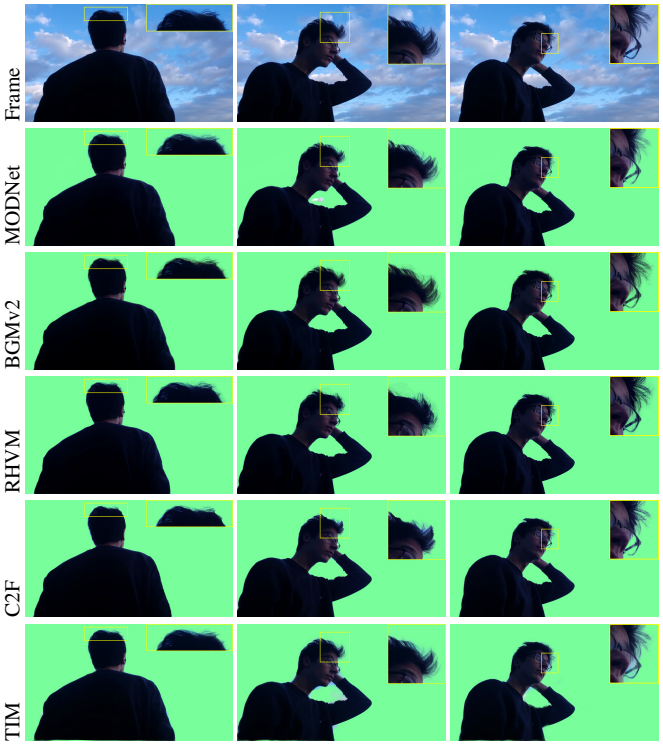



Fig. 17. Matting results on cropped video frames. The given text description is "a man with glasses".

potentially leading to a more refined cross-modal self-attention mechanism and improved positional embedding.

## Acknowledgments

We would to thank Jiatong Huang and Xinchen Liu for their help on the dataset preparation.

## References

[1] N. Aafaq, A. Mian, W. Liu, N. Akhtar, and M. Shah, "Cross-domain modality fusion for dense video captioning," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 763–777, 2021.

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

[3] H. Cai, F. Xue, L. Xu, and L. Guo, "Transmatting: Enhancing transparent objects matting with transformers," in *Proceedings of European Conference on Computer Vision*. Springer, 2022, pp. 253–269.

[4] 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 IEEE International Conference on Computer Vision*, 2019, pp. 8819–8828.

[5] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, October 2019.

[6] X. Chen, D. Zou, S. Zhiying Zhou, Q. Zhao, and P. Tan, "Image matting with local and nonlocal smooth priors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1902–1907.

[7] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 162–11 173.

[8] H. Ding, S. Cohen, B. Price, and X. Jiang, "Phraseclick: toward achieving flexible interactive segmentation by phrase and click," in *Proceedings of European Conference on Computer Vision*. Springer, 2020, pp. 417–435.

[9] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 16 321–16 330.

[10] ——, "Vlt: Vision-language transformer and query generation for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7900–7916, 2023.

[11] H. Ding, H. Zhang, C. Liu, and X. Jiang, "Deep interactive image matting with feature propagation," *IEEE Transactions on Image Processing*, vol. 31, pp. 2421–2432, 2022.

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

[13] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proceedings of European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 108–124.

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

[15] 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.

[16] L. Karacan, A. Erdem, and E. Erdem, "Alpha matting with kl-divergence-based sparse sampling," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4523–4536, 2017.

[17] 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 AAAI Conference on Artificial Intelligence*, 2022.

[18] U.-H. Kim, Y. Hwang, S.-K. Lee, and J.-H. Kim, "Writing in the air: Unconstrained text recognition from finger movement using spatio-temporal convolution," *IEEE Transactions on Artificial Intelligence*, 2022.

[19] Y. Lee and S. Yang, "Parallel block sequential closed-form matting with fan-shaped partitions," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 594–605, 2017.

[20] 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.

[21] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," in *Proceedings of International Joint Conference on Artificial Intelligence*, 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

[22] ——, "Referring image matting," *arXiv preprint arXiv:2206.05149*, 2022.

[23] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[24] X. Li, K. Liu, Y. Dong, and D. Tao, "Patch alignment manifold matting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 3214–3226, 2018.

[25] Y. Li and H. Lu, "Natural image matting via guided contextual attention," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 450–11 457.

[26] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8762–8771.

[27] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2022, pp. 238–247.

[28] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proceedings of IEEE International Conference on Computer Vision*, Oct 2017.

[29] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8560–8569.

[30] Y. Liu, D. Dai, X. Tang, S. Xia, and G. Wang, "Bi-lstm sequence modeling for on-the-fly fine-grained sketch-based image retrieval," *IEEE Transactions on Artificial Intelligence*, 2022.

[31] Y. Liu, J. Xie, Y. Qiao, Y. Tang, and X. Yang, "Prior-induced information alignment for image matting," *IEEE Transactions on Multimedia*, 2021.

[32] Y. Liu, J. Xie, X. Shi, Y. Qiao, Y. Huang, Y. Tang, and X. Yang, "Tripartite information mining and integration for image matting," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 7555–7564.

[33] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings ofAnnual Meeting of the Association for Computational Linguistics*, 2018, pp. 2247–2256.

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

[35] L. Lu, W. Zhai, H. Luo, Y. Kang, and Y. Cao, "Phrase-based affordance detection via cyclic bilateral interaction," *IEEE Transactions on Artificial Intelligence*, 2022.

[36] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[37] Y. Niu, P. Liu, T. Zhao, and Y. Fan, "Matting-based residual optimization for structurally consistent image color correction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3624–3636, 2019.

[38] 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.

[39] Z. Pei, X. Chen, and Y.-H. Yang, "All-in-focus synthetic aperture imaging using image matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 2, pp. 288–301, 2016.

[40] S. M. Prabhu and A. Rajagopalan, "Natural matting for degraded pictures," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3647–3653, 2011.

[41] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020, pp. 13 673–13 682.

[42] A. Rao, L. Xu, Z. Li, Q. Huang, Z. Kuang, W. Zhang, and D. Lin, "A coarse-to-fine framework for automatic video unscreen," *IEEE Transactions on Multimedia*, 2022.

[43] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013–2016.

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

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

[46] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Proceedings of European Conference on Computer Vision*, September 2018.

[47] A. R. Smith and J. F. Blinn, "Blue screen matting," in *Proceedings of Annual Conference on Computer Graphics and Interactive Techniques*, 1996, pp. 259–268.

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

[49] 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 (CVPR)*, June 2021, pp. 11 120–11 129.

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

[51] E. S. Varnousfaderani and D. Rajan, "Weighted color and texture sample selection for image matting," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4260–4270, 2013.

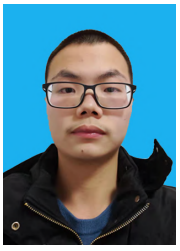
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

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

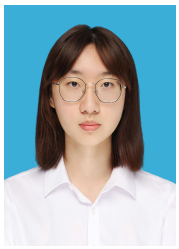
[54] Y. Wang, Y. Niu, P. Duan, J. Lin, and Y. Zheng, "Deep propagation based image matting." in *Proceedings of International Joint Conference on Artificial Intelligence*, vol. 3, 2018, pp. 999–1006.

[55] T. Wei, D. Chen, W. Zhou, J. Liao, H. Zhao, W. Zhang, and N. Yu, "Improved image matting via real-time user clicks and uncertainty estimation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 374–15 383.
[56] 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, 2014.
[57] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2970–2979.
[58] 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*, pp. 1–1, 2021.
[59] L. Ye, Z. Liu, and Y. Wang, "Dual convolutional lstm network for referring image segmentation," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3224–3235, 2020.
[60] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.
[61] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.
[62] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 69–85.
[63] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 1154–1163.
[64] 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 IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7469–7478.
[65] W. Zhao, D. Zhou, B. Cao, K. Zhang, and J. Chen, "Adversarial modality alignment network for cross-modal molecule retrieval," *IEEE Transactions on Artificial Intelligence*, 2023.
[66] Y. Zheng, Y. Yang, T. Che, S. Hou, W. Huang, Y. Gao, and P. Tan, "Image matting with deep gaussian process," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[67] 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, 2021.
[68] Q. Zhu, L. Shao, X. Li, and L. Wang, "Targeting accurate object extraction from an image: A comprehensive study of natural image matting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 185–207, 2014.

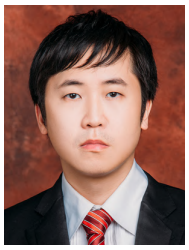
**Yong Xu** received the B.S., M.S., and Ph.D. degrees in mathematics from Nanjing University, Nanjing, China, in 1993, 1996, and 1999, respectively. He was a Postdoctoral Research Fellow of computer science with the South China University of Technology, Guangzhou, China, from 1999 to 2001, where he became a Faculty Member and is currently a professor with the School of Computer Science and Engineering. He is the Dean of Guangdong Big Data Analysis and Processing Engineering & Technology Research Center. His current research interests include computer vision, pattern recognition, image processing, and big data. He is a senior member of the IEEE Computer Society and the ACM. He has received the New Century Excellent Talent Program of MOE Award.



**Xin Yao** is currently a M.Sc candidate of Computer Science at South China University of Technology. He is working on image processing and computer vision.



**Baoling Liu** is currently a M.Sc candidate of Computer Science at South China University of Technology. She is working on machine learning and image processing.



**Yuhui Quan** received the Ph.D. degree in Computer Science from South China University of Technology in 2013. He worked as the postdoctoral research fellow in Mathematics at National University of Singapore from 2013 to 2016. He is currently the associate professor at School of Computer Science and Engineering in South China University of Technology. His research interests include computer vision, image processing and sparse representation.

**Hui Ji** received his B.Sc. degree in Mathematics from Nanjing University, China, and his Ph.D. in Computer Science from the University of Maryland, College Park, Maryland, United States. He joined the National University of Singapore as an assistant professor in 2006. Currently, he is a professor in applied mathematics at the same university. His research interests include computational harmonic analysis, imaging science, machine learning, and computational vision.