Tripartite Information Mining and Integration for Image Matting

Yuhao Liu[1,*], Jiake Xie[2,*], Xiao Shi[3], Yu Qiao[1], Yujie Huang[4], Yong Tang[2,†] Xin Yang[1,†]
[1] Dalian University of Technology, [2] PicUp.AI
[3] Agricultural Information Institute of CAAS, [4] Fudan University
{yuhaoLiu7456, coachqiao2018}@gmail.com, {jxie, yt}@picup.ai
sixiaosmile@outlook.com, 19112020091@fudan.edu.cn, xinyang@dlut.edu.cn

Figure 1: Visual comparisons of our method and CAM [17] on the Real-World images. Both results are produced from the model trained on DIM [56] dataset. Please zoom in to see the fine details.

Abstract

*With the development of deep convolutional neural networks, image matting has ushered in a new phase. Regarding the nature of image matting, most researches have focused on solutions for transition regions. However, we argue that many existing approaches are excessively focused on transition-dominant local fields and ignored the inherent coordination between global information and transition optimisation. In this paper, we propose the Tripartite Information Mining and Integration Network (**TIMI-Net**) to harmonize the coordination between global and local attributes formally. Specifically, we resort to a novel 3-branch encoder to accomplish comprehensive mining of the input information, which can supplement the neglected coordination between global and local fields. In order to achieve effective and complete interaction between such multi-branches information, we develop the Tripartite Information Integration (TI^2) Module to transform and integrate the interconnections between the different branches. In addition, we built a large-scale human matting dataset (Human-2K) to advance human image matting, which consists of 2100 high-precision human images (2000 images for training and 100 images for test). Finally, we conduct extensive experiments to prove the performance of our proposed **TIMI-Net**, which demonstrates that our method performs favourably against the SOTA approaches on the alphamatting.com (**Rank First**), Composition-1K (**MSE-0.006, Grad-11.5**), Distinctions-646 and our Human-2K. Also, we have developed an online evaluation website to perform natural image matting.*

1. Introduction

The digital matting is one of the important tasks in computer vision, which aims to accurately estimate the opacity of foreground objects in images and video sequences. It has a wide range of applications, particularly in the fields of

*Joint first authors. †Joint corresponding authors. Project page: https://wukaoliu.github.io/TIMI-Net.

film production and digital image editing. Formally, the input image is modeled as a linear combination of foreground and background colours [41], as shown below:

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

where $\alpha_i \in [0, 1]$ denotes the opacity at pixel i in the input image, F_i and B_i refer to the Foreground (Fg) and Background (Bg) colour. The problem is highly ill-posed. As for a pixel in a typical *3*-channel (*e.g.* RGB) image, *7* unknown values (*i.e. 3* F, *3* B and *1* $alpha$) need to be solved, but there are only *3* known quantities (*3* I).

To solve the problem, the classical methods [7, 23, 46, 52] utilized trimap as a kind of constraint information to reduce the solution space. The trimap consists of three parts, white, black, and gray, representing the foreground, background and transition regions, separately. Many methods [2, 6, 31, 38, 43, 62] attempted to go about predicting alpha mattes without auxiliary cues. Although they can produce promising results, the gap between real-world and synthetic images remains and can be magnified due to composited artefacts, which can lead to inferior generalisation.

Typically, almost all trimap-based methods [56, 34, 49, 3, 17] perform transition optimisation explicitly by concatenating RGB images and trimap to feed the network. Other trimap-free methods [6, 62, 31] construct some transition variants (pseudo-trimap) implicitly with the assistance of loss functions to guide local region optimisation. However, within the constraints of the transition, these approaches focus excessively on local regions, which may neglect the coordination between global and local attributes (*e.g.* texture similarity, location correlation, *etc.*), thus leading to incomplete information mining.

In this paper, we propose a Tripartite Information Mining and Integration Network (TIMI-Net) that can capture sufficient global information by mining and integrating multimodal information from RGB and Trimap. As for information mining, we consider that features from different modalities have complementary global information. RGB images can provide detailed low-level appearance (*e.g.* texture and colour similarity), while high-level positional relevance (semantics, shape, *etc.*) can be found in trimap. Therefore, we construct two functionally specific units (termed RGB unit and Trimap unit) to perform separate mining.

Regarding the information integration, we can intuitively add or concatenate features from multi-branches like [17, 62]. However, this would lead to incomplete integration due to the differentiated characteristics of different types of global information and initial local information. To combine them effectively, inspired by Non-Local [53], we have developed a Tripartite Information Integration (TI^2) module that transforms and integrates two streams of bilateral relations RGB-Trimap branch and RGB-Unit, RGB-Trimap branch and Trimap-Unit. In this way, global information can be employed to guide the propagation of local information, thus facilitating the coordination of the both.

Our major contributions can be summarized as follows:

- We propose a Tripartite Information Mining and Integration Network (**TIMI-Net**) with a Tripartite Information Integration (TI^2) module for image matting, which can sufficiently mine and integrate complementary global information from the RGB image and trimap.
- We build a large-scale human matting dataset with *2,000* training images and *100* test images. To the best of our knowledge, this is the largest high-accuracy human image matting dataset. We will open it to the public to advance the human image matting task.
- Experimental results demonstrate that the proposed **TIMI-Net** can achieve SOTA performance on synthetic and real-world images, which proves the effectiveness and superiority of the proposed method.

2. Related Work

In this section, we will briefly review the image matting from the three aspects: Traditional and Deep learning-based approaches, and Matting dataset.

Traditional methods that solved this ill-posed problem mainly rely on trimap and scribble constraint information, and they fall into two main categories: sampling-based methods and affinity-based methods. Sampling-based methods [9, 12, 13, 19, 40, 18, 52, 44, 51] collect a set of known foreground and background samples to find candidate colours for the foreground and background of a given pixel. Alpha mattes can then be calculated by applying a local smoothness assumption on the image statistics. Affinity-based methods [23, 1, 7, 22, 46, 15, 24] reconstruct Eq.1 so that it can propagate a known alpha value from known to unknown regions using the affinity of neighbouring pixels.

Deep Learning-based algorithms achieved great success on many tasks due to the advancement of deep convolution neural networks (*e.g.*, object detection [50], image restoration [54, 59, 55] and Specific Region Segmentation [35, 36]). In image matting, Shen *et al.* [45] firstly applied CNN in the portrait matting. DCNN [8] combined the results from [23] and [7], and fused them with a CNN to get the final alpha. For facilitating the end-to-end training, Xu *et al.* [56] proposed the first synthetic dataset and achieved fine performance. Later, Generative Adversarial Network (GAN) [14] was introduced by AlphaGAN [34] to improve the alpha mattes. Subsequently, a range of methods [49, 4, 17, 33, 26, 60, 32, 27, 10, 47] have made different improvements for acquiring better results. Adamatting [3] and CAM [17] explored the position information (semantic and shape) in trimap and the global context information

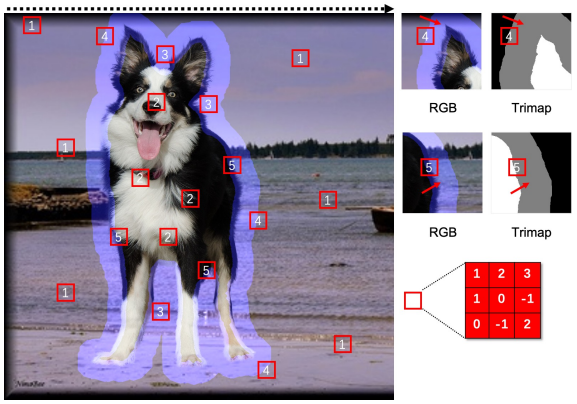

Figure 2: Classification of information transmission pattern. **1** and **2** mean that Convolution operation focus on known regions(Bg and Fg), **3** act on the pure unknown areas. **4** and **5** mean that information flows from Bg and Fg to the transition region, respectively. **Black** dashed line indicates the direction of convolution sliding. **Blue** region refers to the transitional areas. **Red** box shows a random set of 3×3 Convolution weight. For better visualization, we expand the width of unknown regions two times.

(colour and texture) from RGB. GCA [26] mimics image inpainting [61] to transmit context information from known regions, while HDMatt [60] takes the patch to learn cross-patch information between known and unknown areas. The global and local coordination is still ignored by them, although some of their methods seek some relevant information from the background area. In our method, we harmonize the coordination between global and local information, and then mine and integrate the local and complementary global information in a tripartite collaborative manner.

There are also many approaches [6, 58, 39, 57, 38, 31, 43, 28, 20, 25] that can acquire alpha mattes without using trimap as an additional constraint. Although sometimes pleasant results can be achieved, there are still some underlying problems, such as inferior generalisation to the real-world image due to the enlargement of the gap between the composited image and the real-world image, the inability of the user to select the region of interest, the requirement for other additional information (*e.g.* background, segmentation maps for other tasks, *etc.*). Hence, we focus on trimap-based image matting in this paper.

Matting Dataset is rarely available due to the extreme difficulty of obtaining alpha matte and high commercial value, which also leads to the training and inference difficulties for data-driven methods. As shown in Tab. 1, the first matting dataset was proposed in DAPM [45], which focuses solely on portrait images. Later, two datasets were constructed, Composition-1K [56] with *216* human images for training, and the other is Distinctions-646 [38] with *343* human images (*333* for training, *10* for test). However, there still lacks a uniform benchmark for human image matting due to quality and quantity issues. To alleviate this gap, we build a large-scale human image matting dataset containing *2000* and *100* high-quality human images with human-annotated alpha mattes for training and test, respectively.

Image Matting Dataset		V	H	R
DAPM [45]	Train	1700	1700	800*600
	Test	300	300	800*600
Composition-1K [56]	Train	431	216	1256*1048
	Test	50	11	1655*1380
Distinctions-646 [38]	Train	596	333	1758*1573
	Test	50	10	1361*1477
Human-2K (Ours)	Train	**2000**	**2000**	**2560*1440**
	Test	100	100	2560*1440

Table 1: Comparison between different public matting datasets. V, H, and R refer to the total volume, the number of human, and the average resolution.

3. Methodology

3.1. Motivation

For trimap-based methods, most traditional methods solve for unknown alpha based on various local information (*e.g.* local smoothing assumptions [40]). For deep learning-based methods, trimap or transition variants (pseudo trimap) [6, 62, 31] are used to constrain the solution region, thus helping the network to optimise the transition region. However, both types pay close attention to the local areas around the transition regions and may ignore the coordination between global and local information (texture and colour similarity, positional correlation, *etc.*).

As shown in Fig. 2, the convolution kernel slides from left to right when performing a convolution operation. Types 1, 2 and 3 focus only on locally uncrossed fields (unknown **or** known regions). Only 4 and 5 perform enlightened transfers, with information flowing from Bg and Fg to the transition region, respectively. This paradigm therefore focuses more on local features and is similar to the partial convolution [30]. In addition, as the network deepens and the resolution decreases, it leads to a substantial loss of global positioning guides in trimap, which further weakens the affinity of unknown regions with known information.

Therefore, we analyse and propose an information mining and integration network in this paper. It can complement the neglected coordination between the global and local fields by mining and integrating multi-modal information from the input RGB image and trimap. Specifically, while retaining the mainstream RGB-Trimap branch for local information acquisition, we designed two functionally specific units RGB-Unit and Trimap-Unit based on the differ-

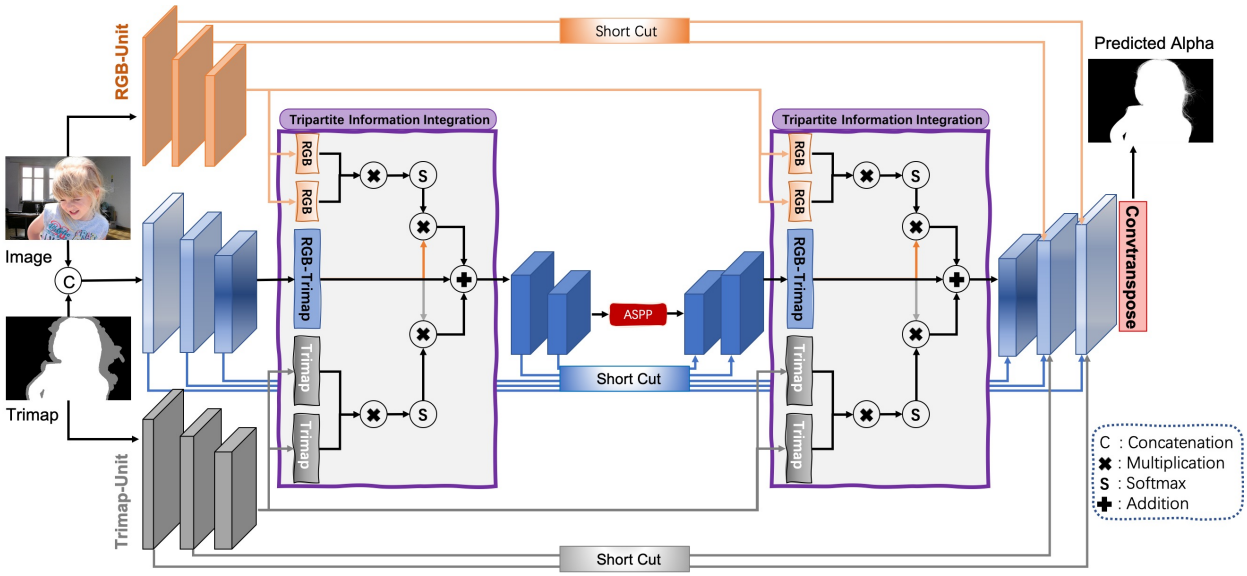

Figure 3: Pipeline of the proposed TIMI-Net. RGB-Unit (Orange Rectangle) and Trimap-Unit (Gray Rectangle) receive rgb and trimap respectively for mining the global information. The blue rectangle shows the RGB-Trimap processing of the mainstream local information. ResNet-18 [16] and ResNet-34 [16] are employed as the encoder for Trimap-Unit and RGB-Trimap branch, respectively. The Tripartite Information Integration (TI^2) module receives three inputs from RGB-Trimap branch, RGB-Unit and Trimap-Unit. Also, ASPP [5] is utilized to extract multi-scale contextual information.

ent effects of RGB and trimap for detailed appearance mining and related location guidance. After having obtained the respective functions from the two separate units, how to integrate them functionally with the mainstream branches is a key issue. Addition or concatenation is probably the most straightforward way. However, they tend to yield sub-optimal results due to undistinguished features. Instead, we have developed a Tripartite Information Integration (TI^2) that allows sufficient integration of complementary features. Regarding the complementarity of RGB-Unit and Trimap-Unit with the RGB-Trimap branch, TI^2 can transform the global information by using two different attentions computed from the RGB-Triamp branch and RGB-Unit, and the RGB-Trimap branch and Trimap-Unit. In this way, complementary features can be captured efficiently, thus harmonizing global and local information coordination.

3.2. Network Structure

The overall architecture of the proposed method is shown in Fig. 3. Our method uses the U-Net [42] structure with the short-cut (Blue) in each encoder block and decoder block as the baseline, and was used for acquiring the local information in the RGB-Trimap branch, which has been recognized by other methods [33, 26]. On the basis of it, we develop a RGB-Unit and a Trimap-Unit for their individually global information mining. There is also a short-cut (Orange or Gray) between each block from the two units and decoder. Then, the features from the RGB-Trimap branch and two other units are integrated in Tripartite Information Integration (TI^2). For enhancing the representation capabilities of the RGB-Trimap branch, we introduce the ASPP [5] to extract high-level semantic information.

RGB-Unit. We use three consecutive convolutional operations with a kernel size is *3* to achieve this. The number of channels is *16*, *64* and *128*, respectively. In addition, given the location and computational burden of the TI^2, we set the stride of the three convolutions to *2*, thus achieving *8x* downsampling for compatibility with resolution and computation. With this skin-deep design pattern, the global appearance, especially colour and texture information from unblended RGB fields, can be preserved, allowing for good disambiguation when foreground and background are locally similar to each other. [17].

Trimap-Unit. Position correlation is also important for modelling the long-range semantic and shape from trimap, especially for images where almost all regions are transition regions (*e.g.* nets, translucency, *etc.*). However, that character is under-utilized in the basic RGB-Trimap structure. To this end, we resort to a relatively deep network, ResNet-18 [16], to extract the high-level global representations for modelling position attributes. Meanwhile, in order to maintain the same resolution as the features at the mainstream RGB-Trimap branch, we only use the first three blocks (conv-1, res-2, res-3), and do not change the kernel size and number of channels in ResNet-18 [16]. Notably, we kept the MaxPool to increase the receptive field for a more global view, while the other two downsamplings are performed on the first convolution of res-2 and res-3.

Tripartite Information Integration. When we acquire the distinctive features from the RGB-Trimap branch, RGB-Unit, and Trimap-Unit, the major issue becomes how to integrate them efficiently. In general, addition or concatenation is a simplistic way, but they tend to treat the features from different modalities equally. Inspired by the Non-Local model [53] and in order to integrate the complementary multi-modal global information, we utilise the two attention maps gained from two separate units as indexes and sufficiently fetch the information about their properties from the mainstream RGB-Trimap branch separately.

We firstly briefly review the Non-Local model, which can be regularly defined as:

$$Y = g(X)A(X), \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times C}$ is the input features and output attentive features, where H, W, and C denote their height, width, and the number of channel, respectively. $A(X)$ is a normalization function that outputs an attention map:

$$A(X) = \text{softmax}\left(\theta(X)^T\phi(X)\right), \tag{3}$$

the g, θ and ϕ are learnable embedding functions, and the X is the feature extracted from one same domain.

As we can see from Eq. 2 and 3, the Non-Local focus on an identical feature X and it computes a self-attentive map in a bilinear projection way. While in our case, the RGB image and Trimap is a kind of cross-modal information. The RGB image is rich in global appearance (colour, texture, *etc.*), while more comprehensive high-level information (semantic and shape attributes) can be seen in trimap. Therefore, we incorporate the characteristics complement using supplementary information from different modalities. Formally, the features from RGB-Trimap branch, RGB-Unit and Trimap-Unit are depicted as $\boldsymbol{X}_{R\_T} \in \mathbb{R}^{H\times W\times C_{R\_T}}$, $\boldsymbol{X}_R \in \mathbb{R}^{H\times W\times C_R}$, and $\boldsymbol{X}_T \in \mathbb{R}^{H\times W\times C_T}$. With the two features X_R and X_T from the RGB and Trimap modality in mind, we integrated them into X_{R_T}:

$$Output_{TI^2} = X_{R\_T}(A(X_R) + A(X_T) + 1), \tag{4}$$

as for X_R and X_T, we embed them into the θ and ϕ spaces and obtain their attentive feature, respectively. While for X_{R_T}, instead of performing the linear embedding g on it, we apply residuals to superimpose the complementary information. Consequently, the global information can be utilized to steer the local information throughout the optimization of transition. In view of the computation costs, we choose to deploy it only at the stage in encoder and decoder when $output\_stride = 8$.

3.3. Loss Function

In order to verify the validity of this pattern and to prevent bias caused by other losses, we use only the alpha loss in all experiments:

$$L_\alpha = |\alpha_g^i - \alpha_p^i| \tag{5}$$

where i refers to the pixel position. The g and p denotes the Ground Truth and predicted alpha, respectively.

4. Experiments

4.1. Experiment Settings

To verify the effectiveness of the proposed method, we evaluate the performance on the following four datasets.

Alphamatting.com. It is an online benchmark website, which provides *27* images and alpha mattes for training and *7* images for evaluation.

Composition-1K. It contains **431** and *50* sets of foreground images and alpha mattes for training and test. And they are used to composite new images combined with background images from COCO [29] and VOC [11] in a ratio of *1* : *100* and *1* : *20*, respectively.

Distinctions-646. This dataset provides *596* and *50* sets of foreground images and alpha mattes with more challenging and diverse training and test objects. It takes the same rule and ratio as Composition-1K.

Human-2K. Although some public datasets we can use for human image matting task, quantity and quality remain an issue. Besides, we lack a uniform benchmark for comparison. Instead, our Human-2K provides *2100* high-accuracy images and alpha masks, which are good enough to be used as a benchmark for training (*2000*) and test (*100*). The same rules and ratios as Composition-1K [56] are used in our Human-2K to composite new images.

Implementation Details. The proposed framework is built on the public PyTorch [37] toolbox and is trained on a *24*-core i9-9920X *3.5*GHZ CPU, *128* GB RAM, and an NVIDIA Tesla V100 GPU. We use the Adam [21] optimizer for all the network training with an initial learning rate of *0.01* and batch size of *16*. The learning rate is divided by *10* at the epoch of {*20*, *30*, *40*}, {*60*, *80*}, and {*90*, *100*, *120*} for Composition-1K [56], Distinctions-646 [38], and our Human-2K dataset, respectively. It took *5*, *10*, and *15* days for the above three datasets to train *50* epochs, *100* epochs, and *150* epochs, respectively. We follow [56, 62, 38] to carry out the data augmentation. For training, we randomly cropped the input images and trimaps to a resolution of 512×512, 640×640, and 960×960, and then, random scaling, flipping, and rotation between $[-60, 60]$ degrees are applied to them. When do inference, we feed full-resolution images and trimaps to network to predict alpha mattes.

Evaluation Metrics. We follow [17, 3, 33, 26] to use th following four metrics to make comparisons. Namely the **S**um of **A**bsolute **D**ifferences, **M**ean **S**quare **E**rror, the **Grad**ient and **Conn**ectivity error.

SAD ↓	Average Rank				Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
	O	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
Ours	3.0	3.8	3.3	1.9	8.3	8.7	9.0	4.4	4.7	4.4	2.8	2.9	2.0	1.0	1.1	1.3	4.7	5.2	6.2	1.8	1.9	2.3	15.9	16.2	15.5	16.6	19.2	18.0
HDMatt [60]	8.4	10.3	7.0	8.0	9.5	10	10.7	4.7	4.8	5.8	2.9	3.0	2.6	1.1	1.2	1.3	5.2	5.9	6.7	2.4	2.6	3.1	17.3	17.3	17.0	21.5	22.4	23.2
AdaMatting [3]	10.3	9.1	9.3	12.6	10.2	11.1	10.8	4.9	5.4	6.6	3.6	3.4	3.4	0.9	0.9	1.8	4.7	6.8	9.3	2.2	2.6	3.3	19.2	19.8	18.7	17.8	19.1	18.6
BgMatting [43]	10.5	8.1	8.4	14.9	9.3	10.0	10.1	4.5	5.1	6.7	2.9	3.3	2.9	1.0	1.2	2.2	5.7	6.0	7.8	2.8	3.4	4.3	16.4	17.3	16.4	19.5	20.9	27.9
SampleNet [49]	10.7	8.6	10.3	13.3	9.1	9.7	9.8	4.3	4.8	5.1	3.4	3.7	3.2	0.9	1.1	2.0	5.1	6.8	9.7	2.5	4.0	3.7	18.6	19.3	19.1	20.0	21.6	23.2
GCA [26]	11.9	12.9	9.0	13.8	8.8	9.5	11.1	4.9	4.8	5.8	3.4	3.7	3.2	1.1	1.2	1.3	5.7	6.9	7.6	2.8	3.1	4.5	18.3	19.2	18.5	20.8	21.7	24.7
DIM [56]	13.5	14.9	12.5	13.0	10.7	11.2	11.0	4.8	5.8	5.6	2.8	2.9	2.9	1.1	1.1	2.0	6.0	7.1	8.9	2.7	3.2	3.9	19.2	19.6	18.7	21.8	23.9	24.1
IndexNet [33]	16.9	19.5	15.6	15.6	12.6	13.4	11.4	4.8	4.9	5.7	3.3	4.0	3.0	1.1	1.5	1.6	6.4	7.5	8.9	3.4	4.0	4.1	18.6	19.1	18.5	23.4	25.1	29.3
AlphaGAN [34]	18.5	19.5	18.8	17.3	9.6	10.7	10.4	4.7	5.3	5.4	3.1	3.7	3.1	1.1	1.3	2.0	6.4	8.3	9.3	3.6	5.0	4.3	20.8	21.5	20.6	25.7	28.7	26.7
Context-Aware [17]	21.0	25.0	19.0	18.9	10.4	11.1	10.1	6.4	7.4	7.1	4.1	4.5	3.8	2.3	3.1	3.0	7.1	8.2	9.1	3.5	5.5	4.1	18.3	19.2	16.5	21.1	23.3	24.6

Table 2: Comparison between our method and nine representative algorithms using the SAD metric. "O" represents overall rank, "**S**", "**L**", and "**U**" represent performance corresponding to the trimaps with different difficulty levels. Our method rank first regardless of the quality of the trimap (**S**mall, **L**arge, or **U**ser).

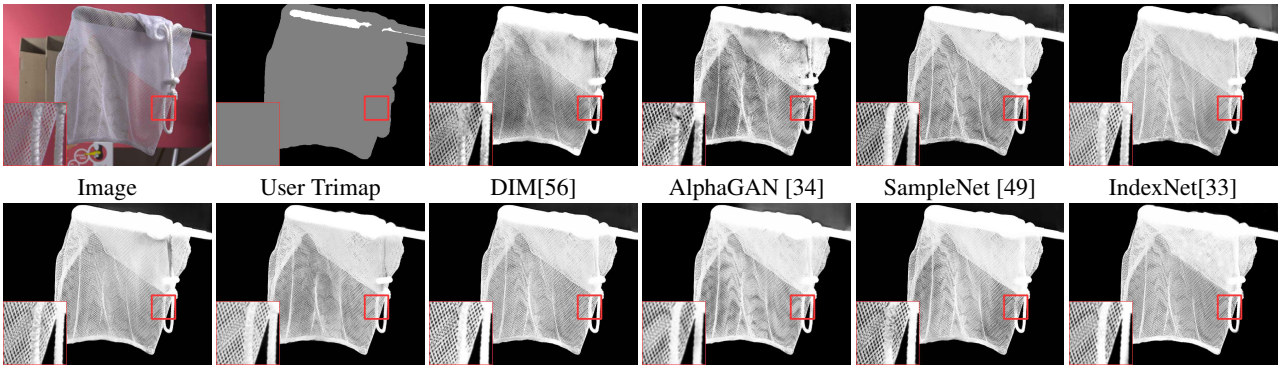

Figure 4: Visual comparison of TIMI-Net against SOTA methods results on the Alphamatting.com test set. All the results are obtained from the alphamatting.com website. More visual comparison can be seen in the supplementary materials.

4.2. Comparison to Prior Work

To evaluate the performance of the proposed method, we quantitatively and visually compare our method with other *2* classical and *9* SOTA deep learning-based image matting methods with available codes or results, including KNN [7], Closed-Form [23], DCNN [8], DIM [56], AlphaGAN [34], IndexNet [33], CAM [17], SampleNet [49], GCA [26], BgMatting [43] and HDMatt [60].

Tab. 2, 3, 4, 5 tabulate the quantitative results of our model and SOTA methods on four datasets. Our model rank first on the public benchmark alphamatting.com and outperforms all of them in all metrics on the Composition-1K, Distinctions-646 datasets, and our human image matting benchmark. Compared to the HDMatt [60] using patches, our method yields a result of *29.08* and *11.5* in terms of SAD and Conn on the Composition-1K test set, which brings *4.42* and *4.54* improvements. Meanwhile, our model outperforms GCA [26] by a large margin, with *6.22* and *5.40* improvements regarding SAD and Grad on the Composition-1K test set. The same improvements can be seen in diverse Distinctions-646 and Human-2K datasets, proving the superiority of our method in harmonizing the global and local information from the complementary RGB and Trimap modalities. We also give visual comparisons in Fig. 4, 5, 6 and 7. It can be obviously seen that our method can acquire fine details, such as the hair tip sites, the fingertip slits in Fig. 7.

4.3. Internal Analysis

We also validate the effectiveness of each component in TIMI-Net on three datasets (Tab. 3, 4, 5). *Basic* denotes the U-Net [42] structure with the shortcut for local information acquisition, and the RGB-Unit and Trimap-Unit are used to mine global appearance and position correlation, respectively. *SI* refers to the additive fusion of the local information from the RGB-Trimap branch with the global information from RGB-Unit and Trimap-Unit. TI_E^2 and TI_D^2 indicate the TI^2 is applied in Encoder and Decoder.

As shown in Tab. 3, we take the results on Composition-1K as an example. (i) Compared to the baseline model, the addition of our RGB unit reduced SAD and Conn by *5.48* and *6.76*, respectively, providing strong evidence that global appearance in the RGB domain is essential to guide the transition optimization, particularly for modelling colour and texture similarity. (ii) Trimap-Unit also improves the results, showing that pixel position correlation between transition and known regions (Fg and Bg) is necessary. (iii) We

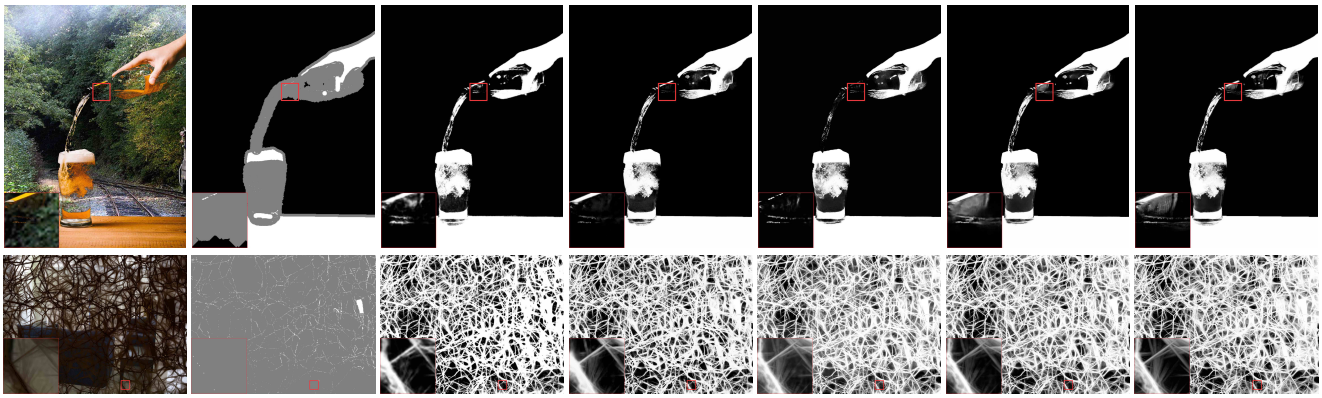

Figure 5: Visual comparison of TIMI-Net against SOTA methods results obtained on the Adobe Composition-1K test set.

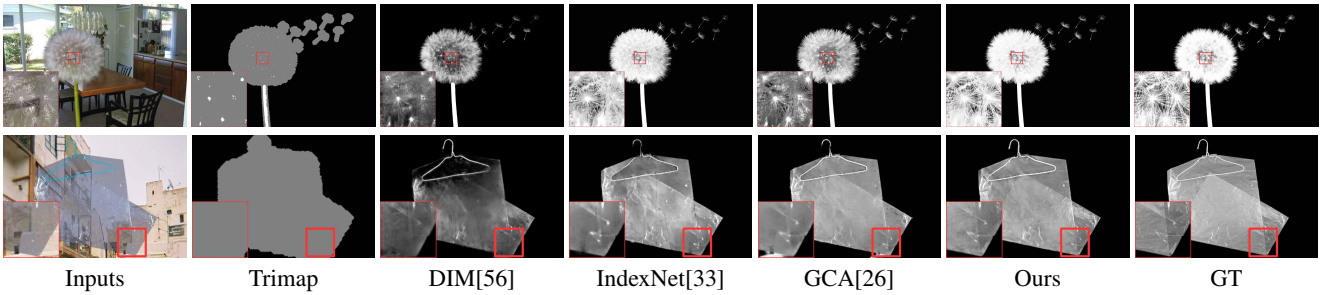

Figure 6: Visual comparison of TIMI-Net against SOTA methods results obtained on the Distinctions-646 test set.

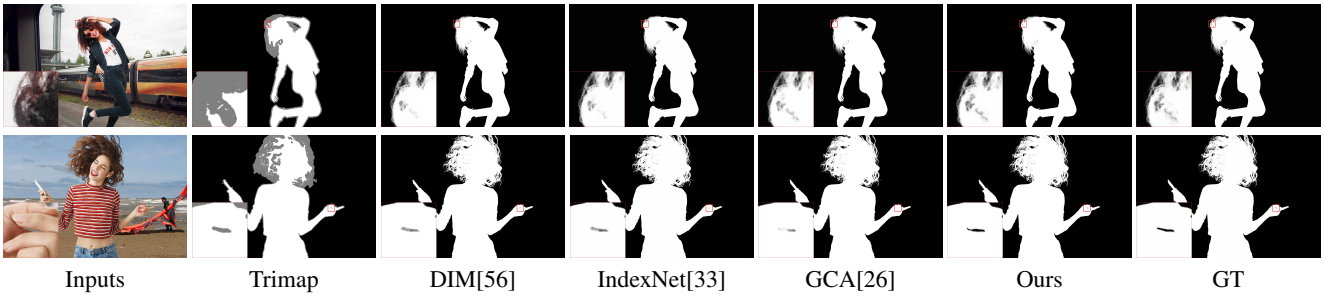

Figure 7: Visual comparison of TIMI-Net against SOTA methods results obtained on our Human-2K test set.

can also see that the results are further improved by incorporating RGB-Unit and Trimap-Unit into the mainstream branch compared to each separate unit, which validates the complementarity between the two, one for low-level details and the other for higher-level positioning. (vi) Compared to S_I, the proposed TI^2 allows for better integration of global and local information from the two complementary modalities, as it can sufficiently model the interrelationship between mainstream global and local information for each modality. It is worth noting that whether our TI^2 is applied in Encoder or Decoder both show growth. Our TI^2_D results show increases of *5.6%* and *7.2%* in the MSE and Grad metrics, and *11%* and *12%* increases can be seen in TI^2_E. (v) The multiplexing of information at the decoder stage can further harmonize the coordination between global and local information, thus better results are achieved. Similar results can also be seen in Tab. 4 and 5.

4.4. Generalization Analysis

For proving the generalization of our method and Human-2K, we have done cross-comparison experiments using different models on different datasets. We train each model using the entire training set of Composition-1K [56] and Human-2K dataset. For test, we selected only the images with human from the Composition-1K [56] test set. As shown in Tab. 6, the performance of both the representative methods DIM [56] and IndexNet [33] and our TIMI-Net were improved, especially for the DIM with MSE and Grad

Methods	SAD↓	MSE↓	Grad↓	Conn↓
KNN [7]	175.4	0.103	124.1	176.4
Closed-Form [23]	168.1	0.091	126.9	167.9
DCNN [8]	161.4	0.087	115.1	161.9
DIM [56]	54.4	0.014	31.0	50.8
IndexNet [33]	45.8	0.013	25.9	43.7
CAM [17]	35.8	0.008	17.3	33.2
SampleNet [49]	40.4	0.010	$*$	$*$
GCA [26]	35.3	0.009	16.9	32.5
HDMatt(Patch) [60]	33.5	0.007	14.5	29.9
Basic	38.60	0.0107	19.11	36.62
Basic + RGB-Unit	33.12	0.0081	14.48	29.86
Basic + Trimap-Unit	36.77	0.0097	16.67	34.34
Basic + S_I	31.21	0.0072	13.84	27.89
TI^2_E	30.18	0.0064	12.20	26.74
TI^2_D	30.60	0.0068	12.90	27.29
TIMI-Net (Ours)	**29.08**	**0.0060**	**11.50**	**25.36**

Table 3: Quantitative results on the Composition-1K [56] test set. $*$ means the results were not shown in their paper. *Basic* and **+** denote our baseline network and the addition operation. *S_I* denotes and the straightforward additive integration between the RGB-Trimap branch, RGB-Unit, and Trimap-Unit. TI^2_E and TI^2_D refer to the TI^2 being applied in Encoder and Decoder.

Methods	SAD↓	MSE↓	Grad↓	Conn↓
DIM [56]	44.15	0.031	39.08	44.65
IndexNet [33]	34.47	0.019	28.31	33.37
GCA [26]	26.59	0.015	19.50	25.23
Basic	32.20	0.0163	20.07	28.75
Basic + RGB-Unit	26.93	0.0140	18.51	25.57
Basic + Trimap-Unit	27.67	0.0150	19.47	26.41
Basic + S_I	25.96	0.0135	16.01	24.68
TI^2_E	25.27	0.0124	15.32	24.22
TI^2_D	25.56	0.0131	15.86	24.09
TIMI-Net (Ours)	**22.28**	**0.0107**	**14.38**	**20.49**

Table 4: Quantitative results on the Distinctions-646 [38] test set.

improving by *0.005* and *4.67*, respectively, which demonstrates that the generalisation of our dataset is robust and can be used as a benchmark. Meanwhile, our model still turn out to be optimal, which also implies the superiority.

4.5. Real-World Image Matting

In practice, to facilitate selecting areas of interest, novice users are asked to draw trimaps based on known and unknown regions. As shown in Fig. 1, the quality of these trimaps is inferior. However, as our method harmonises more global information, our results are better than those of CAM [17]. Notably, both models used were trained using solely the Composition-1K [56] training set.

Methods	SAD↓	MSE↓	Grad↓	Con↓
DIM [56]	7.53	0.008	6.4	6.7
IndexNet [33]	6.55	0.006	4.5	5.5
GCA [26]	5.18	0.004	3.0	4.0
Basic	5.87	0.0047	3.68	4.81
Basic + RGB-Unit	5.16	0.0037	2.78	4.03
Basic + Trimap-Unit	5.45	0.0041	3.03	4.36
Basic + S_I	4.93	0.0034	2.59	3.71
TI^2_E	4.65	0.0031	2.43	3.47
TI^2_D	4.83	0.0033	2.57	3.62
TIMI-Net (Ours)	**4.20**	**0.0026**	**2.06**	**2.95**

Table 5: Quantitative results on our Human-2K test set.

Model		SAD↓	MSE↓	Grad↓	Conn↓
DIM [56]	C_C	15.25	0.0150	10.99	14.41
	H_C	11.46	0.0100	6.32	10.04
IndexNet [33]	C_C	11.27	0.0086	6.01	9.64
	H_C	10.57	0.0070	5.30	9.00
Our	C_C	8.11	0.0046	3.12	6.24
	H_C	7.41	0.0040	2.78	5.55

Table 6: Generalization analysis of our Human-2K dataset. C and H refer to the Composition-1K [56] and our Human-2K datasets. C_C and H_C mean the model trained on C and H, then they are tested on C.

5. Conclusion and Future Work

In this paper, we have observed that previous image matting methods pay more attention to local areas closed to transitional regions, which potentially ignore the coordination between global and local information. Based on this observation, we have proposed a novel tripartite information mining and integration model to sufficiently supplement the ignored harmonization. To advance the development of the human image matting task, we have prepared a new large-scale high-accuracy human image matting dataset (Human-2K). Finally, we have conducted extensive experiments to verify the effectiveness of the proposed method against SOTA approaches.

Our method does have limitations, the parameters of two units and computation cost of TI^2 limit its application to real-time. In the future, we will explore how to exploit other techniques to model long-range information in a light way to image and video matting [48, 63].

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China under Grant 61972067, the Innovation Technology Funding of Dalian (2020JJ26GX036), and PicUP.Ai project of the Winroad Holdings Ltd.

References

[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 228–236, 2017.

[2] Yağız Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Trans. Graph.*, 37(4), 2018.

[3] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 8818–8827, 2019.

[4] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Tom-net: Learning transparent object matting from a single image. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9233–9241, 2018.

[5] Liang Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. 2017.

[6] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *ACM Int. Conf. Multimedia*, page 618–626, 2018.

[7] Qifeng Chen, Dingzeyu Li, and Chi Keung Tang. Knn matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2175–2188, 2013.

[8] Donghyeon Cho, Yu-Wing Tai, and In So Kweon. Deep convolutional neural network for natural image matting using initial alpha mattes. *IEEE Trans. Image Process.*, 28(3):1054–1067, 2019.

[9] Yung Yu Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages II–II, 2003.

[10] Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6841–6850, 2021.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. Jour. Comput. Vis.*, 88(2):303–338, 2010.

[12] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *Proc. Eur. Conf. Comput. Vis.*, pages 204–219, 2016.

[13] Eduardo S. L. Gastal and Manuel M. Oliveira. Shared sampling for real-time alpha matting. *Comput. Graph. Forum*, 29(2):575–584, 2010.

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Xu Bing, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Int. Conf. Neural Inf. Process. Systems*, page 2672–2680, 2014.

[15] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proc. VIIP*, pages 423–429, 2005.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016.

[17] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4129–4138, 2019.

[18] Jubin Johnson, Ehsan Shahrian Varnousfaderani, Hisham Cholakkal, and Deepu Rajan. Sparse coding for alpha matting. *IEEE Trans on Image Process.*, 25(7):3032–3043, 2016.

[19] Levent Karacan, Aykut Erdem, and Erkut Erdem. Alpha matting with kl-divergence-based sparse sampling. *IEEE Trans. Image Process.*, 26(9):4523–4536, 2017.

[20] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. Is a green screen really necessary for real-time portrait matting? *arXiv preprint arXiv:2011.11961*, 2020.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[22] P Lee and Ying Wu. Nonlocal matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2193–2200, 2011.

[23] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):228–242, 2007.

[24] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10):1699–1712, 2008.

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

[26] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proc. AAAI Conf. Artif. Intell.*, pages 11450–11457, 2020.

[27] Yaoyi Li, Qingyao Xu, and Hongtao Lu. Hierarchical opacity propagation for image matting. *arXiv preprint arXiv:2004.03249*, 2020.

[28] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. pages 8762–8771, 2021.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

[30] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proc. Eur. Conf. Comput. Vis.*, pages 85–100, 2018.

[31] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-Sheng Hua. Boosting semantic human matting with coarse annotations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.

[32] Yuhao Liu, Jiake Xie, Yu Qiao, Yong Tang, and Xin Yang. Prior-induced information alignment for image matting. *IEEE Trans. Multimedia*, 2021.

[33] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3265–3274, 2019.

[34] Sebastian Lutz, Konstantinos Amplianitis, and Aljoscha Smolic. Alphagan: Generative adversarial networks for natural image matting. In *Proc. Bri. Mach. Vis. Conf.*, page 259, 2018.

[35] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2021.

[36] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2021.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. 2019.

[38] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 13676–13685, 2020.

[39] Yu Qiao, Yuhao Liu, Qiang Zhu, Xin Yang, Yuxin Wang, Qiang Zhang, and Xiaopeng Wei. Multi-scale information assembly for image matting. *Computer Graphics Forum*, 39(7):565–574, 2020.

[40] C. Rhemann and C. Rother. A global sampling method for alpha matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2049–2056, 2011.

[41] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1826–1833, 2009.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[43] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2291–2300, 2020.

[44] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 636–643, 2013.

[45] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *Proc. Eur. Conf. Comput. Vis.*, pages 92–107, 2016.

[46] Jian Sun, Jiaya Jia, Chi Keung Tang, and Heung Yeung Shum. Poisson matting. *ACM Trans. Graph.*, 23(3):315–321, 2004.

[47] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 11120–11129, 2021.

[48] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6975–6984, 2021.

[49] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3050–3058, 2019.

[50] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection, 2020.

[51] Jue Wang and Michael F. Cohen. An iterative optimization approach for unified image segmentation and matting. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 936–943, 2005.

[52] Jue Wang and Michael F. Cohen. Optimized color sampling for robust matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2007.

[53] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7794–7803, 2018.

[54] Ke Xu, Xin Wang, Xin Yang, Shengfeng He, Qiang Zhang, Baocai Yin, Xiaopeng Wei, and Rynson WH Lau. Efficient image super-resolution integration. *The Visual Computer*, 34(6):1065–1076, 2018.

[55] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2281–2290, 2020.

[56] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 311–320, 2017.

[57] Xin Yang, Yu Qiao, Shaozhe Chen, Shengfeng He, Baocai Yin, Qiang Zhang, Xiaopeng Wei, and Rynson W. H. Lau. Smart scribbles for image matting. *ACM TOMM*, 16(4), 2020.

[58] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. In *Proc. Int. Conf. Neural Inf. Process. Systems*, pages 4590–4600, 2018.

[59] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating hdr transformation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1798–1807, 2018.

[60] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. 2021.

[61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5505–5514, 2018.

[62] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7461–7470, 2019.

[63] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporal coherent video object matting. *arXiv preprint arXiv:2105.11427*, 2021.