Prior-Induced Information Alignment for Image Matting

Yuhao Liu*, Jiake Xie*, Yu Qiao, and Yong Tang†, Xin Yang†

Abstract—Image matting is an ill-posed problem that aims to estimate the opacity of foreground pixels in an image. However, most existing deep learning-based methods still suffer from the coarse-grained details. In general, these algorithms are incapable of felicitously distinguishing the degree of exploration between deterministic domains (e.g. certain FG and BG pixels) and undetermined domains (e.g. uncertain in-between pixels), or inevitably lose information in the continuous sampling process, leading to a sub-optimal result. In this paper, we propose a novel network named Prior-Induced Information Alignment Matting Network (PIIAMatting), which can efficiently model the distinction of pixel-wise response maps and the correlation of layer-wise feature maps. It mainly consists of a Dynamic Gaussian Modulation mechanism (DGM) and an Information Alignment strategy (IA). Specifically, the DGM can dynamically acquire a pixel-wise domain response map learned from the prior distribution. The response map can present the relationship between the opacity variation and the convergence process during training. On the other hand, the IA comprises an Information Match Module (IMM) and an Information Aggregation Module (IAM), jointly scheduled to match and aggregate the adjacent layer-wise features adaptively. Besides, we also develop a Multi-Scale Refinement (MSR) module to integrate multi-scale receptive field information at the refinement stage to recover the fluctuating appearance details. Extensive quantitative and qualitative evaluations demonstrate that the proposed PIIAMatting performs favourably against state-of-the-art image matting methods on the Alphamatting.com, Composition-1K and Distinctions-646 dataset.

Index Terms—Image Matting, Gaussian Distribution, Information Alignment.

I. INTRODUCTION

THE digital matting is one of the essential 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, especially in the field of digital image editing and film production. Formally, the input image can be modelled as a linear combination of foreground and background colours, as shown below:

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

where \(i\) refers to the pixel position in the input image. \(F_i, B_i\) and \(\alpha_i\) denote the foreground, background colour, and alpha matte separately. Given an input image \(I\), the image matting algorithms intend to solve \(F, B, \alpha\) simultaneously. Based on Eq. 1 for a pixel of a typical 3-channel (e.g. RGB) input image, 7 unknowns (i.e. 3 \(F\) values, 3 \(B\) values and 1 \(\alpha\) value) need to be solved, but there are only 3 known quantities (3 \(I\) values). Therefore, the problem is highly ill-posed.

Natural image matting is essentially the computation of the opacity of each pixel in an image \([1, 4]\). According to Eq. 1, the pixels in an image can be divided into three groups where \(\alpha_i = 1\) and \(\alpha_i = 0\) refer to the foreground and background regions, respectively, while \(\alpha_i \in (0, 1)\) represent the transition regions (usually the area that needs to be solved in image matting) which can be decoupled to the deterministic domains (certain foreground and background pixels) and undetermined domains (uncertain in-between pixels). In general, some user-specified constraint information is utilized, such as trimap and scribble, which are helpful to reduce the solution space of the ill-posed problem for solving Eq. 1. The trimap is composed of three parts, white, black, and grey, indicating the foreground, background, and transition regions separately. While the scribble can be regarded as a simplified version of trimap, which specifies the foreground and background regions using few sparse scribbles. With the trimap as assistance, early works attempted to solve Eq. 1 using the colour distribution, leading to blurred or chunky artefacts \([1]\).

Motivated by the success of deep learning, the digital image matting methods based on Deep Convolutional Neural Networks (DCNNs) have been proposed. Cho et al. \([5]\) took the...
results of [6] and [7] combined with the RGB colours as input to predict an alpha matte by using the CNN. Then, Deep Image Matting (DIM) [1] proposed the first large-scale image matting dataset and trained a model in an end-to-end fashion. Most of the following methods [8]–[11] endeavoured to design a more complicated network to implicitly constrain Eq. 1 for obtaining high-quality alpha mattes. However, all the above methods suffer two possible limitations: 1) the disproportion between pixel domains. 2) the information discrepancy between the process of sampling. A visual comparison of different methods can be seen in Fig. 1.

To address the above issue-1, the LFM [12] combined two classification networks and a fusion network under the supervision of a hybrid loss to jointly predict alpha mattes with a single RGB image as input. The hybrid loss function intends to calculate the different regions in varied manners, especially L1 loss for transition regions and L2 loss for the rest of the regions. While this setting can attenuate the disproportion of pixels at the different domains to some extent, it neglects the internal distribution imbalance of the transition regions and can only be regarded as a static pre-defined partition that does not properly accommodate varieties in opacity. As for the information discrepancy, the index map generated in indexNet [2] is similar to an attention map, which can learn the index from specific feature maps and propagate the crucial information to the corresponding upsampling operation in the decoder stage, thereby preventing the loss of details. However, it ignores the complementarity between adjacent layers in information selection and ensemble.

Overall, the domain disproportion and the information discrepancy can bring necessary details lost for image matting. On the one hand, the pixel distribution of unknown regions is unbalanced (see the second column of Fig. 2), the number of pixels in the deterministic domains (e.g., foreground (FG) pixels (green colour) and background (BG) pixels (red colour)) and the undetermined domains (uncertain in-between pixels (blue colour)) is highly biased. Meanwhile, due to the constraint of trimap, the opacity of large amount of foreground and background pixels are already given. Hence the information dissemination is more conducive to the deterministic domains in unknown regions according to the local smoothness assumption [13]. On the other hand, the continuous sampling operation is likely to produce information discrepancy and deteriorate the details of the boundary. Consequently, we argue that both biased pixel distribution and information discrepancy may produce a negative influence on pixel regression and sampling propagation, resulting in deficient alpha mattes.

To address the limitations mentioned above, we propose a Prior-Induced Information Alignment network for image matting (PIIAMatting) as depicted in Fig. 3. As for the disproportion of the pixel domains, we propose the Dynamic Gaussian Modulation mechanism (DGM) that can adjust the degree of exploration of pixels at different positions by a domain response map learned from the prior distribution in Ground Truth. The response map can better adapt to the opacity variation, thereby facilitating the entire convergence process during training. To further alleviate the information discrepancy, we propose an Information Alignment strategy (IA) to sufficiently explore the complementarity between adjacent layer-wise features. It is decoupled into Information Match Module (IMM) and Information Aggregation Module (IAM), which performs guided information matching and information ensemble. Moreover, a Multi-Scale Refinement (MSR) module is exploited to integrate multi-scale information to improve the alpha matte in a parallel residual fashion. Finally, extensive experiments on three datasets demonstrate that our method can achieve consistently superior performance over recent state-of-the-art approaches.

Our contributions are summarized as follows:

1) We propose a Dynamic Gaussian Modulation mechanism (DGM), which can distribute the adaptive responses to each pixel according to the domain response map learned from the prior distribution. This mechanism is very efficient for the highly translucent pixels and can stabilize the process of training.
2) We propose an Information Alignment strategy consisting of an Information Matching Module (IMM) and an Information Aggregation Module (IAM). Both of them are combined to commit to preserving details by matching and aggregating the features of two adjacent layer-wise features.
3) Experimental results demonstrate that the proposed PIIAMatting can achieve state-of-the-art performance on three datasets, proving the effectiveness and superiority of the proposed method.

II. RELATED WORKS

In this section, we will briefly review the image matting from the following three categories: sampling-based, affinity-based, and deep learning-based approaches.

**Sampling-based** approaches collect a set of known foreground and background samples to find candidate colours for the foreground and background of a given pixel. According to the local smoothness assumption [13] on the image statistics, which requires these sampling colours should be "close" to the true foreground and background colours. Once the foreground and background colours are determined, the corresponding alpha values can be computed based on the Eq. 1. Bayesian
Matting [16], Cluster Matting [17], Shared Matting [18], K-L divergence Matting [19], Global Matting [13], Comprehensive Matting [20], Robust Matting [21], Iterative Matting [22] are some typical representative methods follow this assumption.

Affinity-based methods take advantage of affinities of neighboring pixels to propagate the known alpha value from the known regions into unknown regions. ClosedForm Matting [7], Spectral Matting [23], Information-Flow Matting [24], KNN Matting [6], Random Walk Matting [25], Non-Local Matting [26], Poisson Matting [27], etc. are some of the known propagation methods introduced in this direction. The most representative method is ClosedForm Matting [7], which is derived from the matting Laplacian and acquires the globally optimal alpha matte by solving a sparsely linear system of equations. While the affinity-based approaches can obtain successful alpha matte, they typically encounter the difficulty of high computational complexity and memory limitations.

Deep learning-based algorithms have recently received considerable attention. Benefiting from the effective feature representation of the Deep Convolutional Neural Networks (DCNNs), a few approaches have obtained promising performance. Shen et al. [28] proposed the first automatic matting method for portrait photos in an end-to-end manner. DCNN matting [29] combined the results of [6] and [7] with the input image to predict the final alpha matte. Deep image matting [1] firstly introduced a large-scale dataset and utilized the Seg-Net [30] with a refinement to jointly estimate the alpha mattes. AlphaGAN [31] firstly introduced the generative adversarial network (GAN) [32], [33] to generate alpha mattes using a discriminator, indicating the GAN is capable to process well for pixel-wise regression tasks. SampleNet [8] and GCA Matting [3] learned from image context to leverage the background information to guide the foreground prediction. IndexNet matting [2] introduced the idea of indexing information to integrate upsampling operators for improving the ability to retain details. AdaMating [10] divided the matting task into two branches (i.e., trimap adaption and natural image matting), which can work together to jointly refine the alpha matte. Similarly, the Context-Aware matting [9] dismantled the image matting into foreground and alpha estimation, then pushed them to generate the alpha matte and foreground image simultaneously. However, almost all the above methods turned to a static optimization without taking the opacity variation into account, leading to some obvious artifacts. Moreover, LFM matting [12] combined two classification networks and a fusion network under the supervision of a hybrid loss to jointly predict alpha mattes with single RGB image as input. Similar to LFM without auxiliary assistance, the HAttMatting [11] employed spatial and channel-wise attention to integrate appearance cues and advanced semantics to achieve better results. Besides, there are some other deep learning-based approaches [34]–[45] to solve image matting and the field-specific tasks.

Fig. 3: Schematic illustration of the proposed PIIAMatting. We employ the ResNet-50 [14] as the backbone and utilize the ASPP [15] to extract multi-scale contextual information. According to the depth of the network from shallow to deep, we split the ResNet-50 into five blocks from block0 to block4. Information Match Module (IMM) and Information Aggregation Module (IAM) are concrete implementations of the Information Alignment strategy (IA), represented by the green box and blue box, respectively. The yellow box denotes the Multi-Scale Refinement module (MSR).
III. Method

In this section, we first describe the approach overview briefly in III-A. Then, we discuss the interrelation between different loss functions and introduce our Dynamic Gaussian Modulation mechanism (DGM) in III-B. Subsequently, we give a detailed depiction of our network structure in III-C and elaborate on the loss function we use during the training in III-D. The overall architecture is shown in Fig. 3.

A. Overview

The deep learning-based image matting methods have shown their advantages compared with the original colour-based methods. However, most of them either treat pixels at different domains equally or lose potential boundary details during the sampling operation, resulting in the sub-optimal alpha matte. Consequently, we propose a PIIAMatting to model the pixel-wise domain modulation and the information alignment within a single network in an end-to-end manner. On the one hand, we argue that the pixels in different regions (such as the $\alpha_i \in (0, 1)$ and the $\alpha_i = 0$ or 1) should be treated as distinguished so that the model can be driven to focus more on the hard-to-mine samples. On the other hand, a well-balanced model should also explore layer-wise features to enhance the ability to match and integrate fine-grained details, which is critical in image matting, especially in boundary areas.

B. Dynamic Gaussian Modulation

With the trimap as assistance, the target of image matting is to predict the opacity of each pixel within the unknown regions. As shown in the second column of Fig. 2, typically, the pixels in unknown regions could be divided into the deterministic and undetermined domains. The deterministic domains can be split into certain FG and BG pixels, while the undetermined domains can be denoted as highly uncertain in-between pixels directly. As a common practice, it is solved as a regression problem using L1 or L2 Loss, but it has limitations to a certain extent. On the one hand, the distribution of the pixels is highly biased between the deterministic and undetermined domains. Normally, the proportion of the FG and BG pixels is far less than the in-between pixels. This disproportion may cause instability during model learning and potentially deteriorate the quality of the alpha matte. On the other hand, the exploration of opacity is varied according to the positions of pixels domains. Based on the convolution slide window and local smoothness assumption, the pixels at the deterministic domains (green and red colour) are directly connected to the known regions. Thus the opacity solution process on the deterministic domains is more accessible than the undetermined domains (blue colour).

Without regard to varying network structures [1], [9], [12], the solving process of the pixel opacity is mainly optimized by the objective function. In addition to the L1 and L2 loss, composition loss [1] is also widely used in the RGB domain to improve the alpha mattes. Although the blended loss function can improve the result, it does not solve the problem of disproportion essentially.

Another solution was proposed in LFM [12], which employs distinct losses to supervise different areas:

$$L^i_{\alpha} = \begin{cases} |\alpha^{i}_{p} - \alpha^{i}_{g}|, & 0 < \alpha^{i}_{g} < 1, \\ (\alpha^{i}_{p} - \alpha^{i}_{g})^2, & \alpha^{i}_{g} \in [0, 1]. \end{cases}$$ (2)

where $\alpha^{i}_{p}$ and $\alpha^{i}_{g}$ refer to the predicted alpha matte and ground truth alpha matte at pixel $i$ separately. However, since the transparency is continuously distributed, this solution can only be expressed as a static pre-defined partition, which does not completely solve the dynamically varied opacity. Furthermore, it is difficult to draw a clear demarcation line to separate the deterministic and undetermined domains.

Based on the above observations, we propose our Dynamic Gaussian Modulation mechanism (DGM), which naturally fits the task of image matting. As shown in Fig. 4, the DGM is evolved from the standard Gaussian distribution for which $\mu = 0$, $\sigma^2 = 1$. The different colours represent the distinct response scheme. From top to bottom, the response coefficient for the undetermined domains is getting smaller and smaller. Specifically, we denote the Ground Truth as $F_{ground-truth} \in \mathbb{R}^{1 \times H \times W}$, and the output feature map holds the same size as the Ground Truth. Since the distribution of pixels in the transition regions varies for different images, we resort to the case-specific prior distribution in the corresponding Ground Truth to acquire the domain response map.

$$R = f_{transform}(F_{ground-truth}) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\alpha^{i}_{g} - \mu)^2}{2\sigma^2}}$$ (3)

where $i$ denotes the position of pixels, $g$ refer to the Ground Truth, and the $\alpha^{i}_{g}$ is the ground truth alpha matte at pixel $i$. The mean $\mu$ and variance $\sigma^2$ of the Gaussian distribution are adjustable parameters, and they were set to 0.5 and 0.25 unless otherwise specified. The Gaussian Distribution will assign sligher responses to the deterministic domains (foreground and background pixels) and stronger responses to the undetermined domains (uncertain in-between pixels).
The higher the uncertainty level of pixels, the stronger the responses. In general, the higher the uncertainty for pixels with opacity closer to 0.5. In this way, every pixel at every image can attain an opacity-adaptive pixel-wise domain response coefficient and contribute to the network to explore more valuable information.

Furthermore, as the training progresses, the adaptability of our model to unknown regions is continuously enhanced. In order to better match the opacity variation according to the fitting ability of the model during the training process, we further propose to adjust the response map dynamically. Intuitively, we can gradually reduce the $\sigma^2$ as the number of training iterations progresses so as to modulate the responsiveness of the deterministic and undetermined domains. The specific formulation is as follows:

$$D(\sigma^2) = \sigma^2 + 0.005 \quad \text{while iteration}\%2000 = 0 \quad (4)$$

the $\sigma^2$ is initialized to 0.25 and increased by 0.005 every 2000 iterations. As the iteration progresses, the response coefficient assigned to the undetermined domains gradually decreases. Hence, we can achieve domain-specific weight distribution through the Dynamic Gaussian Modulation mechanism induced by the prior information, which can stabilize the convergence process while exploring information evenly. Some examples of the domain response map obtained from our DGM are shown in the third column in Fig. 3. Since the pixels are continuously distributed inside the unknown regions, the domain response coefficient varies according to the different opacity of pixels. From blue to orange, the deeper the colour, the stronger the response.

C. Model Architecture

As mentioned above, we explicitly assign the dynamic responses to different pixels for exploring various types of regions. However, as the network deepens, the resolution of the feature map decreases gradually. During this process, the down-sampling operation inevitably results in information loss, especially the average pooling that discard the high-frequency information and only retain the low-frequency information [46], leading to ambiguity in the boundary regions. Same as average pooling, the other two prevalent options for down-sampling are max-pooling and convolution with stride larger than 2, which also incur different degrees of information loss. Therefore, the information discrepancy will occur in the corresponding up-sampling stage in the decoder, which will bring accuracy loss and potentially affect the convergence of the model. Consequently, we argue that designing a strategy to match valid information for aggregation effectively can improve the model’s overall performance.

In the following, we firstly describe the overall network structure. Then, the Information Alignment strategy is introduced detailiedly in both the Information Match Module and Information Aggregation Module. At last come to introduce our MSR.

Overall network structure: Fig. 3 shows the pipeline of our model, which uses the original image concatenated with trimap as input. We use ResNet-50 [14] as our backbone, and it is split into five blocks from block0 to block4 according to the depth of the network from shallow to deep. In order to fuse multi-scale contextual information and enhance the feature representation capability, we extract the features from the block4 and transmit them to the ASPP [15] to aggregate features of different receptive fields. Specifically, considering the computation overhead, we exploit three parallel atrous convolution layers with different rates and an average pooling layer followed by a convolution operation with a kernel size of $1 \times 1$ to capture multi-scale information. The dilation rates are set to 1, 8 and 16, respectively. Then, the features of the four branch are concatenated and fed into a convolution operation with a kernel size of $3 \times 3$ followed by a BatchNorm and ReLu.

To ensure the integrity and validity of the sampled features, we utilize the IMM and IAM to jointly perform the adjacent layer-wise information match and aggregation. As the information alignment has been decoupled into information match and aggregation, the information aggregation is automatically fulfilled once the information match is complete. Besides, the MSR applied dilated convolutions with different rates in a parallel residual manner to complement more details and generate high-quality results.

Information Alignment strategy: Skip-Connection [47] is a compelling feature enhancement method in different kinds of vision tasks [15, 48, 49], which bridge the information gap by concatenating or adding features between encoder and decoder at each down-sampling and up-sampling stage. Although it can attenuate the information loss to some extent, it uses only the peer-to-peer level features and ignores the relationship between adjacent layer-wise features (e.g. features from block0 and block1), thus inevitably introducing superfluous bias during the constant sampling operation. Therefore, it is only direct reuse of features without discrimination and does not radically address the information discrepancies. Moreover, the low-level and high-level features are intrinsically different but complementary to each other, one for the details preserved while the other for semantic retaining, then jointly contribute to the generation of alpha mattes. In the IA, we perform the IMM to match the valuable information at each stage in the encoder between adjacent layer-wise layers during down-sampling. As shown in Tab. VII the experimental analysis also confirms the effectiveness of our Information Alignment strategy.

Information Match Module:

We feed two inputs to the IMM: $F_{\text{low}} \in R^{H \times W \times C}$ and $F_{\text{high}} \in R^{\frac{H}{2} \times \frac{W}{2} \times C}$. $H$, $W$ and $C$ represent the height, width and number of channels respectively. Firstly, we transfer $F_{\text{high}}$ to the transposed convolution for up-sampling. Then, the product of $F_{\text{high}}$ and $F_{\text{low}}$ is used to match the valuable details, clear foreground, and background features into the IAM module. The implementation process is shown in the green box in Fig. 3 and specific operations can be depicted as follows:

$$F^{i}_{\text{imm}} = F^{i-1}_{\text{low}} \ast T(F^{i}_{\text{high}}) \quad (5)$$

where $i$ refers to the index of the block in ResNet, $T(\cdot)$ is the upsampling operation and $\ast$ denotes the element-wise product.
TABLE I: Comparison of our method with eight top-performing algorithms on the Alphamatting.com dataset [4]. “O” represents overall rank, “S”, “L”, and “U” represent performance corresponding to the trimaps with three difficulty levels: small, large, and user. Please see Alphamatting.com for more results.

| Methods         | Overall | S  | L  | U  | Overall | S  | L  | U  | Overall | S  | L  | U  |
|-----------------|---------|----|----|----|---------|----|----|----|---------|----|----|----|
| Our             | 7.6     | 3.8| 9.8| 9.4| 4.9     | 3.1| 6.0| 5.6| 8.0     | 5.3| 7.8| 11.0|
| AdaMatting [10] | 8.5     | 6.5| 7.8| 11.4| 7.5     | 6.8| 6.5| 9.4| 8.1     | 5.0| 5.6| 13.6|
| SampleNet [8]   | 9.7     | 6.6| 9.6| 12.8| 8.4     | 6.6| 8.0| 10.6| 9.7     | 5.9| 7.4| 15.8|
| GCA [3]         | 10.0    | 10.0|8.6| 11.3| 9.3      | 10.1|6.8|11.0| 9.3     | 10.6 |9.9|7.3|
| Context-Aware [9]| 12.2 | 15.8|13.4|7.5| 18.1    | 22.0|16.0|16.3| 9.3     | 10.6 |9.9|7.3|
| InformationFlow [24]| 17.8 | 20.4|16.1|16.8| 14.1    | 16.5|12.8|13.1| 13.3    | 12.3 |11.8|16 |
| DIM [1]         | 14.5    | 17.3|13.8|12.5| 13.1    | 14.3|13.8|11.4| 20.1    | 23.0 |18.8|18.6|
| AlphaGAN [31]   | 18.8    | 19.3|19.9|17.4| 15.7    | 16.5|15.8|14.8| 18.1    | 17.0 |15.9|21.4|

Information Aggregation Module: The IAM also takes two inputs, that is $F^{i-1}_{iam} \in \mathbb{R}^{H \times W \times C}$ generated from the preceding IAM and $F^{imm}_{imm} \in \mathbb{R}^{H \times W \times C}$ obtained from IMM. Specifically, the IAM superimposes these two features by element-wise summation and then extracts the most critical features by element-wise product. The operations can be formally defined as follows:

$$F^{i}_{iam} = \begin{cases} (F^{i}_{imm} + T(F^{aspp}_{aspp})) \ast (F^{aspp}_{aspp}), & i = \text{top}. \\ (F^{i}_{imm} + T(F^{i-1}_{iam})) \ast (F^{i-1}_{iam}), & i = \text{else}. \end{cases}$$  \hspace{2cm} (6)

where $T(\cdot)$, $i$, and $\ast$ hold the same meaning in IMM and IAM, and top refers to the same depth as block4 in ResNet-50. As shown in the blue box in Fig. 3, we adopt the output features of ASPP [15] as the initialization for the whole decoder process, and it is sufficient to directly aggregate the initialization with features from the IMM since features from the encoder have been matched through our IMM.

In this way, via the utilization of relatively high feature information by our IA strategy, valuable information can be effectively matched and retained between adjacent cascading features, resulting in a more accurate aggregation of information at the sampling stage.

Muti-Scale Refinement Module: In the encoder-decoder structure, the incorporation of trimap helps to guide the model to learn significant information. However, the unknown regions only account for a small portion of the overall image, leading to the delicate details under natural scenes that are still difficult to be learned. Therefore, some information may be potentially lost after the process of the first stage. Inspired by [1], [10], our model is subsequently extended with a Multi-Scale Refinement module further to improve the quality of the estimated alpha mattes. As depicted in the orange box in Fig. 3, we apply Atrous Convolutions [15] with different kernels to the preliminary alpha mattes obtained from the encoder-decoder stage to explore multi-scale features. The original image was also introduced at this stage to extract rich location and colour information to guide the learning process. The accuracy increase improved by our MSR can also be seen in Tab. VI.
D. Loss Function

Considering the Dynamic Gaussian Modulation mechanism, we abandon the Composition-L1 loss in DIM [1] and only resort to the Alpha prediction loss to achieve effective alpha matte optimization in the PIIAMatting. Our loss function is defined as follows:

$$L_{\text{alpha}} = R_i g \ast |\alpha_i^p - \alpha_i^g|.$$  \hspace{1cm} (7)

where $i$ denotes the pixel position, $\ast$ means the element-wise product, and the $R_i g$ represents the response coefficient at which pixel $i$ obtained according to the prior information in the Ground Truth. The $\alpha_i^p$ and $\alpha_i^g$ are the predicted alpha matte and ground truth alpha matte at pixel $i$, separately.

IV. Experimental Results

In this section, we conducted extensive experiments and evaluated our PIIAMatting on three challenging datasets to prove the effectiveness of it: (1) Alphamatting.com dataset [4], (2) Adobe Composition-1K dataset [1] and (3) Distinctions-646 dataset [11]. Firstly, we compare our PIIAMatting quantitatively and qualitatively with the current state-of-the-art methods. Then, we take Composition-1K [1] as an example and carry out the experimental analysis of the efficacy of each component in our PIIAMatting. Finally, we apply our PIIAMatting to some Real-World images to further validate our method and predict their alpha mattes.
A. Datasets

The first dataset is the Alphamatting.com dataset [4], which is the existing benchmark for image matting. It consists of 27 images with user-defined trimaps and alpha mattes, and 8 test with three different kinds of trimaps, namely, "small", "large", and "user". The second dataset is Composition-1K [1], which is the first public large-scale image matting dataset. The training set is composed of 431 foreground objects with the corresponding alpha mattes. Each foreground image is combined with 100 distinct backgrounds from MS COCO [50] to form the new synthesis. The test set consists of 50 pair of foreground objects and the corresponding alpha matte, and each foreground is composited with 20 different backgrounds from PASCAL VOC [51]. The third dataset is the Distinctions-646, which is the current biggest dataset for image matting. It provides 646 distinct foreground images with the corresponding alpha mattes and takes the same rule as DIM [1] to generate new images. We completely follow the composition rule provided by [1] and [11] when using the datasets.

B. Evaluation metrics

We evaluate the alpha mattes following four common quantitative metrics: the sum of absolute differences (SAD), mean square error (MSE), gradient (Grad), and connectivity (Conn) error proposed by [4]. The first two metrics measure the difference in absolute pixel space, while the last two metrics mean to weigh the visual quality of generated alpha mattes. Anyway, the lower the values of all metrics, the better the predicted alpha mattes.

C. Implementation Details

Our method is implemented with PyTorch [52] Toolbox. For training, we first randomly crop the images with a crop size of 320×320, 512×512, and 640×640, and then resize the images to 320×320. Trimaps are generated using either dilation/corrosion [1] or Euclidean distance [2] based on random selection, and the kernel sizes are randomly selected in [5, 25]. We also perform the data augmentation, including random scaling, flipping, and rotation between -60 and 60 degrees. For all experiments, we train for 50 epochs with a batch-size of 64. The Adam [53] optimizer was adopted with the initial learning rate of 0.001, and it was adjusted to 0.0001 and 0.00001 at the 30th and 40th epochs, respectively. The µ and σ² were initially set to 0.5 and 0.25. With the increase of iterations, the deviation σ² is adapted to make the Gaussian distribution flatter so that the model gradually treats all samples more fairly. The training time of our network on a GPU server (configuration: XEON Gold-6240 CPU, 32G RAM, and 3 Tesla V100 graphics card) is 3 days for the DIM [1] dataset and 3.5 days for the Distinctions-646 [11] dataset. During inference, the full-resolution input images with trimaps are concatenated as a 4-channel input fed into the network.

D. Comparison to Prior Work

1) Results on the Alphamatting.com dataset: We evaluate our method on Alphamatting.com [4], and the rank is shown in Table I. There are some SOTA methods listed for comparison. We can see that our method (PIIAMatting) achieves state-of-the-art performance, ranking in the first place for the average performance on all three metrics at the time of this submission. Some qualitative comparisons are presented in Figure. 5. Generally, our method is more robust to unknown regions of different sizes and can retain more detailed information. Compared with IndexNet [2] and GCA [3], our method has better performance on the "Plastic bag", "Net" and "Troll" image cases, as shown in Fig. 5.

2) Results on the Composition-1K dataset: We compare our method with 16 state-of-the-art methods, including 6 traditional colour-based methods: Shared Matting [18], Learning-based [54], Global Matting [13], Information-Flow [24], ClosedForm [7], KNN Matting [6], and 10 latest deep learning-based approaches: DCNN [5], DIM [1], AlphaGAN [51], SampleNet [8], IndexNet [2], AdaMatting [10], Context-Aware [9] and GCA [3]. For fair comparisons, we use the released codes and their default parameters to implement these methods. In terms of methods without released source code, we use their published results for comparisons. Quantitatively, as shown in Tab. I our approach achieves the SOTA results with respect to four metrics, compared with other counterparts. It demonstrates the superior performance of the proposed PIIAMatting. Specifically, Our model without using additional data augmentation (Ours) has surpassed the DIM, LFM, SampleNet, IndexNet, AdaMatting. Although the result of our model (Ours) is slightly inferior to the Context-Aware and GCA, the former utilizes the ResNet-101 [14] as its two-branch backbone that learns stronger representations, and the latter resorts to the data augmentation for the improvement. To enable a fair competition, we also equipped our model

| Methods          | SAD | MSE | Grad | Conn |
|------------------|-----|-----|------|------|
| Shared Matting [18] | 128.9 | 0.091 | 126.5 | 135.3 |
| Learning Based [54] | 113.9 | 0.048 | 91.6 | 122.2 |
| Global Matting [13] | 133.6 | 0.068 | 97.6 | 133.3 |
| Information-Flow [24] | 75.4 | 0.066 | 63.0 | - |
| DCNN [5] | 161.4 | 0.087 | 115.1 | 161.9 |
| KNN Matting [6] | 175.4 | 0.103 | 124.1 | 176.4 |
| ClosedForm [7] | 168.1 | 0.091 | 126.9 | 167.9 |
| AlphaGAN [51] | - | 0.031 | - | - |
| DIM [1] | 54.6 | 0.017 | 36.7 | 55.3 |
| LFM [12] | 49.0 | 0.020 | 34.3 | 50.6 |
| HAttMatting [11] | 48.8 | 0.016 | 25.3 | 48.6 |
| SampleNet [8] | 40.4 | 0.010 | - | - |
| IndexNet [2] | 45.8 | 0.013 | 25.9 | 43.3 |
| AdaMatting [10] | 41.7 | 0.010 | 16.8 | - |
| Context-Aware [9] | 35.8 | 0.008 | 17.3 | 33.2 |
| GCA [3] | 35.3 | 0.009 | 16.9 | 32.6 |
| PIIAMatting (Ours) | 36.4 | 0.009 | 16.9 | 31.5 |
| PIIAMatting (Ours-Big) | 35.8 | 0.009 | 16.2 | 34.7 |
| PIIAMatting (Ours-Data) | 35.2 | 0.009 | 16.9 | 33.1 |
with data augmentation and ResNet-101 \cite{resnet} denoted as Ours-Data, Ours-Big respectively. As expected, both Ours-Data and Ours-Big outperform the GCA and Context-Aware.

Qualitatively, Fig. 6 shows a visual comparison of our method against six top-performing approaches. For better visualization, we highlight the salient differences in the red circle on each image. It is worth noting that even the IndexNet \cite{indexnet} and GCA \cite{gca} can get smooth and performing results, but the effect is not ideal for the highly translucent pixels, and some artefacts can still be observed on their alpha mattes. On the contrary, the result of our model is more delicate on the undetermined domains, mainly due to the fact that our method takes the pixel domains disproportion and information alignment into account during the training. The DGM mechanism can alleviate the disproportion of discrepant pixel domains, while the IA strategy can lighten the information discrepancy. Both of them contribute together to achieve such good results.

3) Results on the Distinction-646 dataset: For the Distinctions-646 dataset, we take the same standards as above to make the comparisons, and here we mainly compare the following most representative methods: Learning Based \cite{learning}, Shared Matting \cite{shared}, Global Matting \cite{global}, DIM \cite{dim}, IndexNet \cite{indexnet} and HAttMatting \cite{hattmatting}. We first visually compare the results of the proposed method to the state-of-the-art image matting methods. Fig. 7 shows the results of DIM \cite{dim} and IndexNet \cite{indexnet} on the Distinctions-646 test set. We can see that DIM \cite{dim} and IndexNet \cite{indexnet} are incapable of handling the situation very well when the undetermined domain accounts for most of the unknown regions (see the first and second row). The DIM \cite{dim} and IndexNet \cite{indexnet} treat the distinct domains equally and thus neglect the opacity variation. While the IndexNet \cite{indexnet} can somewhat enhance the boundary details, it did not fully consider the relationship between semantic context and texture information on the adjacent layer-wise features. In contrast, our results show that the proposed method can notably enhance the adaptability of the model to opacity variation while retaining details that are easy to lose. We have also quantitatively compared our method to these methods. As shown in Tab. III the PIIAMatting consistently outperforms other methods on the Distinctions-646 test set with respect to four metrics, demonstrating the superior performance of our method on image matting.

E. Internal Analysis

In this section, with the Composition-1K dataset \cite{composition} as the training set, we detailedly explore the validity and generalization of different components in the proposed PIIAMatting. Analyses are conducted on the following three aspects:
TABLE IV: Analytical experiments of our model with different loss functions on the Composition-1K dataset [1]. Comp+ Alpha means the compositional loss+alpha loss. L1+L2 refers to the L2 loss is calculated when alpha is equal to 0 and 1, and L1 loss is calculated on the rest. Gaussian L1 represents the L1 loss combined with the static Gaussian mechanism. Gaussian L1-D means that the Dynamic Gaussian Modulation mechanism is attached to Gaussian-L1. All experiments here abandoned the use of the MSR.

| Distinct Loss Function | MSE↓ | SAD↓ |
|------------------------|------|------|
| Comp + Alpha Loss [1] | 0.0118 | 43.2 |
| L1 + L2 Loss [12]     | 0.0107 | 41.5 |
| Gaussian L1 (Ours)    | 0.0105 | 40.6 |
| Gaussian L1-D (Ours)  | 0.0102 | 40.3 |

TABLE V: Analysis of the DGM on the Composition-1K dataset. Ori loss represents the primal loss function used in the method. DGM indicates our Dynamic Gaussian Modulation mechanism, and ‘wo’ means disable the data augmentation.

| Baseline Model | Ori Loss | DGM | MSE | SAD |
|----------------|----------|-----|-----|-----|
| DIM [1]        | ✓        | ✓   | 0.0171 | 54.6 |
| DIM [1]        | ✓        | ✓   | 0.0152 | 52.4 |
| IndexNet [2]   | ✓        | ✓   | 0.0132 | 45.8 |
| IndexNet [2]   | ✓        | ✓   | 0.0128 | 45.2 |
| GCA (wo) [3]   | ✓        | ✓   | 0.0126 | 47.1 |
| GCA (wo) [3]   | ✓        | ✓   | 0.0117 | 44.7 |

(i) Dynamic Gaussian Modulation mechanism (DGM), (ii) Information Alignment strategy (IM) and (iii) Multi-Scale Refinement module (MSR).

Analysis on DGM. In this subsection, we investigate the role of DGM, which is also our core component.

1). The validity of DGM: In the following, we illustrate this efficacy by comparing different loss functions on our model. For a fair comparison and to prevent the disturbance from other components, we removed the MSR and only utilized the rest of components.

As shown in Tab. II, Alpha+Comp refers to the loss function applied in DIM [1], L1+L2 indicates the loss function in LFM [12], the Gaussian L1 is the static Gaussian Mechanism that $\mu = 0.5$ and $\sigma = 0.25$ in this submission. Gaussian L1-D denotes that the Dynamic Modulation mechanism is attached to Gaussian-L1. Our model combined with Gaussian L1 is capable of getting obvious improvements compared with the Comp+Alpha, especially the decrement of SAD from 43.2 to 40.6, which validates the ability of DGM to adapt to the opacity variation.

With the incorporation of the Gaussian-L1 dynamic mechanism, the degree of exploration of our model in the undetermined domain decreased progressively with the training progresses. We can observe that our model can obtain further improvement (e.g. the SAD decreased from 40.6 to 40.3). Meanwhile, the SAD of our PIIAMatting decreased by 1.2 compared to the L1 + L2 that used in LFM. This is mainly due to the L1 + L2 can only be regarded as a pre-defined static weighting loss function based on the different partition. Hence it can not handle properly with the variation of opacity at distinct domains.

2). The generalization of DGM: In what follows, we embed DGM in other SOTA methods (e.g. DIM [1], IndexNet [2], and GCA [3]) to prove its effectiveness and generalization.

As shown in Tab. V after combination with the DGM, all three methods achieved a certain degree of improvement. For DIM, the metric of MSE and SAD decreased by 0.0019 and 2.2, respectively. While for GCA, MSE and SAD decreased by 0.0009 and 2.3. Due to IndexNet’s low-capacity nature of its model demonstrated by the authors, the two evaluation metrics were just reduced by 0.0004 and 0.6 on IndexNet + DGM, respectively. In a word, we can see that the DGM can greatly increase the adaptability of the model by modulating the disproportionalitity between different domains, which also confirms the effectiveness.

Analysis on IA. To demonstrate the proposed the Information Alignment strategy efficiently retained the valuable details, we validate it from the following two sides: 1). Self-comparison with and without IA. 2). The generality of IA to other methods. We disable the MSR and employ only the Composition+Alpha loss on our model for a fair comparison.
TABLE VI: Analysis of the IA and MSR on the Composition-1K [1] dataset. ’w’ and ‘wo’ mean that our model combines and disables the corresponding components, respectively. The Refine refers to the refinement module in DIM [1].

| Methods         | SAD↓ | MSE↓ | Grad↓ | Conn↓ |
|-----------------|------|------|-------|-------|
| DIM [1]         | 54.6 | 0.017| 36.7  | 55.3  |
| DIM-IA [1]      | 39.9 | 0.015| 36.6  | 49.5  |
| DIM-MSR [1]     | 49.2 | 0.015| 34.5  | 51.7  |
| Ours-w/IA       | 43.2 | 0.012| 22.6  | 44.9  |
| Ours-wo/IA      | 53.8 | 0.015| 33.1  | 52.2  |
| Ours-wo/MSR     | 40.3 | 0.010| 20.4  | 37.1  |
| Ours-Refine     | 43.3 | 0.010| 19.5  | 36.0  |
| Ours (w/MSR)    | 36.4 | 0.009| 16.9  | 31.5  |

Since the IMM and IAM are the concrete implementations of IA, we remove the IMM and IAM and take the decoder in U-net [47] which is denoted Ours-wo/IA. As shown in Tab.1, the four metrics of Ours-w/IA decreased a lot compared to Ours-wo/IA after removing the IA, particularly the SAD dropped from 53.8 to 43.2 and the MSE decreased from 0.015 to 0.012. Such results confirm that our Information Alignment strategy is beneficial for image matting due to the ability to match and aggregate useful information during the sampling process. Furthermore, we apply our IA to DIM [1] to prove its generality. Specifically, we switch the original decoder in DIM with our IA and take the same training parameters and policy. We can see that the result of DIM-IA is far superior to the result of the fundamental DIM. It is worth noting that the improvements of 0.002 and 14.7 with respect to the metrics of MSE and SAD, which further demonstrate that the ability of IA to excavate valuable information is effective and efficient.

Analysis on MSR. In what follows, we evaluate the effectiveness of the Multi-Scale Refinement module compared to the Refinement module in DIM. Attention, all the following experiments are conducted on the Composition-1K dataset. We first use the refinement module of DIM to replace our MSR and denote it as Our-Refine. As shown in Tab. VI our MSR can bring improvements in all four metrics compared to the original Refinement module in DIM. In terms of the SAD and Grad, Ours (w/MSR) improves the Ours-Refine by 3.9 and 2.6, proving the necessity of multi-scale information for image matting. We further validate the availability by replacing the original refinement module in DIM with MSR, which was indicated as DIM-MSR. Since the original refinement module in DIM simply concatenated the image with the result of preceding stage to refine the alpha matte and neglects the multi-scale spatial information, the result of DIM-MSR is deservedly better than the vanilla DIM, especially the metrics increased by 0.002 and 2.2 on MSE and Grad.

F. Results on Real-World images

Fig. 8 shows how our algorithm performs on some internet images with a given trimap. All the results are acquired by our PIIAMatting that train on the Composition-1k [1] dataset. The results show that our method can produce delicate details around the boundary regions, such as the hairs and branches. Furthermore, due to the Dynamic Gaussian Modulation mechanism, our model can pay more attention to the highly uncertain regions and contribute to the excavation of structural details on varied domains.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a Prior-Induced Information Alignment network for image matting. It utilizes the Dynamic Gaussian Modulation mechanism to regulate the pixel-wise response according to the prior information and employs the Information Alignment strategy to match and aggregate potential valuable information efficiently. Finally, extensive experiments demonstrate that the proposed model can enhance the boundary details significantly and achieve a new SOTA performance on Alphamatting.com, Composition-1K, and Distinctions-646 dataset.

In the future, we will explore other efficient strategies to achieve high-quality alpha mattes. The recent evolves, such as NAS [55] and learnable parameters [56], can eliminate the limitations of the hand-designed strategy and adapt to the opacity variation more efficiently. It is also appealing to explore how to apply the dynamic Gaussian Modulation mechanism to video matting.

REFERENCES

[1] N. Xu, B. Price, S. Cohen, and T. Huang, “Deep image matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2970–2979.
[2] H. Lu, Y. Dai, C. Shen, and S. Xu, “Index networks,” IEEE Trans. Pattern Anal. Mach. Intell., 2020.
[3] Y. Li and H. Lu, “Natural image matting via guided contextual attention,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11 450–11 457.
[4] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, “A perceptually motivated online benchmark for image matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1826–1833.
[5] D. Cho, Y.-W. Tai, and I. S. Kweon, “Deep convolutional neural network for natural image matting using initial alpha mattes,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 1054–1067, Sep. 2018.
[6] Q. Chen, D. Li, and C.-K. Tang, “Knn matting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 9, pp. 2175–2188, Sep. 2013.
[7] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 2, pp. 228–242, Mar. 2007.
[8] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, “Learning-based sampling for natural image matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3055–3063.
[9] Q. Hou and F. Liu, “Context-aware image matting for simultaneous foreground and alpha estimation,” in Proc. Int. Conf. Comput. Vision, 2019, pp. 4130–4139.
[10] S. Cai, X. Zhang, H. Fan, H. Huang, J. Liu, J. Liu, J. Liu, J. Wang, and J. Sun, “Disentangled image matting,” in Proc. Int. Conf. Comput. Vision, 2019, pp. 8819–8828.
[11] Y. Qiao, Y. Liu, X. Yang, D. Zhou, M. Xu, Q. Zhang, and X. Wei, “Attention-guided hierarchical structure aggregation for image matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13 670–13 685.
[12] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu, “A late fusion cnn for digital matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7469–7478.
[13] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, “A global sampling method for alpha matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 2049–2056.
[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vision, 2018, pp. 801–818.
[16] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, “A bayesian approach to digital matting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 2, 2001, pp. II–II.
X. Feng, X. Liang, and Z. Zhang, “A cluster sampling method for image matting via sparse coding,” in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 204–219.

E. S. Gastal and M. M. Oliveira, “Shared sampling for real-time alpha matting,” in *Comput. Graph. Forum*, vol. 29, no. 2, May. 2010, pp. 573–584.

L. Karaman, A. Erdem, and E. Erdem, “Image matting with k1-divergence based sparse sampling,” in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 424–432.

E. Shahriari, D. Rajan, B. Price, and S. Cohen, “Improving image matting using comprehensive sampling sets,” in *Proc. IEEE Conf. Vis. Pattern Recognit.*, 2013, pp. 636–643.

J. Wang and M. F. Cohen, “Optimized color sampling for robust matting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

—, “An iterative optimization approach for unified image segmentation and matting,” in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 936–943.

A. Levin, A. Rav-Acha, and D. Lischinski, “Spectral matting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1699–1712, Oct. 2008.

Y. Aksoy, T. Ozan Aydin, and M. Pollefeys, “Designing effective iterative pixel information flow for natural image matting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 29–37.

L. Grady, T. Schiwietz, S. Aharon, and R. Westermann, “Random walks for interactive alpha-matting,” in *Proc. VIIHP*, May. 2005, pp. 423–429.

P. Lee and Y. Wu, “Nonlocal matting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2193–2200.

J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum, “Poisson matting,” in *ACM Trans. Graph.*, vol. 24, no. 3, pp. 315–321.

X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, “Deep automatic portrait matting,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 92–107.

D. Cho, S. Kim, Y.-W. Tai, and I. S. Kweon, “Automatic trimap generation and consistent matting for light-field images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1504–1517, Sep. 2016.

V. Badrinarayan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

S. Lutz, K. Ampliamitis, and A. Smolick, “Alphagan: Generative adversarial networks for natural image matting,” *arXiv preprint arXiv:1807.10088*, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. IEEE Conf. Adv. Neural Inf. Process. Sys.*, 2014, pp. 2672–2680.

X. Han, H. Yang, G. Xing, and Y. Liu, “Asymmetric joint gans for normalizing face illumination from a single image,” *IEEE Trans. Comput. Vis. Pattern Recognit.*, 2020.

G. Chen, K. Han, and K.-Y. K. Wong, “Learning transparent object matting,” *Int. J. Comp. Vis.*, vol. 127, no. 10, pp. 1527–1544, 2019.

Q. Chen, T. Ge, Y.-Xu, Z. Zhang, X. Yang, and K. Gai, “Semantic human matting,” Aug. 2004, pp. 315–324.

X. Yang, K. Xu, S. Chen, S. He, B. Y. Yin, and R. Lau, “Active matting,” in *Proc. IEEE Conf. Adv. Neural Inf. Process. Sys.*, 2018, pp. 4950–4960.

Y. Wang, Y. Niu, P. Duan, J. Lin, and Y. Zheng, “Deep propagation based image matting,” in *Proc. Int. Joint. Conf. Artif. Intell.*, 2018, pp. 990–1006.

Y. Aksoy, T.-H. Oh, S. Paris, M. Pollefeys, and W. Matusik, “Semantic soft segmentation,” *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–13, Jul. 2018.

J. Liu, Y. Yao, B. Hu, M. Cui, X. Xie, C. Zhang, and X.-s. Hua, “Boosting semantic human matting with coarse annotations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8563–8572.

S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Background matting: The world is your green screen,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2291–2300.

S. He, Z. Zhou, F. Farhat, and J. Z. Wang, “Discovering triangles in portraits for supporting photographic creation,” *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 496–508, Aug. 2017.

H. Li, K. N. Ngan, and Q. Liu, “Faceseg: automatic face segmentation for real-time video,” *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 77–88, Dec. 2008.

H. Zhu, J. Lu, J. Cai, J. Zheng, S. Lu, and N. M. Thalmann, “Multiple human identification and cosegmentation: A human-oriented crf approach with poselets,” *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1516–1530, May. 2016.

Y. Qiao, Y. Liu, Q. Zhu, X. Yang, Y. Wang, Q. Zhang, and X. Wei, “Multi-scale information assembly for image matting,” vol. 39, no. 7, pp. 565–574, Nov. 2020.

X. Yang, Y. Qiao, S. Chen, S. He, B. Yin, Q. Zhang, X. Wei, and R. W. H. Lau, “Smart scribbles for image matting,” *ACM TOMM*, vol. 16, no. 4, 2020.

M. Ehrlich and I. S. Davis, “Deep residual learning in the jpeg transform domain,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3484–3493.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Int. Conf. Med. Image Comput. Assisted Intervention*, 2015, pp. 234–241.

X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. Wei, “Drfn: Deep recurrent fusion network for single-image super-resolution with large factors,” *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 328–337, 2018.

K. Xu, X. Yang, B. Yin, and R. W. Lau, “Learning to restore low-light images via decomposition-and-enhancement,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2281–2290.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *Int. J. Comp. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., “Pytorch: An imperative style, high-performance deep learning library,” 2019.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Int. Conf. Learn. Repres.*, 2015.

Y. Zheng and C. RamaChammettu, “Learning based digital matting,” in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 889–896.

P. Liu, B. Wu, H. Ma, and M. Seok, “Memnas: Memory-efficient neural architecture search with grow-trim learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.

A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.

Yu Hao Liu received the B.Eng. degree from Zhengzhou University in 2019, and now he is a second-year master student majoring in computer science at Dalian University of Technology. His current research interests focus on low level computer vision problems and deep learning.

Jiakie Xie is the Computer Vision Algorithm Engineer in Hangzhou Winroad Holdings Ltd. He received his bachelor’s degree from Zhengzhou University of Light Industry. He specializes in computer vision related deep learning algorithms with research interests in object detection, semantic segmentation, salient object detection, image matting, and video matting.

Yu Qiao received his B.Eng from Dalian University of Technology in 2017, and takes a successive postgraduate and doctoral program in computer science. His research interest includes computer vision and image processing.
Yong Tang is the Vice President in Hangzhou Winroad Holdings Ltd, where he leads the Libai Artificial Intelligence Lab and the department of Human-Machine Engineering. He received his PHD degree from the Pennsylvania State University majoring in Computational Science and Civil Engineering. He obtained M.S. and B.E. degrees from Tsinghua University. He published articles in the world’s top magazines and conference journals and was awarded with the Best Paper Awards for multiple times. He also peer reviewed articles for magazines in world-renowned professional fields. His research interests include computer vision, deep learning, text to speech synthesis, natural language processing, multi-objective optimization, and parallel computing.

Xin Yang is a Professor in the Department of Computer Science at Dalian University of Technology, China. Xin received his B.S. degree in Computer Science from Jilin University in 2007. From 2007 to June 2012, he was a joint Ph.D. student in Zhejiang University and UC Davis for Graphics, and received his Ph.D. degree in July 2012. His research interests include computer graphics and robotic vision.