Natural Image Matting via Guided Contextual Attention

Yaoyi Li, Hongtao Lu*
Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{dsamuel, htlu}@sjtu.edu.cn

Abstract

Over the last few years, deep learning based approaches have achieved outstanding improvements in natural image matting. Many of these methods can generate visually plausible alpha estimations, but typically yield blurry structures or textures in the semitransparent area. This is due to the local ambiguity of transparent objects. One possible solution is to leverage the far-surrounding information to estimate the local opacity. Traditional affinity-based methods often suffer from the high computational complexity, which are not suitable for high resolution alpha estimation. Inspired by affinity-based method and the successes of contextual attention in inpainting, we develop a novel end-to-end approach for natural image matting with a guided contextual attention module, which is specifically designed for image matting. Guided contextual attention module directly propagates high-level opacity information globally based on the learned low-level affinity. The proposed method can mimic information flow of affinity-based methods and utilize rich features learned by deep neural networks simultaneously. Experiment results on Composition-1k testing set and alphamatting.com benchmark dataset demonstrate that our method outperforms state-of-the-art approaches in natural image matting. Code and models are available at https://github.com/Yaoyi-Li/GCA-Matting.

Introduction

The natural image matting is one of the important tasks in computer vision. It has a variety of applications in image or video editing, compositing and film post-production (Wang, Cohen, and others 2008; Aksoy, Ozan Aydin, and Pollefeys 2017; Lutz, Amplianitis, and Smolic 2018; Xu et al. 2017; Tang et al. 2019). Matting has received significant interest from the research community and been extensively studied in the past decade. Alpha matting refers to the problem that separating a foreground object from the background and estimating transitions between them. The result of image matting is a prediction of alpha matte which represents the opacity of a foreground at each pixel.

Mathematically, the natural image I is defined as a convex combination of foreground image F and background image B at each pixel i as:

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

where α_i is the alpha value at pixel i that denotes the opacity of the foreground object. If α_i is not 0 or 1, then the image at pixel i is mixed. Since the foreground color F_i, background color B_i and the alpha value α_i are left unknown, the expression of alpha matting is ill-defined. Thus, most of the previous hand-crafted algorithms impose a strong inductive bias to the matting problem.

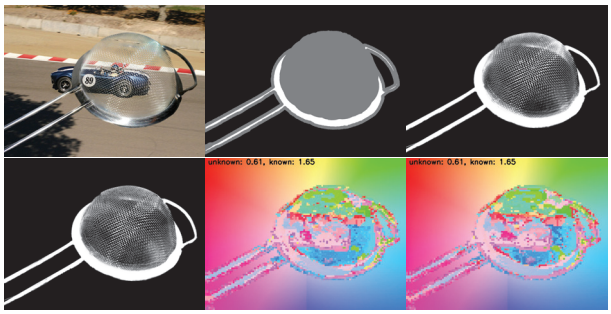

Figure 1: The visualization of our guided contextual attention map. Top row from left to right, the image, trimap and ground-truth. Second row, the alpha matte prediction, attention offset map from first GCA block in the encoder, offset from GCA block in the decoder.

One of the basic idea widely adopted in both affinity-based and sampling-based algorithms is to borrow information from the image patches with similar appearance. Affinity-based methods (Levin, Lischinski, and Weiss 2008; Chen, Li, and Tang 2013; Aksoy, Ozan Aydin, and Pollefeys 2017) borrow the opacity information from known patches with the similar appearance to unknown ones. Sampling-based approaches (Wang and Cohen 2007; Gastal and Oliveira 2010; He et al. 2011; Feng, Liang, and Zhang 2016) borrow a pair of samples from the foreground and background to estimate the alpha value at each pixel in the unknown region based on some specific assumption. One obstacle of the previous affinity-based and sampling-based methods is that they cannot handle the situation that there are only background and unknown areas in the trimap. It is

*Corresponding author

because that these methods have to make use of both foreground and background information to estimate the alpha matte.

Benefiting from the Adobe Image Matting dataset (Xu et al. 2017), more learning-based image matting methods (Xu et al. 2017; Lutz, Amplianitis, and Smolic 2018; Lu et al. 2019; Tang et al. 2019) has emerged in recent years. Most of learning-based approaches use network prior as the inductive bias and predict alpha mattes directly. Moreover, SampleNet (Tang et al. 2019) proposed to leverage deep inpainting methods to generate foreground and background pixels in the unknown region rather than select from the image. It provides a combination of the learning-based and sampling-based approach.

In this paper, we propose a novel image matting method based on the opacity propagation in a neural network. The information propagation has been widely adopted within the neural network framework in recent years, from natural language processing (Vaswani et al. 2017; Yang et al. 2019), data mining (Kipf and Welling 2016; Veličković et al. 2017) to computer vision (Yu et al. 2018; Wang et al. 2018). SampleNet Matting (Tang et al. 2019) indirectly leveraged the contextual information for foreground and background inpainting. In contrast, our proposed method conducts information flow from the image context to unknown pixels directly. We devise a guided contextual attention module, which mimic the affinity-based propagation in a fully convolutional network. In this module, the low-level image features are used as a guidance and we perform the alpha feature transmission based on the guidance. We show an example of our guided contextual attention map in Figure 1 and more details in the section of results. In the guided contextual attention module, features from two distinct network branches are leveraged together. The information of both known and unknown patches are transmitted to feature patches in the unknown region with similar appearance.

Our proposed method can be viewed from two different perspectives. On one hand, the guided contextual attention can be elucidated as an affinity-based method for alpha matte value transmission with a network prior. Unknown patches share high-level alpha features with each other under the guidance of similarity between low-level image features. On the other hand, the proposed approach can also be seen as a guided inpainting task. In this aspect, image matting task is treated as an inpainting task on the alpha image under the guidance of input image. The unknown region is analogous to the holes to be filled in image inpainting. Unlike inpainting methods which borrows pixels from background of the same image, image matting borrows pixel value 0 or 1 from the known area in the alpha matte image under the guidance of original RGB image to fill in the unknown region.

Related Work

In general, natural image matting methods can be classified into three categories: sampling-based methods, propagation methods and learning-based methods.

Sampling-based methods (Wang and Cohen 2007; Gastal and Oliveira 2010; He et al. 2011; Feng, Liang, and Zhang 2016) solve combination equation (1) by sampling colors from foreground and background regions for each pixel in the unknown region. The pair of foreground and background samples are selected under different metrics and assumptions. Then the initial alpha matte value is calculated by the combination equation. Robust Matting (Wang and Cohen 2007) selected samples along the boundaries with confidence. The matting function was optimized by a Random Walk. Shared Matting (Gastal and Oliveira 2010) selected the best pairs of samples for a set of neighbor pixels and reduced much of redundant computation cost. In Global Matting (He et al. 2011), all samples available in image were utilized to estimate the alpha matte. The sampling was achieve by a randomized patch match algorithm. More recently, CSC Matting (Feng, Liang, and Zhang 2016) collected a set of more representative samples by sparse coding to avoid missing out true sample pairs.

Propagation methods (Levin, Lischinski, and Weiss 2008; Chen, Li, and Tang 2013; Aksoy, Ozan Aydin, and Pollefeys 2017), which are also known as affinity-based methods, estimate alpha mattes by propagating the alpha value from foreground and background to each pixel in the unknown area. The Closed-form Matting (Levin, Lischinski, and Weiss 2008) is one of the most prevailing algorithm in propagation-based methods. It solved the cost function under the constraint of local smoothness. KNN Matting (Chen, Li, and Tang 2013) collected matching nonlocal neighborhoods globally by K nearest neighbors. Moreover, the Information-flow Matting (Aksoy, Ozan Aydin, and Pollefeys 2017) proposed a color-mixture flow which combined the local and nonlocal affinities of colors and spatial smoothness.

Due to the tremendous success of deep convolutional neural networks, learning-based methods achieve a dominate position in recent natural image matting (Cho, Tai, and Kweon 2016; Xu et al. 2017; Lutz, Amplianitis, and Smolic 2018; Lu et al. 2019; Tang et al. 2019). DCNN Matting (Cho, Tai, and Kweon 2016) is the first method that introduced a deep neural network into image matting task. It made use of the network to learn a combination of results from different previous methods. Deep Matting (Xu et al. 2017) proposed a fully neural network model with a large-scale dataset for learning-based matting methods, which was one of the most significant work in deep image matting. Following Deep Matting, AlphaGan (Lutz, Amplianitis, and Smolic 2018) explored the deep image matting within a generative adversarial framework. More subsequent work like SampleNet Matting (Tang et al. 2019) and IndexNet (Lu et al. 2019) with different architectures also yielded appealing alpha matte estimations.

Baseline Network for Deep Image Matting

Our proposed model uses the guided contextual attention module and a customized U-Net (Ronneberger, Fischer, and Brox 2015) architecture to perform deep natural image matting. We first construct our customized U-Net baseline for matting, then introduce the proposed guided contextual attention (GCA) module.

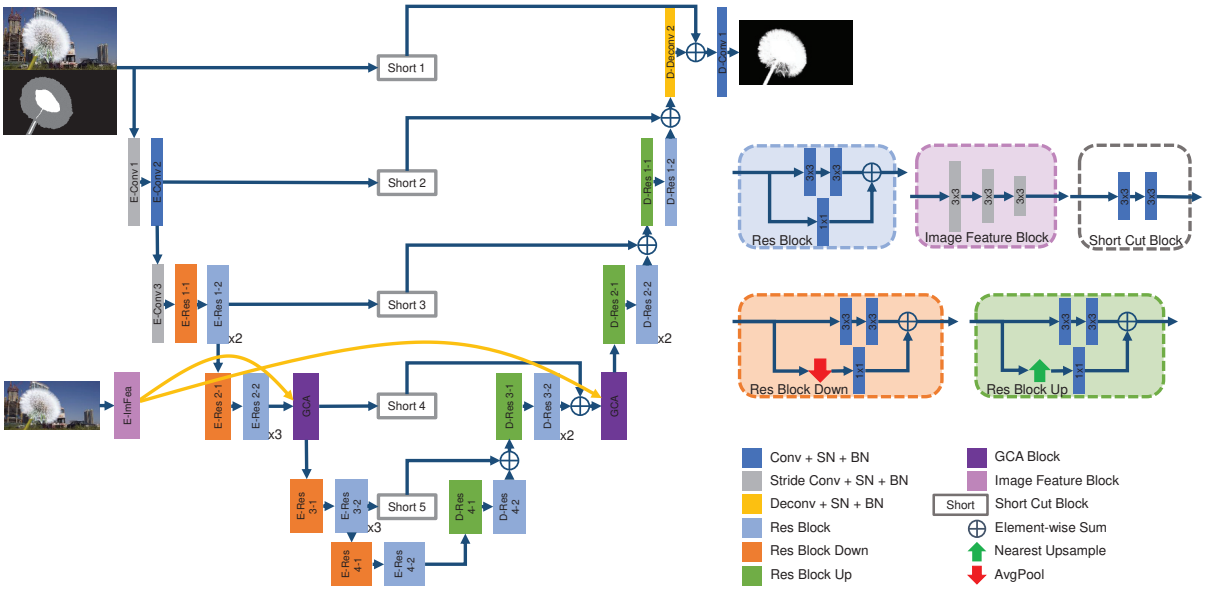

Figure 2: Overview of our proposed guided contextual attention matting framework. The baseline model shares the same architecture without GCA blocks and image feature block. Original image and trimap are the inputs of alpha feature. Image feature block and GCA blocks only takes the original merged image as input. The blue arrows denote alpha feature flow and yellow arrows denote low-level image feature flow. GCA: guided contextual attention; SN: spectral normalization; BN: batch normalization; $\times N$: replicate N times.

Baseline Structure

The U-Net (Ronneberger, Fischer, and Brox 2015) like architecture are prevailing in recent matting tasks (Lutz, Amplianitis, and Smolic 2018; Tang et al. 2019; Lu et al. 2019) as well as image segmentation (Long, Shelhamer, and Darrell 2015), image-to-image translation (Isola et al. 2017) and image inpainting (Liu et al. 2018). Our baseline model shares almost the same network architecture with guided contextual attention framework in Figure 2. The only difference is that the baseline model replaces GCA blocks with identity layers and has no image feature block. The input to this baseline network is a cropped image patch and a 3-channel one-hot trimap which are concatenated as a 6-channel input. The output is corresponding estimated alpha matte. The baseline structure is built as an encoder-decoder network with stacked residual blocks (He et al. 2016).

Since the low-level features play a crucial role in retaining the detailed texture information in alpha mattes, in our customized baseline model, the decoder combines encoder features just before upsampling blocks instead of after each upsampling block. Such a design can avoid more convolutions on the encoder features, which are supposed to provide lower-level feature. We also use a two layer short cut block to align channels of encoder features for feature fusion. Moreover, in contrast to the typical U-Net structure which only combines different middle-level features, we directly forward the original input to the last convolutional layer through a short cut block instead. These features do not share any computation with the stem. Hence, this short cut branch only focuses on detailed textures and gradients.

In addition to the widely used batch normalization (Ioffe and Szegedy 2015), we introduce the spectral normalization (Miyato et al. 2018) to each convolutional layer to add a constraint on Lipschitz constant of the network and stable the training, which is prevalent in image generation tasks (Brock, Donahue, and Simonyan 2019; Zhang et al. 2019).

Loss Function

Our network only leverages one alpha prediction loss. The alpha prediction loss is defined as an absolute difference between predicted and ground-truth alpha matte averaged over the unknown area:

$$\mathcal{L} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} |\hat{\alpha}_i - \alpha_i|, \tag{2}$$

where $\mathcal{U}$ indicates the region labeled as unknown in the trimap, $\hat{\alpha}_i$ and α_i denote the predicted and ground-truth value of alpha matte as position i.

There are some losses proposed in prior work for the deep image matting tasks, like compositional loss (Xu et al. 2017), gradient loss (Tang et al. 2019) and Gabor loss (Li et al. 2019). Compositional loss used in Deep Matting (Xu et al. 2017) is the absolute difference between the original input image and predicted image composited by the ground-truth foreground, background and the predicted alpha mattes. The gradient loss calculates the averaged absolute difference between the gradient magnitude of predicted and ground-truth alpha mattes in the unknown region. Gabor loss proposed in (Li et al. 2019) substitutes the gradient operator with a bundle of Gabor filters and aims to have a

Table 1: Ablation study on data augmentation and different loss functions with baseline structure. The quantitative results are tested on Composition-1k testing set. Aug: data augmentation; Rec: alpha prediction loss; Comp: compositional loss; GradL: gradient loss; Gabor: Gabor loss.

Aug	Rec	Comp	GradL	Gabor	MSE	Grad
✓	✓				0.0106	21.53
✓	✓	✓			0.0107	21.85
✓	✓		✓		0.0108	22.51
✓	✓			✓	0.0109	20.66
	✓				0.0146	32.01

more comprehensive supervision on textures and gradients than gradient loss.

We delve into these losses to reveal whether involving different losses can benefit the alpha matte estimation in our baseline model. We provide an ablation study on Composition-1k testing set (Xu et al. 2017) in Table 1. As Table 1 shows, the use of compositional loss does not bring any notable difference under MSE and Gradient error, and both errors increase when we incorporate the gradient loss and alpha prediction loss. Although the adoption of Gabor loss can reduce the Gradient error to some degree, it also slightly increases the MSE. Consequently, we only opt for the alpha prediction loss in our model.

Data Augmentation

Since the most dominant image matting dataset proposed by Xu et al. only contains 431 foreground objects for training. We treat the data augmentation as a necessity of our baseline model. We introduce a sequence of data augmentation.

Firstly, following the data augmentation in (Tang et al. 2019), we randomly select two foreground object images with a probability of 0.5 and combine them to obtain a new foreground object as well as a new alpha image. Subsequently, the foreground object and alpha image will be resized to 640×640 images with a probability of 0.25. In this way, the network can nearly see the whole foreground image instead of a cropped snippet. Then, a random affine transformation are applied to the foreground image and the corresponding alpha image. We define a random rotation, scaling, shearing as well as the vertical and horizontal flipping in this affine transformation. Afterwards, trimaps are generated by a dilation and an erosion on alpha images with random number of pixels ranging from 5 to 29. With the trimap obtained, we randomly crop one 512×512 patch from each foreground image, corresponding alpha and trimap respectively. All of the cropped patches are centered on an unknown region. The foreground images are then converted to HSV space, and different jitters are imposed to the hue, saturation and value. Finally, we randomly select one background image from MS COCO dataset (Lin et al. 2014) for each foreground patch and composite them to get the input image.

To demonstrate the effectiveness of data augmentation, we conduct an experiment with minimal data augmentation. In this case, only two necessary operations, image cropping and trimap dilation are retained. More augmentations like random image resize and flipping, which are widely used in most of previous deep image matting methods (Xu et al. 2017; Lutz, Amplianitis, and Smolic 2018; Tang et al. 2019; Lu et al. 2019), are not included in this experiment. We treat this experiment setting as no data augmentation. The experimental results are also listed in Table 1. We can see that without additional augmentation, our baseline model already achieves comparable performance with Deep Matting.

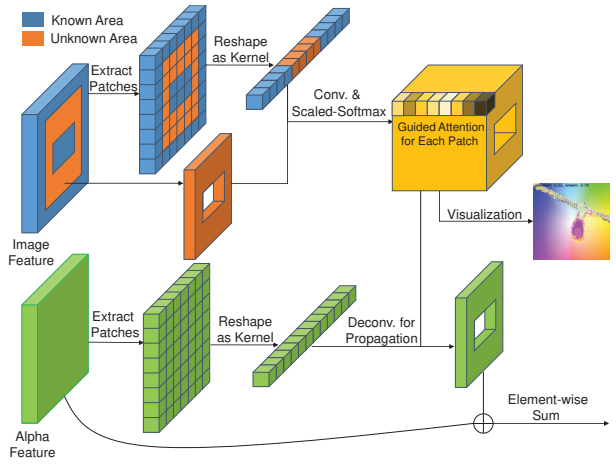

Figure 3: The illustration of the guided contextual attention block. Computation is implemented as a convolution or a deconvolution. Two additional 1×1 convolutional layers for adaptation are not shown in this figure to keep neat. One is applied to the input image feature before extracting patches, and the other one is applied to the result of propagation before the element-wise summation.

Guided Contextual Attention Module

The guided contextual attention module contains two kinds of components, an image feature extractor block for low-level image feature and one or more guided contextual attention blocks for information propagation.

Low-level Image Feature

Most of the affinity-based approaches have a basic inductive bias that local regions with almost identical appearance should have similar opacity. This inductive bias allows the alpha value propagates from the known region of a trimap to the unknown region based on affinity graph, which can often yields impressive alpha matte prediction.

Motivated by this, we define two different feature flows in our framework (Figure 2): alpha feature flow (blue arrows) and image feature flow (yellow arrows). Alpha features are generated from the 6-channel input which is a concatenation of original image and trimap. The final alpha matte can be predicted directly from alpha features. Low-level image features contrast with the high-level alpha features. These features are generated only from the input image by a sequence of three convolutional layer with stride 2, which are analogous to the local color statistics in conventional affinity-based methods.

In other words, the alpha feature contains opacity information and low-level image feature contains appearance information. Given both opacity and appearance information, we can build an affinity graph and carry out opacity propagation as affinity-based methods. Specifically, we utilize the low-level image feature to guide the information flow on alpha features.

Guided Contextual Attention

Inspired by the contextual attention for image inpainting proposed in (Yu et al. 2018), we introduce our guided contextual attention block.

As shown in Figure 3, the guided contextual attention leverages both the image feature and alpha feature. Firstly, the image feature are divided into known part and unknown part and 3×3 patches are extracted from the whole image feature. Each feature patch represents the appearance information at a specific position. We reshape the patches as convolutional kernels. In order to measure the correlation between an unknown region patch $U_{x,y}$ centered on (x, y) and an image feature patch $I_{x',y'}$ centered on (x', y') , the similarity is defined as the normalized inner product:

$$s_{(x,y),(x',y')} = \begin{cases} \lambda & (x,y) = (x',y'); \\ \langle \frac{U_{x,y}}{\|U_{x,y}\|}, \frac{I_{x',y'}}{\|I_{x',y'}\|} \rangle & \text{otherwise}, \end{cases} \quad (3)$$

where $U_{x,y} \in \mathcal{U}$ is also an element of the image feature patch set $\mathcal{I}$, i.e. $\mathcal{U} \subseteq \mathcal{I}$. The constant λ is a punishment hyperparameter that we use -10^4 in our model, which can avoid a large correlation between each unknown patch and itself. In implementation, this similarity is computed by a convolution between unknown region features and kernels reshaped from image feature patches. Given the correlation, we carry out a scaled softmax along (x', y') dimension to attain the guided attention score for each patch as following,

$$a_{(x,y),(x',y')} = \text{softmax}(w(\mathcal{U}, \mathcal{K}, x', y') s_{(x,y),(x',y')}), \quad (4)$$

$$w(\mathcal{U}, \mathcal{K}, x', y') = \begin{cases} \text{clamp}(\sqrt{\frac{|\mathcal{U}|}{|\mathcal{K}|}}) & I_{x',y'} \in \mathcal{U}; \\ \text{clamp}(\sqrt{\frac{|\mathcal{K}|}{|\mathcal{U}|}}) & I_{x',y'} \in \mathcal{K}, \end{cases} \quad (5)$$

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

in which $w(\cdot)$ is a weight function and $\mathcal{K} = \mathcal{I} - \mathcal{U}$ is the set of image feature patches from known region. As distinct from image inpainting task, the area of unknown region in a trimap is not under control. In many input trimaps, there are overwhelming unknown region and scarcely any known pixel. Thus, typically it is not feasible that only propagate the opacity information from the known region to unknown part. In our guided contextual attention, we let the unknown part borrow features from both known patches and unknown ones. Different weights are assigned to known and unknown patches based on the area of each region as the weight function defined in Eq. (5). If the area of known region is larger, the known patches can convey more accurate appearance information which exposes the difference between foreground and background, hence we weigh known patches with a larger weight. Whereas, if the unknown region has an overwhelming area, the known patches only provide some local appearance information, which may harm the opacity propagation. Then a small weight is assigned to known patches.

When we get guided attention scores from image features, we do the propagation on alpha features based on the affinity graph defined by guided attention. Analogous to image features, patches are extracted and reshaped as filter kernels from alpha features. The information propagation is implemented as a deconvolution between guided attention scores and reshaped alpha feature patches. This deconvolution yields a reconstruction of alpha features in the unknown area and the values of overlapped pixels in the deconvolution are averaged. Finally, we combine the input alpha features and the propagation result by an element-wise summation. This element-wise summation works as a residual connection which can stable the training.

Table 2: The quantitative results on Composition-1k testing set. Best results are emphasized in bold. (- indicates not given in the original paper.)

Methods	MSE	SAD	Grad	Conn
Learning Based Matting	0.048	113.9	91.6	122.2
Closed-Form Matting	0.091	168.1	126.9	167.9
KNN Matting	0.103	175.4	124.1	176.4
Deep Matting	0.014	50.4	31.0	50.8
IndexNet Matting	0.013	45.8	25.9	43.7
SampleNet Matting	0.0099	40.35	-	-
Baseline	0.0106	40.62	21.53	38.43
Ours	**0.0091**	**35.28**	**16.92**	**32.53**

Network with Guided Contextual Attention

Most of the affinity-based matting methods result in a closed-form solution based on the graph Laplacian (Levin, Lischinski, and Weiss 2008; Lee and Wu 2011; Chen, Li, and Tang 2013). The closed-form solution can be seen as a fixed point of the propagation or a limitation of infinite propagation iterations (Zhou et al. 2004). Motivated by this, we stick two guided contextual attention blocks to the encoder and decoder symmetrically in our stem. It aims to propagate more times in our model and take full advantage of the opacity information flow.

When we compute the guided contextual attention on higher-resolution features, more detailed appearance information will be attended. However, on the other hand, the computational complexity of the attention block is $O(c(hw)^2)$, where c, h, w are the channels, height and width of the feature map respectively. Therefore, we append two guided contextual attention blocks to the stage with 64×64 feature maps.

The network is trained for $200,000$ iterations with a batch size of 40 in total on the Adobe Image Matting dataset (Xu et al. 2017). We perform optimization using Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is initialized to 10^{-4}. Warmup and cosine decay (Loshchilov and Hutter 2016; Goyal et al. 2017; He et al. 2019) are applied to the learning rate.

Figure 4: The visual comparison results on Adobe Composition-1k. From left to right, the original image, trimap, ground-truth, Deep Matting (Xu et al. 2017), IndexNet Matting (Lu et al. 2019), baseline and ours.

Results

In this section we report the evaluation results of our proposed model on two datasets, the Composition-1k testing set and alphamatting.com dataset. Both quantitative and qualitative results are shown in this section. We evaluate the quantitative results under the Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient error (Grad) and Connectivity error (Conn) proposed by (Rhemann et al. 2009).

Composition-1k Testing Dataset

The Composition-1k testing dataset proposed in (Xu et al. 2017) contains 1000 testing images which are composed from 50 foreground objects and 1000 different background images from Pascal VOC dataset (Everingham et al. 2015).

We compare our approach and the baseline model with three state-of-the-art deep image matting methods: Deep Matting (Xu et al. 2017), IndexNet Matting (Lu et al. 2019) and SampleNet Matting (Tang et al. 2019), as well as three conventional hand-crafted algorithms: Learning Based Matting (Zheng and Kambhamettu 2009), Closed-Form Matting (Levin, Lischinski, and Weiss 2008) and KNN Matting (Chen, Li, and Tang 2013). The quantitative results are shown in Table 2. Our method outperforms all of the state-of-the-art approaches. In addition, our baseline model also get better results than some of the top performing methods. The effectiveness of the proposed guided contextual attention can be validated by the results displayed in Table 2.

Some qualitative results are given in Figure 4. The results of Deep Matting and IndexNet Matting are generated by source codes and pretrained models provided in (Lu et al. 2019). As displayed in Figure 4, our approach achieves better performance on different foreground objects, especially in the semitransparent regions. Advantages are more obvious with a larger unknown region. This good performance profits from the information flow between feature patches with similar appearance features.

Additionally, our proposed method can evaluate each image in Composition-1k testing dataset as a whole on a single Nvidia GTX 1080 with 8GB memory. Since we take each image as a whole in our network without scaling, the guided contextual attention blocks are applied to feature maps with a much higher resolution than 64×64 in training phase. This results in a better performance in the detailed texture.

Alphamatting.com Benchmark dataset

The alphamatting.com benchmark dataset (Rhemann et al. 2009) has eight different images. For each testing image, there are three corresponding trimaps, namely, ”small”, ”large” and ”user”. The methods on the benchmark are ranked by the averaged rank over 24 alpha matte estimations in terms of four different metrics. We evaluate our method on the the alphamatting.com benchmark, and show the scores in Table 3. Some top approaches in the benchmark are also displayed for comparison.

As displayed in Table 3, GCA Matting ranks the first place under the Gradient Error metric in the benchmark. The evaluation results of our method under the ”large” and ”user” trimaps are much better than the other top approaches. The image matting becomes more difficult as the trimap has a larger unknown region. Therefore, we can say that our approach is more robust to changes in the area of unknown region. Additionally, our approach has almost the same overall ranks with the SampleNet under the MSE metric. Generally, the proposed GCA Matting is one of the top performing

Table 3: Our scores in the alpha matting benchmark, S, L and U denote the three trimap types, small, large and user, included in the benchmark. (Bold numbers indicate scores which rank the 1st place in the benchmark at the time of submission)

Gradient Error	Average Rank				Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
	Overall	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	**5.2**	5	4	6.5	0.1	**0.1**	0.2	0.1	0.1	0.3	0.2	0.2	0.2	0.2	0.2	**0.3**	1.3	1.6	1.9	0.7	0.8	1.4	**0.6**	**0.7**	**0.6**	0.4	0.4	0.4
SampleNet Matting	7.2	3.6	4.4	13.6	**0.1**	0.1	0.2	0.1	0.1	0.2	0.2	0.3	0.3	0.1	0.2	0.5	1.1	1.5	2.7	0.6	0.9	1	0.8	0.9	0.9	0.4	0.4	0.4
IndexNet Matting	10.3	8.6	8.8	13.6	0.2	0.2	0.2	0.1	0.1	0.3	0.2	0.2	**0.2**	0.2	0.2	0.4	1.7	1.9	2.5	1	1.1	1.3	1.1	1.2	1.2	0.4	0.5	0.5
AlphaGAN	14.9	13.6	12.5	18.5	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.3	0.3	0.2	0.2	0.4	1.8	2.4	2.7	1.1	1.4	1.5	0.9	1.1	1	0.5	0.5	0.6
Deep Matting	15.6	12	12.3	22.5	0.4	0.4	0.5	0.2	0.2	**0.2**	**0.1**	**0.1**	0.2	0.2	0.2	0.6	1.3	1.5	2.4	0.8	0.9	1.3	0.7	0.8	1.1	0.4	0.5	0.5
Information-flow matting	18.3	21.5	16.5	16.8	0.2	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.4	0.3	0.4	0.4	1.7	1.8	2.2	0.9	1.3	1.3	1.5	1.4	0.8	0.5	0.6	0.5

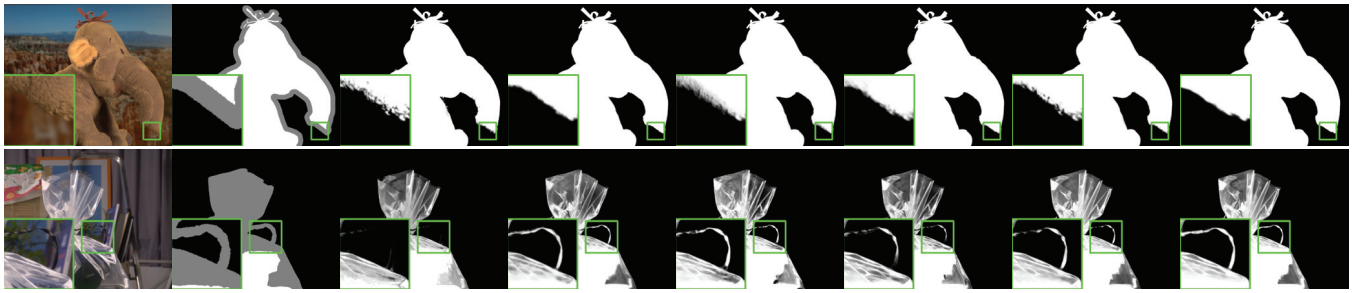

Figure 5: The alpha matte predictions of the test images from alphamatting.com benchmark. From left to right, the original image, trimap, Information-flow Matting (Aksoy, Ozan Aydin, and Pollefeys 2017) , Deep Matting (Xu et al. 2017), AlphaGAN (Lutz, Amplianitis, and Smolic 2018), IndexNet Matting (Lu et al. 2019), SampleNet Matting (Tang et al. 2019) and ours.

method on this benchmark dataset.

We provide some of the visual examples in Figure 5. The results of our method and some top algorithms on "Elephant" and "Plastic bag" are displayed to demonstrate the good performance of our approach. For example, in the test image "Plastic bag", most of the previous methods make a mistake at the iron wire. However, our method learns from the contextual information in the surrounding background patches and predicts these pixels correctly.

Visualization of Attention Map

We visualize the attention map learned in the guided contextual attention block by demonstrating the pixel position with the largest attention score. Unlike the offset map widely used in optical flow estimation (Dosovitskiy et al. 2015; Hui, Tang, and Loy 2018; Sun et al. 2018) and image inpainting (Yu et al. 2018) which indicates the relative displacement of each pixel, our attention map demonstrates the absolute position of the corresponding pixel with highest attention activation. From this attention map, we can easily identify where the opacity information is propagated from for each feature pixel. As we can see in Figure 1, there is no information flow in the known region and feature patches in the unknown region tend to borrow information from the patches with similar appearance. Figure 1 reveals where our GCA blocks attend to physically in the input image. Since there is an adaption convolutional layer in the guided contextual attention block before patch extraction on image features, attention maps from two attention blocks are not identical. The weights of known and unknown part are shown in the top-left corner of the attention map.

From the attention offset map in Figure 1, we can easily recognize the car in the sieve. The light pink patches at the center of the sieve indicate that these features are propagated from the left part of the car. While blue patches show the features which are borrowed from the right-hand side road. These propagated features will assist in the identification of foreground and background in ensuing convolutional layers.

Conclusions

In this paper, we propose to solve the image matting problem by opacity information propagation in an end-to-end neural network. Consequently, a guided contextual attention module is introduced to imitate the affinity-based propagation method by a fully convolutional manner. In the proposed attention module, the opacity information is transmitted between alpha features under the guidance of appearance information. The evaluation results on both Composition-1k testing dataset and alphamatting.com dataset show the superiority of our proposed method.

Acknowledgement

This paper is supported by NSFC (No.61772330, 61533012, 61876109), the advanced research project (No.61403120201), Shanghai authentication key Lab. (2017XCWZK01), Technology Committee the interdisciplinary Program of Shanghai Jiao Tong University (YG2015MS43). We also would like to thank the help and support from Versa.

References

Aksoy, Y.; Ozan Aydin, T.; and Pollefeys, M. 2017. Designing effective inter-pixel information flow for natural image matting. In *CVPR*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale gan training for high fidelity natural image synthesis. In *ICLR*.

Chen, Q.; Li, D.; and Tang, C.-K. 2013. Knn matting. *IEEE TPAMI*.

Cho, D.; Tai, Y.-W.; and Kweon, I. 2016. Natural image matting using deep convolutional neural networks. In *ECCV*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *ICCV*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111(1):98–136.

Feng, X.; Liang, X.; and Zhang, Z. 2016. A cluster sampling method for image matting via sparse coding. In *European Conference on Computer Vision*, 204–219. Springer.

Gastal, E. S., and Oliveira, M. M. 2010. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, 575–584. Wiley Online Library.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

He, K.; Rhemann, C.; Rother, C.; Tang, X.; and Sun, J. 2011. A global sampling method for alpha matting. In *CVPR 2011*, 2049–2056. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; and Li, M. 2019. Bag of tricks for image classification with convolutional neural networks. In *CVPR*.

Hui, T.; Tang, X.; and Loy, C. C. 2018. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 8981–8989.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, P., and Wu, Y. 2011. Nonlocal matting. In *CVPR 2011*, 2193–2200. IEEE.

Levin, A.; Lischinski, D.; and Weiss, Y. 2008. A closed-form solution to natural image matting. *IEEE TPAMI*.

Li, Y.; Zhang, J.; Zhao, W.; and Lu, H. 2019. Inductive guided filter: Real-time deep image matting with weakly annotated masks on mobile devices. *arXiv preprint arXiv:1905.06747*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 85–100.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Loshchilov, I., and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices matter: Learning to index for deep image matting. In *ICCV*.

Lutz, S.; Amplianitis, K.; and Smolic, A. 2018. Alphagan: Generative adversarial networks for natural image matting. In *BMVC*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Rhemann, C.; Rother, C.; Wang, J.; Gelautz, M.; Kohli, P.; and Rott, P. 2009. A perceptually motivated online benchmark for image matting. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Sun, D.; Yang, X.; Liu, M.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*.

Tang, J.; Aksoy, Y.; Öztireli, C.; Gross, M.; and Aydın, T. O. 2019. Learning-based sampling for natural image matting. In *Proc. CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, J., and Cohen, M. F. 2007. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wang, J.; Cohen, M. F.; et al. 2008. Image and video matting: a survey. *Foundations and Trends® in Computer Graphics and Vision* 3(2):97–175.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep image matting. In *CVPR*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *CVPR*.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-attention generative adversarial networks. In *ICML*.

Zheng, Y., and Kambhamettu, C. 2009. Learning based digital matting. In *ICCV*.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*, 321–328.