Deep Propagation Based Image Matting

Yu Wang[1], Yi Niu[1], Peiyong Duan[1,*], Jianwei Lin[1], Yuanjie Zheng[1,2,3,4,*]
[1] School of Information Science and Engineering, Shandong Normal University, China
[2] Key Lab of Intelligent Computing and Information Security in Universities of Shandong, China
[3] Shandong Provincial Key Lab for Distributed Computer Software Novel Technology, China
[4] Institute of Biomedical Sciences, Shandong Normal University, China
{wangyu52, linjianwei}@stu.sdnu.edu.cn, {niuyi, duanpeiyong, yjzheng}@sdnu.edu.cn

Abstract

In this paper, we propose a deep propagation based image matting framework by introducing deep learning into learning an alpha matte propagation principal. Our deep learning architecture is a concatenation of a deep feature extraction module, an affinity learning module and a matte propagation module. These three modules are all differentiable and can be optimized jointly via an end-to-end training process. Our framework results in a semantic-level pairwise similarity of pixels for propagation by learning deep image representations adapted to matte propagation. It combines the power of deep learning and matte propagation and can therefore surpass prior state-of-the-art matting techniques in terms of both accuracy and training complexity, as validated by our experimental results from 243K images created based on two benchmark matting databases.

1 Introduction

Image matting aims to extract a foreground object image F together with its alpha matte α (taking values in [0 1]) from a given image I. Techniques for achieving image matting are mostly founded on the following convex combination of F and a background image B:

$$I = \alpha F + (1 - \alpha)B. \quad (1)$$

A follow-up composition process of image matting is to blend F with a new background image B by using Eq. (1) again for creating a new image. Image matting is critical for commercial television and film production due to its power to insert new elements seamlessly into a scene or transport an actor into a totally new environment [Wang and Cohen, 2007].

Image matting is a highly ill-posed problem because it involves an estimation of seven unknowns (3 color components for each of F and B, plus the α value) from three equations for each pixel as shown in Eq. (1). Among a large variety of matting techniques (as summarized in Sec. 2), propagation-based image matting [Levin *et al.*, 2008; Chen *et al.*, 2013; Zheng and Kambhamettu, 2009] constitutes one of the most prominent matting approaches in literature. The related techniques leverage pixel similarities to propagate the alpha matte values from manually-drawn regions where the alpha values are known to unknown regions. They model a complicated image structure simply by measuring pairwise similarity between pixels, resolve matte typically with a closed-form fashion, are easy to implement and can result a smooth matte.

However, most of the existing propagation-based image matting techniques deteriorate inevitably in practice considering the fact that they are built on a low-level pairwise similarity which is typically measured by using image color or other hand-designed visual features [Levin *et al.*, 2008; Chen *et al.*, 2013]. As widely known, image matting is of a high-level vision task and therefore demands a semantic-level pairwise similarity [Liu *et al.*, 2017]. In order to deal with this limitation, the ubiquitous deep learning techniques have been recently applied to achieving a semantic-level analysis of the image for matting [Cho *et al.*, 2016; Xu *et al.*, 2017; Shen *et al.*, 2016; Aksoy *et al.*, 2017; Liu *et al.*, 2017; Bertasius *et al.*, 2016]. They behave as learning an alpha matte or a pairwise similarity in an end-to-end fashion given an image plus a trimap. However, for the former, the propagation process is not involved and learning an alpha matte directly is well-known to be hard due to the high dimensionality of the parameter space to be specified during training, especially when considering the fact that the size of alpha matte dataset is usually very limited in practice [Xu *et al.*, 2017]. For the latter, the propagation process is treated as a followed but totally-independent procedure. Therefore, the benefits of matting propagation can't be combined with the power of deep learning, which limits the performances of the related approaches.

In this paper, we introduce a novel deep propagation based image matting framework with an motivation to deal with the above challenges by propagating alpha matte values using a propagation principle learned via a deep learning architecture. Different from the existing deep learning based image matting techniques, our deep learning architecture is a concatenation of a deep feature extraction module, an affin-

*Corresponding author

ity learning module and an alpha matte propagation module. These three modules are all differentiable and can therefore be jointly trained using the stochastic gradient descent (SGD). They result in deep image representations adapted to matting propagation and semantic-level pairwise similarities. The complete training pipeline is driven by the fact that the propagated alpha matte via the learned propagation principle should be as close as possible to the ground-truth.

The proposed deep propagation based framework is distinguished from many existing matting techniques for several of its strengths. First, it can learn a semantic-level pairwise similarity for propagation via a deep learning architecture in a data-driven manner. This is in contrast to traditional hand-designed similarities which may not adequately describe pixel-pixel relationships for a high-level vision task. Second, the learned pairwise similarities are adapted to matte propagation, which is superior to techniques of learning similarity directly. Third, it combines both the power of deep learning and propagation by training the image representation and pairwise similarity jointly. Fourth, the dimensionality of its parameter space to be specified during training is significantly smaller than learning an alpha matte directly. Experimental results obtained from 243K images validate that our network obviously outperforms several representative state-of-the-art matting techniques.

2 Related Work

There exist three main approaches to digital matting: sampling based techniques [Chuang *et al.*, 2001; Feng *et al.*, 2016], propagation based frameworks [Levin *et al.*, 2008; Chen *et al.*, 2013; Zheng and Kambhamettu, 2009] and a hybrid of these two [Zheng and Kambhamettu, 2009; Wang and Cohen, 2005]. The sampling based techniques are founded on an assumption that two pixels with similar colors should be close in their alpha matte values. The propagation based framework propagates the alpha values from user-drawn foreground & background scribbles into unknown regions. It leverages a pairwise similarity of pixels to represent a complicated image structure.

Propagation based techniques constitute one of the prominent approaches in literature for not only image matting [Levin *et al.*, 2008; Chen *et al.*, 2013; Zheng and Kambhamettu, 2009] but also image segmentation & editing (very similar problems to image matting) [Chen *et al.*, 2012; Endo *et al.*, 2016]. This benefits from the availability of a clear mathematical formulation determined mainly by an inter-pixel affinity measurement characterizing simply local interactions between neighboring pixels. This formulation can be solved in closed-form and implemented easily [Levin *et al.*, 2008]. However, most of the existing propagation based approaches determine the pairwise similarity by using image color, resulting in a low-level pairwise similarity measurement. This is in essence equivalent to two basic assumptions: the linear alpha-color relation, meaning that the alpha matte value is a linear function of image color for each pixel; and the color-line model, denoting that the local image colors distribute on a line in the color space [Zheng and Kambhamettu, 2009]. These propagation-based techniques deteriorate inevitably when the above assumptions are violated in practice, typically when image color is not good enough for characterizing each pixel. The violation may result in "smearing" or "chunky" artifacts in alpha matte as descried in [Chen *et al.*, 2013; Cho *et al.*, 2016; Xu *et al.*, 2017; Aksoy *et al.*, 2017].

A variety of solutions have been proposed to resolve the challenges in propagation based techniques, including relaxing the local color distribution by searching neighbors in a nonlocal domain [Lee and Wu, 2011; He *et al.*, 2010; Chen *et al.*, 2013], defining more complex affinity measurements by using a feature vector instead of only image color [Chen *et al.*, 2013; Cho *et al.*, 2016; Xu *et al.*, 2017; Shen *et al.*, 2016; Maire *et al.*, 2016], or adapting the measurement to different image regions [Aksoy *et al.*, 2017]. However, these approaches cannot absolutely lead to obvious improvements because neither the specification of proper nonlocal neighbors nor the more effective features for better measuring affinity is a trivial [Cho *et al.*, 2016].

We noticed that the ubiquitous deep learning techniques have recently been exploited to learn a more advanced semantic-level pairwise similarity [Cho *et al.*, 2016; Xu *et al.*, 2017; Shen *et al.*, 2016; Aksoy *et al.*, 2017; Liu *et al.*, 2017; Bertasius *et al.*, 2016; Sui *et al.*, 2017]. This seems to be a promising research direction to resolving the challenges of image matting. However, they are accomplished as learning the affinity of propagation or the matte/segmentation/edit end-to-end given an image plus a trimap/label/edit. The former treats the propagation process as a followed but independent part and therefore the benefits of deep learning and propagation can't be combined. The latter excludes the propagation process and is plagued with the high dimensionality of the parameter space during training, especially when considering the fact that the size of alpha matte dataset is usually very limited in practice [Xu *et al.*, 2017].

3 Method

We construct a DeepMattePropNet to learn deep image representations with an adaption to alpha matte propagation, which results in a more effective pairwise similarity for propagation. Our motivation to design DeepMattePropNet arises from the need to reveal not only pixel-level description but also semantic-level coherence between pixels when measuring the pairwise similarity for matte propagation. As illustrated in Fig. 1, DeepMattePropNet connects a deep feature extraction module to an affinity learning module, followed by a matte propagation module. All modules are differentiable and can be trained jointly using SGD.

3.1 Deep Feature Extraction Module

Our deep feature extraction module is comprised of two branches: a semantic-level feature extraction branch and a low-level feature extraction branch. They learn deep image representations which will be used to measure the pairwise similarity for the matte propagation module. The input to these two branches is both a 4-channel matrix constructed by concatenating the original image and the corresponding manually-drawn trimap along the channel dimension.

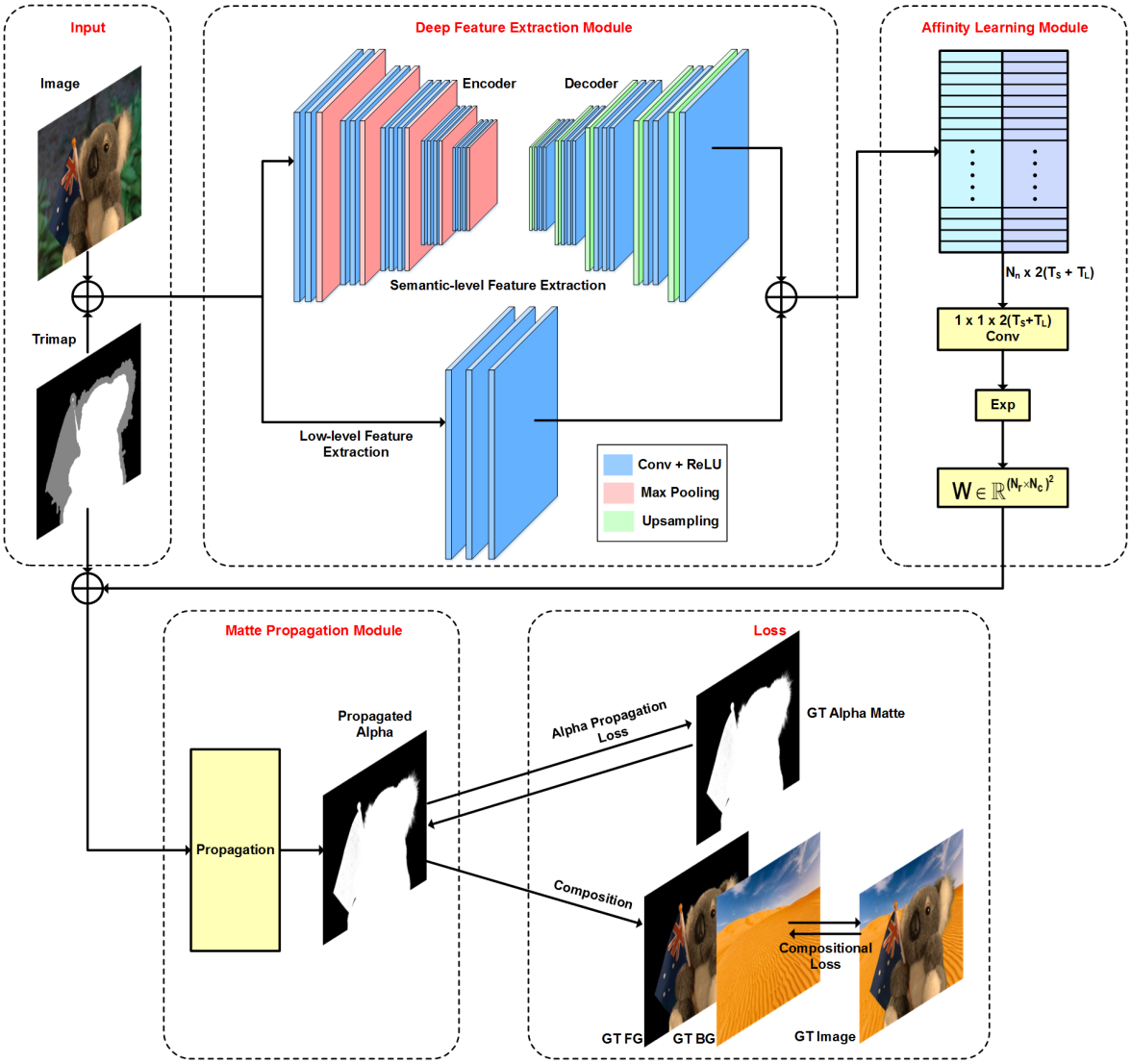


Figure 1: An illustration of our DeepMattePropNet architecture. It consists of 3 modules for deep feature extraction, affinity learning and matte propagation, respectively. They are all differentiable and the parameters of the entire architecture can be jointly optimized.

The network structure of the semantic-level feature extraction branch is identical to the SegNet [Badrinarayanan *et al.*, 2017] which consists in an encoder network and a corresponding decoder network. The encoder network transforms the input into downsampled feature maps through convolutional layers and max-pooling layers. Specifically, it consists in 13 "Conv+ReLU" layers and 5 max-pooling layers. The "Conv+ReLU" layers correspond to the first 13 convolutional layers of the VGG16 network [Simonyan and Zisserman, 2014] and each of them performs convolution with a filter bank to produce a set of feature maps, batch-normalizes the feature maps and then applies an element-wise rectified-linear nonlinearity (ReLU) $max(0, x)$. The max-pooling is carried out with a 2×2 window and stride 2. The followed decoder network semantically upsamples the features learnt by the encoder via unpooling layers and convolutional layers to get dense features for predicting pairwise similarities. It is structured in a fashion that each of its layers corresponds to one encoder lays (e.g. "Conv+ReLU" vs. "Conv+ReLU", unpooling vs. max-pooling). As in [Badrinarayanan *et al.*, 2017], the max-pooling indices are memorized and then resued in the upsampling process in the decoder network.

The branch for low-level features extraction is a network composed of 3 convolutional layers (with a 3×3 kernel), each of which is followed by a nonlinear"ReLU" layer. As shown in [Xu *et al.*, 2017], low-level features can result in more matte details.

The semantic-level feature extraction branch outputs $N_r \times N_c \times T_s$ features, where N_r, N_c and T_s represent the number of rows of the original image, columns of the original image and features output by this branch, respectively. The low-level feature extraction branch produces $N_r \times N_c \times T_l$ features, where T_l denotes the number of output features.

3.2 Affinity Learning Module

The affinity learning module learns pairwise affinity of pixels for propagation and is connected with the semantic-level feature extraction branch and the low-level feature extraction branch of the deep feature extraction module. The input to the affinity learning module is a $N_n \times 2(T_s + T_l)$ matrix for which each row stores the learned $2(T_s + T_l)$ deep features for each pair of neighboring pixels, where N_n denotes the total number of neighboring-pixel pairs. The neighborhood can be defined as 4-connection as in our paper.

The affinity learning module consists of a $1\times1\times2(T_s+T_l)$ convolutional layer and an exponential layer. It predicts the affinity value for each pair of neighboring pixels. All affinity values output from this module form a $(N_rN_c) \times (N_r\ N_c)$ symmetric and sparse affinity matrix $\mathbf{W}$ which will be then fed into the matte propagation module. Note that these twos layers are both differentiable.

3.3 Matte Propagation Module

The matte propagation module propagates alpha matte specified by the input trimap based on the affinity matrix $\mathbf{W}$. It generates a refined matte which will be then attached to the loss module. We provide below the mathematics related to matte propagation and prove that this module is differentiable.

From the affinity matrix $\mathbf{W}$, we define a diagonal matrix $\mathbf{D}$ for which each diagonal element equals to the sum of the corresponding row of $\mathbf{W}$. A typical alpha matte propagation module [Levin *et al.*, 2008] can be expressed as

$$\boldsymbol{\alpha} = \arg \min_{\boldsymbol{\alpha}\in\mathbb{R}^{(N_rN_c)}} \boldsymbol{\alpha}^T\mathcal{L}\boldsymbol{\alpha} + (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{C}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \quad (2)$$

where $\boldsymbol{\alpha}$ denotes a vector (in length N_rN_c) of alpha matte values for all pixels, T means transpose, $\boldsymbol{\alpha}^*$ represents a vector (in length N_rN_c) containing all alpha values known from the trimap, $\mathcal{L} = \mathbf{D} - \mathbf{W}$ is a $(N_rN_c) \times (N_rN_c)$ Laplacian matrix, $\mathbf{C}$ stands for a $(N_rN_c) \times (N_rN_c)$ diagonal matrix for which a diagonal element takes zero if the corresponding pixel belongs to unknown regions and a constant value c otherwise. The value of c adjusts the importance of the labeled pixels when propagating the alpha matte and is set as 0.8 in our paper. The solution to Eq. (2) is written as

$$\begin{aligned}\boldsymbol{\alpha} =&(\mathcal{L}+\mathbf{C})^{-1}\mathbf{C}\boldsymbol{\alpha}^*\\ =&(\mathbf{D}-\mathbf{W}+\mathbf{C})^{-1}\mathbf{C}\boldsymbol{\alpha}^*.\end{aligned} \quad (3)$$

Taking the derivative of $\boldsymbol{\alpha}$ relative to an element of $\mathbf{W}_{ij}$ of $\mathbf{W}$, where i and j index the row and column of $\mathbf{W}$, respectively, we have

$$\begin{aligned}\frac{\partial\boldsymbol{\alpha}}{\partial\mathbf{W}_{ji}} =&\frac{\partial(\mathcal{L}+\mathbf{C})^{-1}}{\partial\mathbf{W}_{ji}}\mathbf{C}\boldsymbol{\alpha}^*\\ =&(\mathcal{L}+\mathbf{C})^{-1}\mathbf{J}_{ij}(\mathcal{L}+\mathbf{C})^{-1}\mathbf{C}\boldsymbol{\alpha}^*\end{aligned} \quad (4)$$

where $\mathbf{J}_{ij}$ is a $(N_rN_c)\times(N_rN_c)$ matrix for which the element corresponding to the ith row and jth column is 1 and all other elements are zero.

In Eq. (4), the resulted matrix from $(\mathcal{L} + \mathbf{C})$ is huge and its inverse is hard to compute. As in [Bertasius *et al.*, 2016], we shrink this matrix using a simple but efficient technique in [Arbeláez *et al.*, 2014]. The work in [Bertasius *et al.*, 2016] requires the inverse of a huge matrix to be computed for each random-walk step and demands a couple of random-walk steps for each image during training. In contrast, our scheme requests only one time of computation. Eq. (4) proves the differentiability of the propagation module.

The dimensionality of the parameter space of DeepMattePropNet is significantly smaller than a network predicting per-pixel alpha matte value (e.g. the works in [Xu *et al.*, 2017; Cho *et al.*, 2016]). It is because the convolutional layer in this affinity learning module of DeepMattePropNet share the same weights across all pixel pairs, as shown in Fig. 1. For example, the work in [Xu *et al.*, 2017] requires a parameter space in a dimensionality equaling to the number of image pixels, e.g. N_rN_c. In contrast, our network only needs to learn the $2(T_s + T_l)$ parameters of the convolutional kernels plus the ones for the exponential layer in the affinity learning module, which are much fewer than the number of pixels.

3.4 Losses

The average over the alpha prediction loss and composition loss in [Xu *et al.*, 2017] is treated as the overall loss for training DeepMattePropNet. These two losses measure the Euclidean distance between the ground-truth and the predicted one for the alpha matte and composited color image, respectively, as shown by the following equation's computation for one pixel:

$$L = \sqrt{(\hat{\alpha}-\alpha^*)^2+\epsilon} + \sqrt{\|\hat{\mathbf{c}}-\mathbf{c}^*\|^2+\epsilon} \quad (5)$$

where $\hat{\alpha}$ and α^* denote the estimated and ground-truth alpha matte values, respectively, ϵ is very small number (e.g. 10^{-12}), and $\hat{\mathbf{c}}$ and $\mathbf{c}^*$ represents the composited and ground-truth image colors, respectively. Note that the loss computation can involve only the pixels in the unknown regions in order to reduce the computational complexity.

3.5 Implementation Details

We train the deep feature extraction module, affinity learning module and matte propagation module in the DeepMattePropNet jointly and hence they are integral parts of the network during testing. The training is carried out in an end-to-end (original image plus trimap to alpha matte) fashion.

We implement the DeepMattePropNet using Caffe and conduct training and testing on a NVIDIA Titan X graphics card. The training is carried out with a fixed learning rate of 0.1 and momentum of 0.9. The overall training phase requires about 20 epochs.

4 Experimental Results

4.1 Dataset

We evaluate the performance of the proposed DeepMattePropNet on two matting tasks, the benchmark alphamatting.com dataset [Rhemann *et al.*, 2009] and our own dataset. As the benchmark challenge for image matting, the alphamatting.com dataset makes both the original images and ground truth mattes available online (at www.alphamatting.com). It includes 27 images for training and 8 images for testing, and for each of which, a low-resolution trimap and a high one are

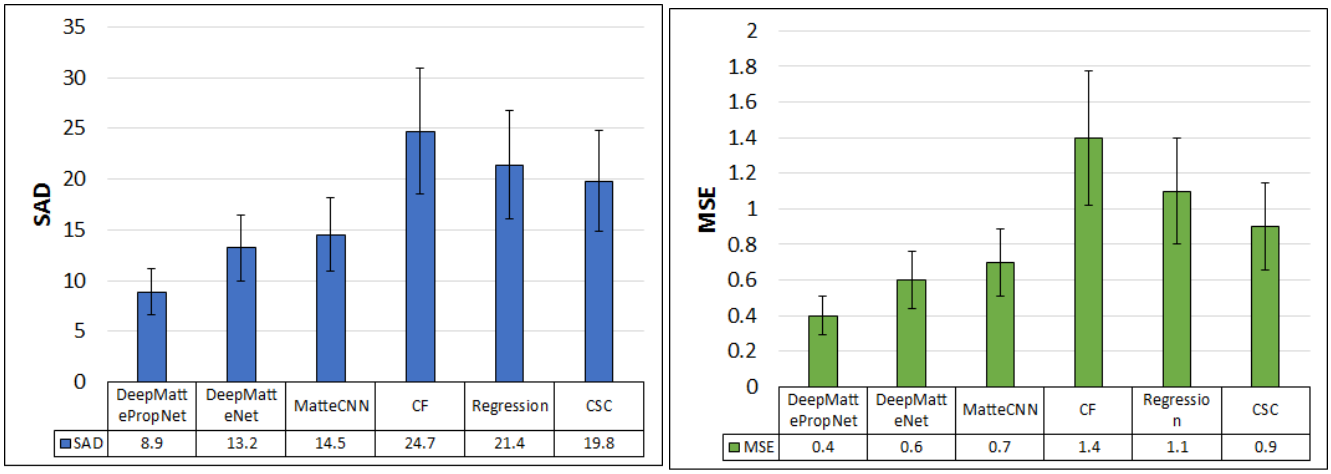
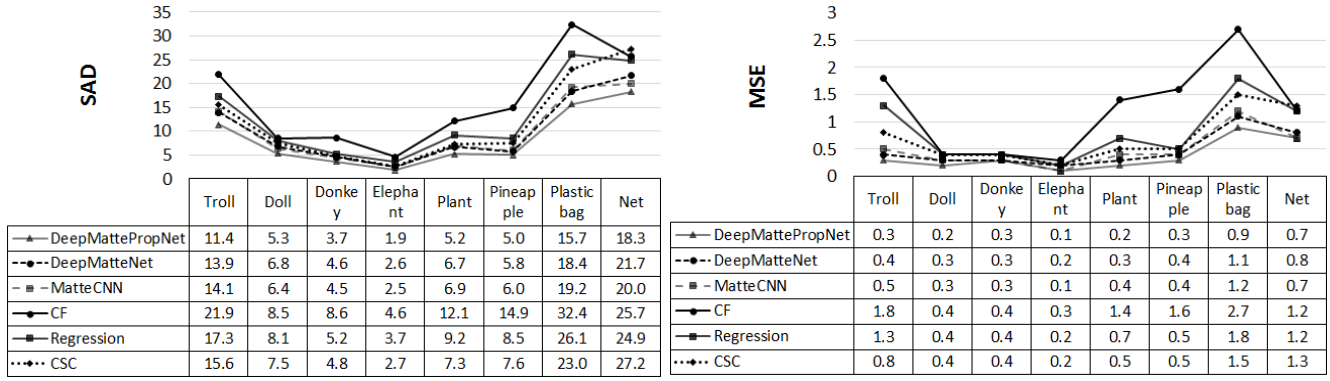


Figure 2: Quantitative comparisons of alpha matte estimation errors on all the testing images in terms of SAD (sum of absolute differences) and MSE (mean square error) between our DeepMattePropNet and 5 representative state-of-the-art techniques. Bars represent the standard deviation of the mean.



SAD	Troll	Doll	Donkey	Elephant	Plant	Pineapple	Plastic bag	Net
DeepMattePropNet	11.4	5.3	3.7	1.9	5.2	5.0	15.7	18.3
DeepMatteNet	13.9	6.8	4.6	2.6	6.7	5.8	18.4	21.7
MatteCNN	14.1	6.4	4.5	2.5	6.9	6.0	19.2	20.0
CF	21.9	8.5	8.6	4.6	12.1	14.9	32.4	25.7
Regression	17.3	8.1	5.2	3.7	9.2	8.5	26.1	24.9
CSC	15.6	7.5	4.8	2.7	7.3	7.6	23.0	27.2

MSE	Troll	Doll	Donkey	Elephant	Plant	Pineapple	Plastic bag	Net
DeepMattePropNet	0.3	0.2	0.3	0.1	0.2	0.3	0.9	0.7
DeepMatteNet	0.4	0.3	0.3	0.2	0.3	0.4	1.1	0.8
MatteCNN	0.5	0.3	0.3	0.1	0.4	0.4	1.2	0.7
CF	1.8	0.4	0.4	0.3	1.4	1.6	2.7	1.2
Regression	1.3	0.4	0.4	0.2	0.7	0.5	1.8	1.2
CSC	0.8	0.4	0.4	0.2	0.5	0.5	1.5	1.3

Figure 3: Quantitative comparisons of alpha matte estimation errors on the 8 alphamatting.com testing images in terms of SAD and MSE between DeepMattePropNet and 5 representative state-of-the-art techniques.

provided. We only use the more challenging low-resolution trimaps in our experiments.

Our own dataset consists of 46 images captured by filming 46 target objects in front of a computer monitor displaying an out-door scene. To obtain their ground-truth alpha matte, we first film these objects in front of five additional constant-color backgrounds and then derive alpha matte by solving an overdetermined linear system of the composition equations (as in Eq. (1)) using singular value decomposition (SVD) [Chuang *et al.*, 2001]. For each of the 46 images, we draw a low-resolution trimap manually. Therefore, we have totally 81 original images for which a ground-truth alpha matte and a low-resolution trimap are available. All images and trimaps are resized to 600×800.

We also augment these original images by composing new images using the corresponding ground-truth alpha matte. The new background images are composed of 500 indoor image selected randomly from the indoor scene recognition database [Quattoni and Torralba, 2009] plus 500 outdoor images chosen randomly from the Places database [Zhou *et al.*, 2014]. All indoor and outdoor images are resized to 600×800. In addition, we also exploit 3 different rotations of the foreground and alpha matte when composition. We finally obtain a total of 243K images. We treat the 50 images including the 8 alphamatting.com testing images plus 42 images selected randomly from other 73 original images and all their composited images as the training set and all the left images as the testing set.

Training the DeepMattePropNet takes around 3$\sim$4 days. For the testing phase, the running time on a 600×800 image is about 1.2 seconds (about 7.4 seconds if conducted on a CPU using the Intel MKL-optimized Caffe).

4.2 Evaluation

We compare the performances of our DeepMattePropNet with several state-of-the-art techniques including the deep image matting (DeepMatteNet) [Xu *et al.*, 2017], the CNN based method (MatteCNN) in [Cho *et al.*, 2016], the closed-form (CF) matting [Levin *et al.*, 2008], the CSC method proposed in [Feng *et al.*, 2016] and the regression algorithm in [Xiang *et al.*, 2010]. All networks are carried out on the same training and testing images. We provide both quantitative as-

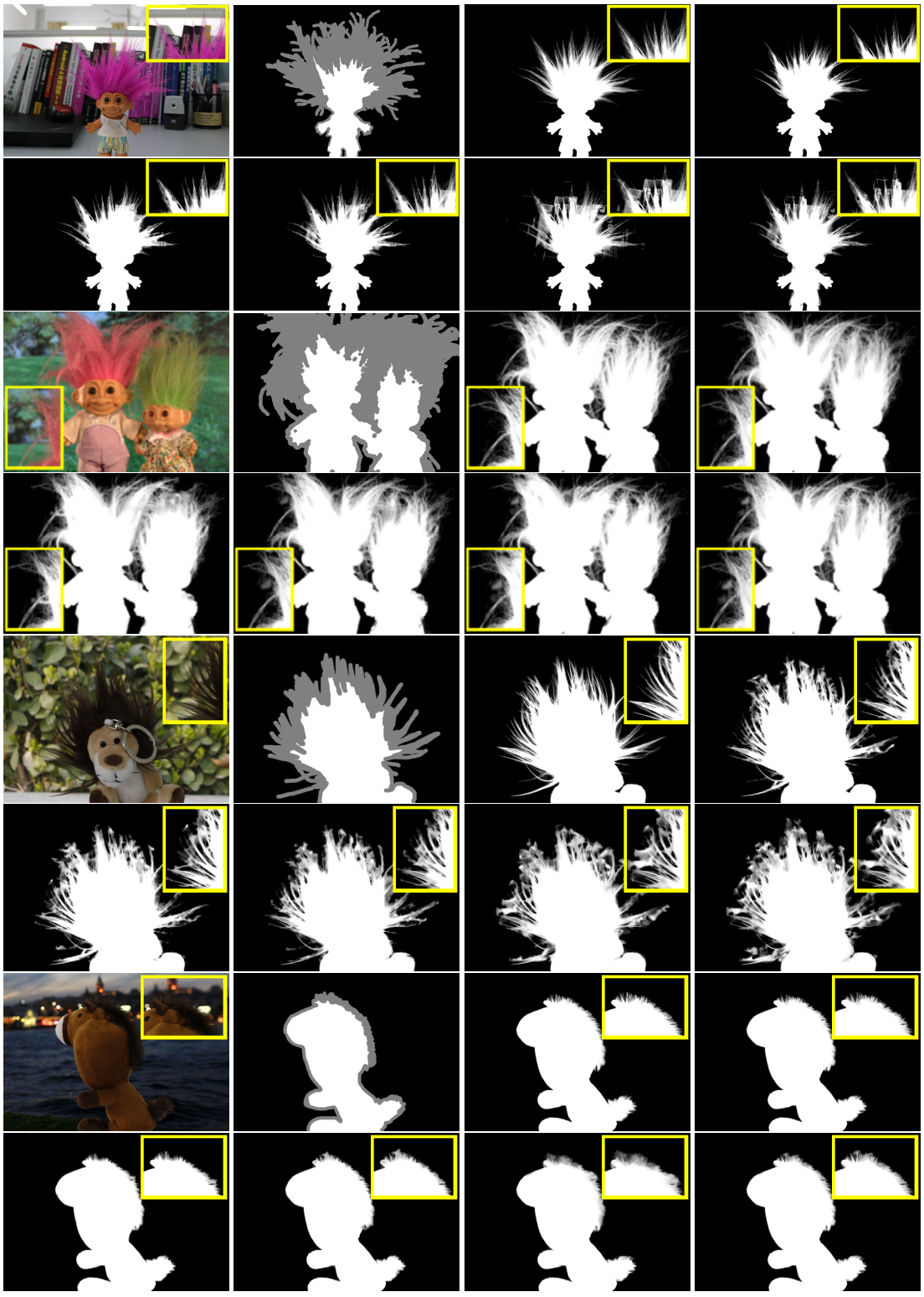


Figure 4: Visual comparison of alpha mattes produced by DeepMattePropNet and 4 state-of-the-art techniques from 4 exampling images. For each exampling image, top to bottom and left to right: original image, trimap, ground-truth matte, mattes from DeepMattePropNet, DeepMatteNet, MatteCNN, CF and CSC, respectively. Yellow rectangles contain a larger view of a local region.

sessments and visual evaluations.

We show SAD (sum of absolute differences) and MSE (mean square error) values over all the testing images in our datasets in Fig. 2 and the ones for each of the alphamatting.com testing images in Fig. 3. From these error statistics, we have at least two findings. First, deep learning based alpha matting techniques perform overwhelmingly better than traditional techniques. This may benefit from the superiority of deep image representations learned from a huge image dataset in contrast to the hand-designed features. Second, our deep matte propagation based method outperforms other two deep learning based matting techniques. This may arise from the fact that our method learns deep image representations for a better matte propagation while others focus on learning the alpha matte directly.

Our visual comparisons show that the DeepMattePropNet is more stable and visually pleasing for various structures of foreground object, including solid boundary, transparent areas, overlapped color distribution between foreground and background and long slim structures. As shown by the results from the upper exampling image in Fig. 4, our network outperforms other techniques obviously especially at the region where the "purple hair" locates in front of the "purple book".

The power of our DeepMattePropNet comes from several of its advantages. First, it propagates matte based on not only low-level but also semantic-level image features. The former enables extraction of matte details while the latter may help to recognize objects and resolve problems such as the one caused by overlapped color distributions between foreground and background. Second, our network predicts similarities instead of alpha matte. This reduces significantly the dimensionality of the parameter space. Third, it combines the strengths of deep learning and propagation.

5 Conclusion

Propagation based image matting techniques treat each pixel of an image as a graph's node and connect each pair of neighboring pixels by the graph's edge. The edge weight measures the affinity between pixels and reflects the pairwise similarity for the image matting task. As pointed out in [Aksoy *et al.*, 2017; Liu *et al.*, 2017], as image matting is a high-level vision task, the affinity used in propagation based matting techniques should reveal semantic-level pairwise similarity. This may be the reason why many of previous image matting techniques can fail in various practical cases because they are mostly based on low-level pairwise similarity.

In this paper, we show that a semantic-level pairwise similarity for propagation based image matting can be learned in a purely data-driven manner via a deep learning mechanism. We carry out our learning process by inserting a similarity learning module and a matte propagation module into the sequence of operations between feature learning and image matte. We also prove that these two modules are both differentiable and therefore the complete deep learning architecture can be optimized jointly using backpropagation and stochastic gradient descent (SGD). Our framework is more efficient than state-of-the-art image matting techniques because it combines the power of deep image representations from deep learning techniques and the propagation based matting techniques. In addition, unlike most deep learning based matting techniques which predict alpha matte directly, the proposed network learns similarity for propagation, which reduces the dimensionality of parameter space significantly. Experimental results from the public alphamatting.com database and our own database show the superiority of the proposed framework against several representative state-of-the-art matting techniques. Especially, we show that the proposed framework can help to deal with difficulties in matting when the foreground colors overlap with background colors.

Our future work would include an extension of our DeepMattePropNet to other propagation based image editing tasks, e.g. image colorization and image segmentation. In addition, we would extend our own matting evaluation database in order to include more images and replace the exponential layer in our affinity learning module for eliminating the potential gradient saturation problem.

Acknowledgments

This work was made possible through support from Natural Science Foundation of China (NSFC) (61572300;61773246) and Taishan Scholar Program of Shandong Province in China (TSHW201502038).

References

[Aksoy *et al.*, 2017] Yagız Aksoy, Tunç Ozan Aydın, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Arbeláez *et al.*, 2014] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[Bertasius *et al.*, 2016] Gedas Bertasius, Lorenzo Torresani, Stella X Yu, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. *arXiv preprint arXiv:1605.07681*, 2016.

[Chen *et al.*, 2012] Xiaowu Chen, Dongqing Zou, Qinping Zhao, and Ping Tan. Manifold preserving edit propagation. *ACM Transactions on Graphics (TOG)*, 31(6):132, 2012.

[Chen *et al.*, 2013] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[Cho *et al.*, 2016] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional

neural networks. In *European Conference on Computer Vision*, pages 626–643. Springer, 2016.

[Chuang *et al.*, 2001] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2001.

[Endo *et al.*, 2016] Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. Deepprop: extracting deep features from a single image for edit propagation. In *Computer Graphics Forum*, volume 35, pages 189–201. Wiley Online Library, 2016.

[Feng *et al.*, 2016] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *European Conference on Computer Vision*, pages 204–219. Springer, 2016.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2165–2172. IEEE, 2010.

[Lee and Wu, 2011] Philip Lee and Ying Wu. Nonlocal matting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2193–2200. IEEE, 2011.

[Levin *et al.*, 2008] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008.

[Liu *et al.*, 2017] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, pages 1519–1529, 2017.

[Maire *et al.*, 2016] Michael Maire, Takuya Narihira, and Stella X Yu. Affinity cnn: Learning pixel-centric pairwise relations for figure/ground embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 174–182, 2016.

[Quattoni and Torralba, 2009] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[Rhemann *et al.*, 2009] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1826–1833. IEEE, 2009.

[Shen *et al.*, 2016] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, pages 92–107. Springer, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sui *et al.*, 2017] Xiaodan Sui, Yuanjie Zheng, Benzheng Wei, Hongsheng Bi, Jianfeng Wu, Xuemei Pan, Yilong Yin, and Shaoting Zhang. Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks. *Neurocomputing*, 237:332–341, 2017.

[Wang and Cohen, 2005] Jue Wang and Michael F Cohen. An iterative optimization approach for unified image segmentation and matting. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 936–943. IEEE, 2005.

[Wang and Cohen, 2007] Jue Wang and Michael F Cohen. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[Xiang *et al.*, 2010] Shiming Xiang, Feiping Nie, and Changshui Zhang. Semi-supervised classification via local spline regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2039–2053, 2010.

[Xu *et al.*, 2017] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. *arXiv preprint arXiv:1703.03872*, 2017.

[Zheng and Kambhamettu, 2009] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *2009 IEEE 12th International Conference on Computer Vision*, pages 889–896. IEEE, 2009.

[Zhou *et al.*, 2014] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.