Semantic Image Matting

Yanan Sun
HKUST
now.syn@gmail.com

Chi-Keung Tang
HKUST
cktang@cs.ust.hk

Yu-Wing Tai
Kuaishou Technology
yuwing@gmail.com

Abstract

Natural image matting separates the foreground from background in fractional occupancy which can be caused by highly transparent objects, complex foreground (e.g., net or tree), and/or objects containing very fine details (e.g., hairs). Although conventional matting formulation can be applied to all of the above cases, no previous work has attempted to reason the underlying causes of matting due to various foreground semantics.

We show how to obtain better alpha mattes by incorporating into our framework semantic classification of matting regions. Specifically, we consider and learn 20 classes of matting patterns, and propose to extend the conventional trimap to semantic trimap. The proposed semantic trimap can be obtained automatically through patch structure analysis within trimap regions. Meanwhile, we learn a multi-class discriminator to regularize the alpha prediction at semantic level, and content-sensitive weights to balance different regularization losses. Experiments on multiple benchmarks show that our method outperforms other methods and has achieved the most competitive state-of-the-art performance. Finally, we contribute a large-scale Semantic Image Matting Dataset with careful consideration of data balancing across different semantic classes. Code and dataset are available at https://github.com/nowsyn/SIM.

1. Introduction

The matting equation models an image I as a linear combination of a foreground image F and a background image B modulated by an alpha matte α:

$$I = \alpha F + (1 - \alpha)B. \tag{1}$$

The natural image matting problem is to extract the alpha matte α from a given image, which has a wide range of applications in image/video editing. Typical foreground objects can belong to a great variety of categories, such as
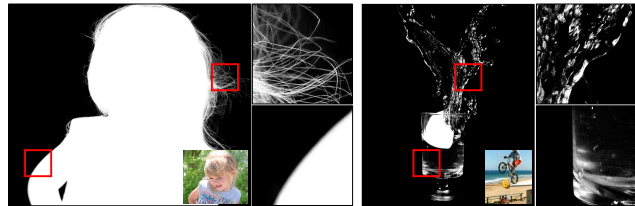
Figure 1. Alpha mattes can have different local shapes and patterns depending on the underlying foreground. Left shows a human matte with both soft hair and sharp body boundaries. Right shows a complex matte where the foreground exhibits different degrees of transparency. Deep reasoning of object semantics in natural image matting can advance the state-of-the-art results.

humans, furry animals, glass objects with transparent/semi-transparent regions, or objects with complex shapes such as net or tree, thus making this research problem still challenging, see Figure 1.

Natural image matting is an ill-posed problem, so most existing methods take both image and user-supplied clues (scribbles, trimaps, etc.) as input to solve the problem. Class-agnostic trimap comprising of foreground, background, and unknown regions is the most common choice among traditional and deep learning-based methods.

Traditional methods [12, 14, 16, 18, 36, 3, 5, 10, 17, 24, 25] mainly rely on low-level features related to image colors or structures. As traditional methods do not take into consideration any semantic information of foreground objects or background scenes, they fail easily in images where foreground pixels mingle deeply with background pixels. This issue has been addressed to a considerable extent with the introduction of deep neural networks. In [2] images are segmented into soft transitional regions based on semantics. However, this method can easily fail in images with entangled colors or large transitional regions. To extract alpha mattes for general objects, deep learning-based methods [11, 43, 31, 6, 20, 26] have contributed various convolution network designs to greatly improve performance benefiting from high-level semantic representations. Recent works such as human matting and transparent object matting [38, 49, 29, 7] have focused on solving specific instances of the matting problem, leveraging prior domain knowledge, and contributing to excellent performance for

these class-specific matting tasks. However, these deep learning-based methods still utilize semantic information only at a data level and do not adequately consider the underlying cause of matting due to different object semantics. Given the unknown region surrounding a foreground object, different types of boundaries or patterns may exist. For example, a portrait usually has both fuzzy hair and sharp boundaries.

In this paper, we propose to incorporate semantic classification of matting regions into our matting framework for extracting better alpha matte. Specifically, we first cluster 20 different matting classes based on the regional matting patterns. Our matting classes cover most typical matting scenarios for various foreground objects. Then, we extend the conventional class-agnostic trimap to semantic trimap, which consists of a 2D confidence map for each matting class in the unknown region. The automatically generated semantic trimap with RGB image is then fed into our framework as input. Meanwhile, we learn a multi-class discriminator for supervision, providing regularization from a semantic level for alpha predictions. In particular, to improve prediction with semantics incorporation, we introduce gradient constraints with content-sensitive weights to balance different regularization losses.

In summary, our main contributions are:

1. We introduce semantics into the matting task and demonstrate how semantic information can be used to achieve the most competitive performance. To our knowledge, this is the *first* paper to consider semantic classification of matting patterns in a natural image matting framework.

2. Our main technical contributions include: the introduction of semantic trimap, the proposal of learnable content-sensitive weights, and the usage of multi-class discriminator to regularize the matting results.

3. We contribute the first large-scale class-balanced Semantic Image Matting Dataset covering a wide range of matting patterns to benefit future matting research. Our new dataset provides new insight and in-depth analysis to the performance of different matting algorithms on different matting classes.

2. Related Work

We first briefly review representative works on general image matting. Then, we will discuss methods that are tailored for humans and other class-specific objects.

Natural Image Matting. Traditional methods on natural image matting are all based on the linear combination equation Eq. 1 except for [39, 23]. Since this is an ill-posed problem, traditional methods often make use of different priors. They can be further divided into sampling-based methods [12, 14, 16, 18, 36] and propagation-based

methods [3, 2, 5, 10, 17, 24, 25] or a combination of both. Sampling-based methods assume color similarities between unknown regions and definitely foreground/background regions to estimate foreground and background colors followed by alpha matte estimation. Propagation-based methods utilize the color line model proposed by [24] to estimate alpha by propagating its values from known regions to unknown regions. A comprehensive review on traditional matting methods can be found in [42, 11].

For deep-learning based methods, Cho *et al*. [11] proposed to apply deep neural networks to fuse the results produced by closed-form matting [24] and KNN matting [10]. Xu *et al*. [43] introduced the Composite-1K dataset and a two-stage encoder-decoder network for general object matting. Lutz *et al*. [31] proposed a generative adversarial network in the encoder-decoder framework to improve matting performance. Cai *et al*. [6] introduced trimap adaptation and applied multi-task learning to matting. Hou *et al*. [20] simultaneously estimated foreground and alpha matte with local matting encoder and global context encoder. Li *et al*. [26] designed a guided contextual attention module to propagate high-level opacity information globally based on the learned low-level affinity. Recently, a number of works [46, 37, 33] have been proposed focusing on relaxing trimap input. Lastly, Forte and Pitié [15] introduced the FBA-net which simultaneously estimates foreground, background and alpha. See supplementary material for comparison with [15][1].

Human Matting. Human matting was first introduced by Shen *et al*. [38] to extract alpha mattes of portrait images. Different from general object matting, human matting only considers hairs and sharp boundaries of human body. The authors formulated the solution as a two-step approach: human detection followed by human matting [49]. Since foreground objects are already known to be human, the trimap can be eliminated [9, 29]. Although these approaches give excellent performance for human subjects, they fall short of matting general objects. In addition, since they were trained using portrait images, the performance would significantly drop when the human detector fails.

Class-Specific Matting. For works targeting at other specific classes of matting, Lin *et al*. [27] introduced motion regularization for matting motion blurred moving objects. Köhler *et al*. [22] proposed to separate motion blurred foreground through explicit modeling of object motion. Amin *et al*. [4] applied image matting to segment out-of-focus regions. Yeung *et al*. [44] proposed attenuation-refraction matte to mask out and re-composite transparent object with refraction consideration. Chen *et al*. [7] proposed a multiscale encoder-decoder network to estimate the refractive flow for transparent object matting.

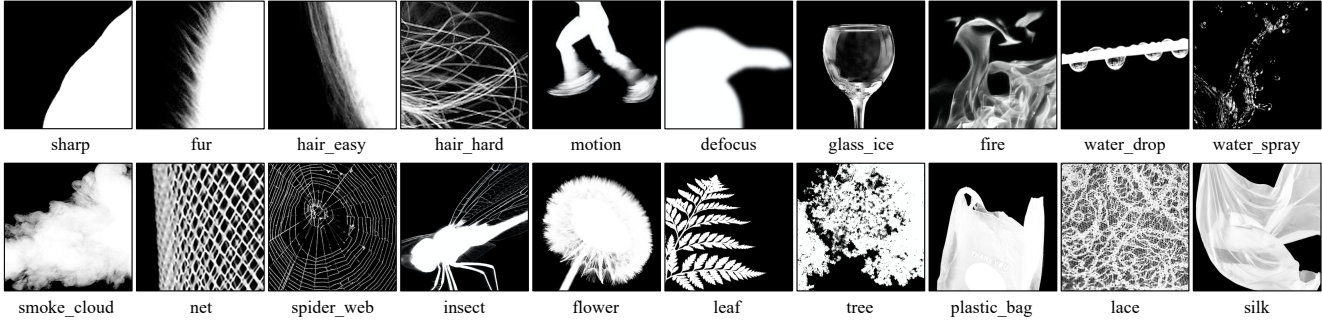[1][15] has not been published in recognized, peer-reviewed venues.

Figure 2. The 20 matting classes with high diversity in appearance across different classes.
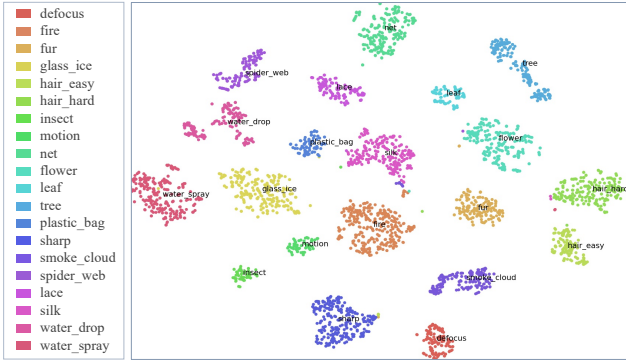

Figure 3. t-SNE visualizations of the class-specific features extracted from our discriminator (for its design see Method section).

In summary, traditional matting methods rely on low-level image cues for solving the matting equation, while more recent methods do not adequately consider or provide a principled approach to semantic image matting, or are limited to one or few classes such as humans.

3. Dataset

Although there exist several large-scale matting datasets for training deep neural networks, they exhibit severe biases with respect to our *matting classes*.

Matting Classes. We study all existing publicly available matting datasets. Alpha mattes can exhibit diverse patterns. We propose 20 matting classes with distinct pattern as shown in Figure 2. These 20 classes cover almost all of the matting scenarios in existing matting datasets. While there exist numerous matting cases in the real world, most of them can be categorized into these 20 classes while the rest are quite rare. Figure 3 shows the t-SNE visualizations of the 20 classes.

Based on these classes, we keep part of foreground objects in Adobe Image Matting Dataset [43]. Additionally, we collect sufficient clean images covering objects with corresponding patterns, and carefully extract their alpha matte with Photoshop. Finally, our Semantic Image Matting Dataset consists of 20 classes with 726 training foregrounds and 89 testing foregrounds.

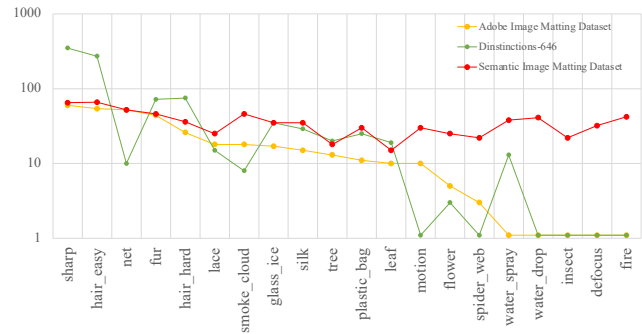To generate the composited training set, we combine

Figure 4. Class distribution of three matting datasets: Adobe Image Matting Dataset [43], Distinctions-646 [33] and our Semantic Image Matting Dataset.
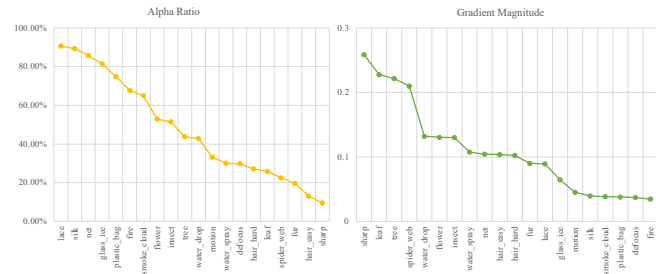

Figure 5. Left: class-specific alpha ratio distribution. Alpha ratio is the quotient between the number of alpha pixels and the total number of all pixels within a region. Right: Alpha gradient magnitude distribution. Y-axis is the mean gradient magnitude of alpha pixels within a region.

training foregrounds with randomly selected background images from COCO [28] dataset in an online manner during training. For the testing set, we follow [43] and composite each testing foreground with 10 background images from PASCAL VOC [13] dataset.

Comparison with Other Matting Datasets. Although large-scale matting datasets, e.g., Adobe Image Matting Dataset [43] and Distinctions-646 [33], have been proposed as benchmarks in matting task, they are heavily biased toward common objects such as humans or animals. Figure 4 compares the statistics of these two representative datasets and our dataset, showing that our Semantic Image Matting Dataset is class-balanced across the 20 matting classes.
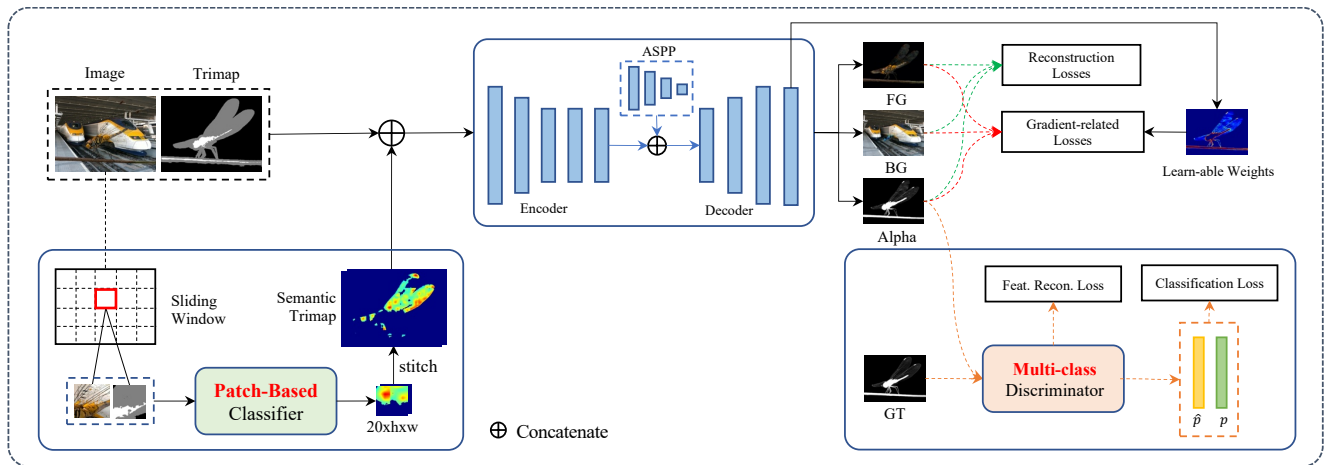
Figure 6. Overall framework. Our method takes as input an RGB image and its semantic trimap automatically generated from a patch-based classifier, which simultaneously predicts the alpha, foreground, background as well as learnable regularization weights. During training, a multi-class discriminator is used to provide supervision including classification loss and feature reconstruction loss from a semantic level.

Analysis. In Figure 5, we compute the average gradient magnitude of alpha mattes as well as foreground alpha ratios within the unknown region in all classes and plot their distributions. From the statistics, we can observe that while some classes have smooth texture with small gradients, such as fire, smoke, and silk, other classes exhibit many fine features and sharp structures, such as net and spider web. This motivates us to exploit more useful information from class-related gradient constraints in training our model. More detailed statistics are included in the supplementary materials.

Different matting classes exhibit quite distinctive alpha patterns and distributions, indicating that taking into consideration such semantics caused by different objects should boost matting performance.

4. Method

Figure 6 shows our whole framework which is an encoder-decoder structure that takes an RGB image as well as its semantic trimap as input, and outputs the alpha predictions. Notably, this network is supervised by our multi-class discriminator as well as reconstruction and gradient-related losses. The network architecture of this multi-class discriminator is the same as the patch-based classifier which generates the semantic trimap, comprising of standard CNN, max-pooling layers, and ResBlocks.

4.1. Framework

Patch-Based Classifier. Semantic trimap is the concatenation of a conventional trimap and n-channel score maps where n is the number of matting classes. The conventional trimap defines definite foreground, definite background and unknown region while the score maps indicate the confidence each unknown pixel belongs to a certain matting class. The score maps are automatically generated through semi-supervised patch-based structure analysis within unknown regions.

Our patch-based classifier is trained on the composited dataset where the images are partitioned with patch labels indicating their respective matting classes. An alpha image usually contains more than one matting pattern, so it is partitioned into multiple regions each of which only belongs to a certain class. When training the classifier, we randomly crop a square patch from unknown regions of a composited image. To make the classifier robust to scale variation, the crop size is random, uniformly chosen within a range of $[160, 640]$ and then the cropped patch is resized to 320. Then, the classifier takes as input this given patch with the corresponding conventional trimap, and predicts the matting class for the given patch. After the classifier is well-trained, we compute the n-channel score maps for a patch by multiplying the last convolutional feature map with the fully connected weights, which is also known as class activation map [48].

During inference, we obtain the semantic trimap for the entire input image through multi-scale patch analysis. In detail, we partition the input image into multi-scale overlapped patches and stitch the score maps of these patches together. Then we treat the stitched scores maps with the conventional trimap as our semantic trimap.

When we train our matting network, the semantic trimap will be concatenated with the RGB image and fed into the encoder, as shown in Figure 6. Compared to conventional trimap, semantic trimap provides prior knowledge for the network to reduce search space for each class, focusing the model to predict more reliable alpha matte compatible to patterns caused by pertinent matting object semantics.

Encoder-Decoder Structure. Our framework for inference stage consists of U-Net [35] like structure. The encoder is adapted from ResNet-50 [19] with a downsampling stride
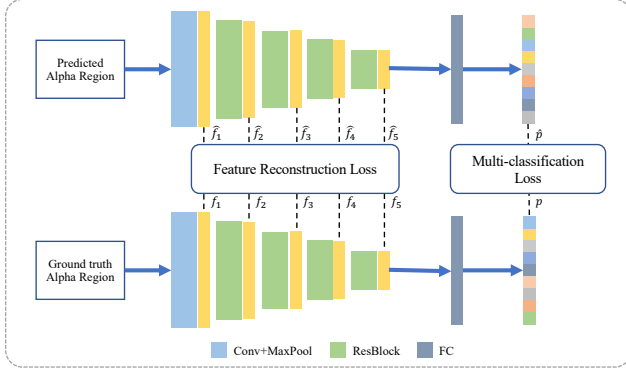
Figure 7. Multi-class discriminator. After feeding the predicted alpha region and groundtruth region to the well-trained discriminator, we compute the classification loss and feature reconstruction loss in order to preserve the distribution of predictions.

of 8. Dilation convolution layers are used to enlarge receptive fields. Before sending encoded features to the decoder, an atrous spatial pyramid pooling (ASPP) module [8] is applied to aggregate features of different receptive fields in order to enhance the feature representation capability. Afterward, three up-conv layers are utilized to recover the spatial information by integrating high-level features as well as high-resolution features from shallow layers. We simultaneously predict F, B, α through 3 prediction heads comprised of two 3×3 convolutional layers.

Multi-Class Discriminator. Previous works [31, 37, 33] also adopt a discriminator to provide additional supervision for obtaining better alpha predictions. However, all of them have not taken inter-class difference into consideration. In contrast, we apply a multi-class discriminator to make the model aware of specific structures and preserve the statistics of relevant alpha patterns. The regularization imposed by the multi-class discriminator enforces the model to better learn the characteristics of each class and consequently give more reasonable predictions.

Overall, the multi-class discriminator shares the same architecture with our semantic trimap classifier, except taking alpha matte as input during training and inference. Figure 7 shows its detailed structure with data flows.

Before training the matting network, we first train this classifier on the groundtruth alpha mattes in Semantic Image Matting Dataset. The training process is similar to that of our semantic trimap classifier. Specifically, in each iteration, we randomly crop a square patch from the unknown class-specific regions and feed it to the classifier. After the classifier has been well trained, it will be deployed as a discriminator in our matting framework to generate classification loss and feature reconstruction loss.

During training, after obtaining the predicted alpha patch from the auto encoder-decoder matting network, the predicted alpha patch and its groundtruth patch are sent to the well-trained classifier so that we can extract for the two

patches the respective multi-classification probabilities $\hat{p}$ and p, as well as the deep encoded multi-level features $\hat{f}_k$ and f_k where $k \in \{1, 2, 3, 4, 5\}$.

To compute the classification loss, we treat $\hat{p}$ as logit and p as label, and apply a cross-entropy loss on them. Note that even though the prediction for the groundtruth patch p is different from its original label, for instance, the original label is *hair_easy* while the predicted label is *hair_hard*, we still use p as the label in the classification loss, since we suppose the distribution of the predicted alpha region should be as close as possible to that of the groundtruth alpha region.

To further improve the consistency, we compute the feature reconstruction loss between the features of predicted alpha and groundtruth alpha, which is a part of the perceptual loss proposed in [21]. Conventionally, L_2 loss is performed between $\hat{f}_k$ and f_k on each level separately.

Content-Sensitive Weights. Each matting class represents a distinct appearance and structure and thus its respective color and alpha exhibit different gradient distributions from other classes. For instance, *Hair* consists of fine structures with large gradients along hair boundaries, while *Fire* exhibits smooth transition across its foreground region.

Based on this observation, we introduce gradient constraints in our framework. The gradient of Eq. 1 is

$$\nabla I = (F - B)\nabla\alpha + \alpha\nabla F + (1 - \alpha)\nabla B \qquad (2)$$

Given an image, within the unknown region, $F - B$ and α are not available to the model. Thus, we learn 2D content-sensitive weights as different regularization coefficients to balance the gradient contributions among F, B, α for I. Specifically, the gradient constraint Eq. 2 is re-formulated as Eq. 3 with 2D learnable weights λ_1, λ_2.

$$\nabla I = \lambda_1\nabla\alpha + (1 - \lambda_2)\nabla F + \lambda_2\nabla B \qquad (3)$$

By introducing gradient constraints with learnable content-sensitive weights, our framework learns the implicit relation between the gradient contribution and semantic representation, which guides the model to distinguish the source of the image structure.

4.2. Losses

We jointly employ multiple losses, including reconstruction losses, classification loss, feature reconstruction loss, and gradient-related losses.

Reconstruction Losses. The reconstruction losses include alpha reconstruction loss L_α and F, B reconstruction loss L_{FB}. Alpha reconstruction loss is composed of difference loss and composition loss as well as Laplacian pyramid loss L_{lap} proposed by Hou *et al.* [20], which is defined as Eq. 4. For foreground and background reconstruction loss, we only consider the pixels within $\widetilde{F}$ and $\widetilde{B}$ respectively, as $\widetilde{F} = F + U$ and $\widetilde{B} = B + U$:
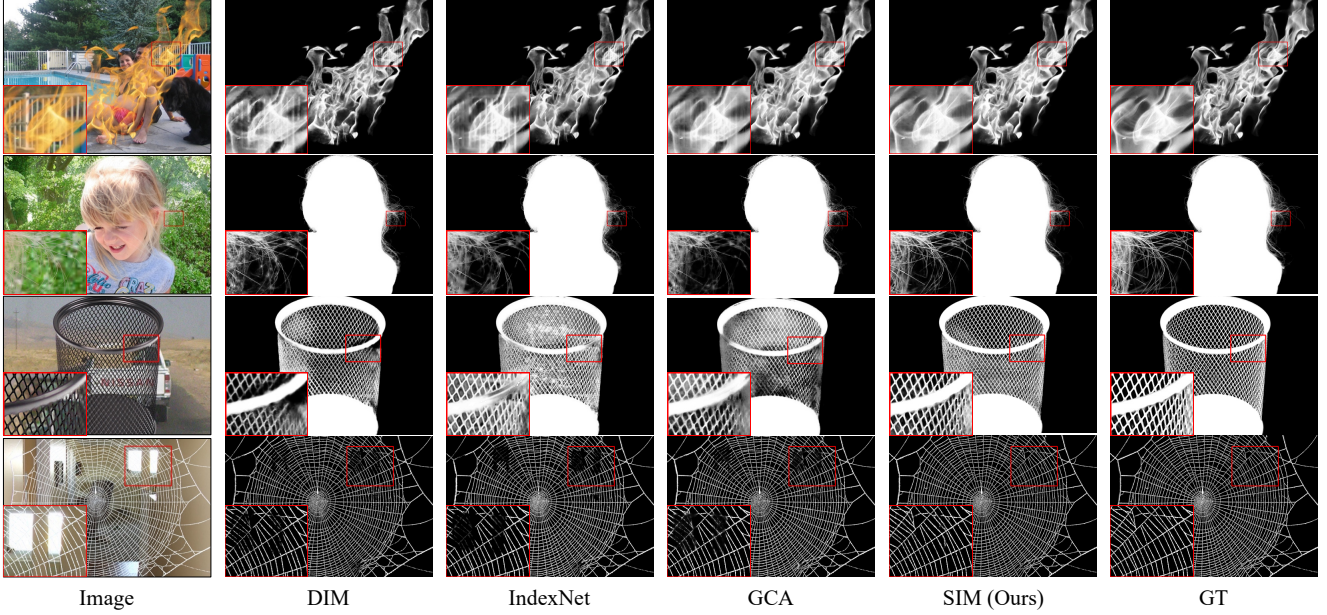
Figure 8. Qualitative results on the Semantic Image Matting test set.

$$L_\alpha = \frac{1}{|U|} \sum_{i \in U} \|\hat{\alpha}_i - \alpha_i\|_1 + \frac{1}{|U|} \sum_{i \in U} \|\hat{I}_i - I_i\|_1 + L_{lap} \quad (4)$$

$$L_{FB} = \frac{1}{|\widetilde{F}|} \sum_{i \in \widetilde{F}} \|\hat{F}_i - F_i\|_1 + \frac{1}{|\widetilde{B}|} \sum_{i \in \widetilde{B}} \|\hat{B}_i - B_i\|_1 \quad (5)$$

Classification and Feature Reconstruction Loss. As discussed above, our multi-class discriminator regularizes the model from a semantic level by classification loss L_c and feature reconstruction loss L_f defined as below:

$$L_c = -\sum_j \hat{p}_j \log p_j \quad (6)$$

$$L_f = \sum_k \frac{1}{|f_k|} \|\hat{f}_k - f_k\|_2 \quad (7)$$

Gradient-related Loss. To enforce the model to obey the gradient constraint, we utilize the loss with learnable content-sensitive regularization weights as Eq. 8:

$$L_g = \frac{1}{|U|} \sum_{i \in U} \|\nabla \hat{I}_i - \nabla I_i\|_1 \quad (8)$$

$$\nabla \hat{I}_i = \lambda_1 \nabla \hat{\alpha}_i + (1 - \lambda_2) \nabla \hat{F}_i + \lambda_2 \nabla \hat{B}_i \quad (9)$$

In addition, to fully exploit the gradient constraint, an exclusion loss used in a similar form in [45] is defined in Eq. 10 to avoid the structure of the image leaking into both foreground and background:

$$L_e = \frac{1}{|U|} \sum_{i \in U} \|\nabla \hat{F}_i\|_1 \|\nabla \hat{B}_i\|_1 + \|\nabla \hat{\alpha}_i\|_1 \|\nabla \hat{B}_i\|_1 \quad (10)$$

Methods	SAD	MSE(10^3)	Grad	Conn
DIM [43]	48.07	15.0	31.67	46.26
IndexNet [30]	51.29	14.0	34.19	48.77
GCA [26]	39.28	11.0	28.70	36.03
SIM (Ours)	**27.87**	**4.7**	**11.57**	**20.83**

Table 1. Comparisons on Semantic Image Matting Dataset.

Total Loss. The total loss is thus the weighted sum of the above losses defined as:

$$L = L_\alpha + 0.2(L_{FB} + L_f + L_g + L_e) + 0.1 L_c \quad (11)$$

5. Experiments

5.1. Implementation Details

We train our models on the Semantic Image Matting Dataset. Specifically, we take the images in COCO dataset [28] as background images. For each foreground object, we randomly select an image from the backgrounds and composite them using Eq. 1. Before composition, we perform various augmentations on both foreground and background. For the foreground, we randomly crop a square patch from size [320, 480, 640] and apply random scaling, rotation, color jittering, and horizontal flipping on the patch. Meanwhile, we randomly crop a patch from the background image and then generate inputs from the two patches. To extend training samples, we also randomly merge two foregrounds into one following [26]. We dilate and erode alpha matte with a random kernel size within [2, 9] and a random iteration within [5, 15] for trimap generation. More details can be found in supplementary materials.
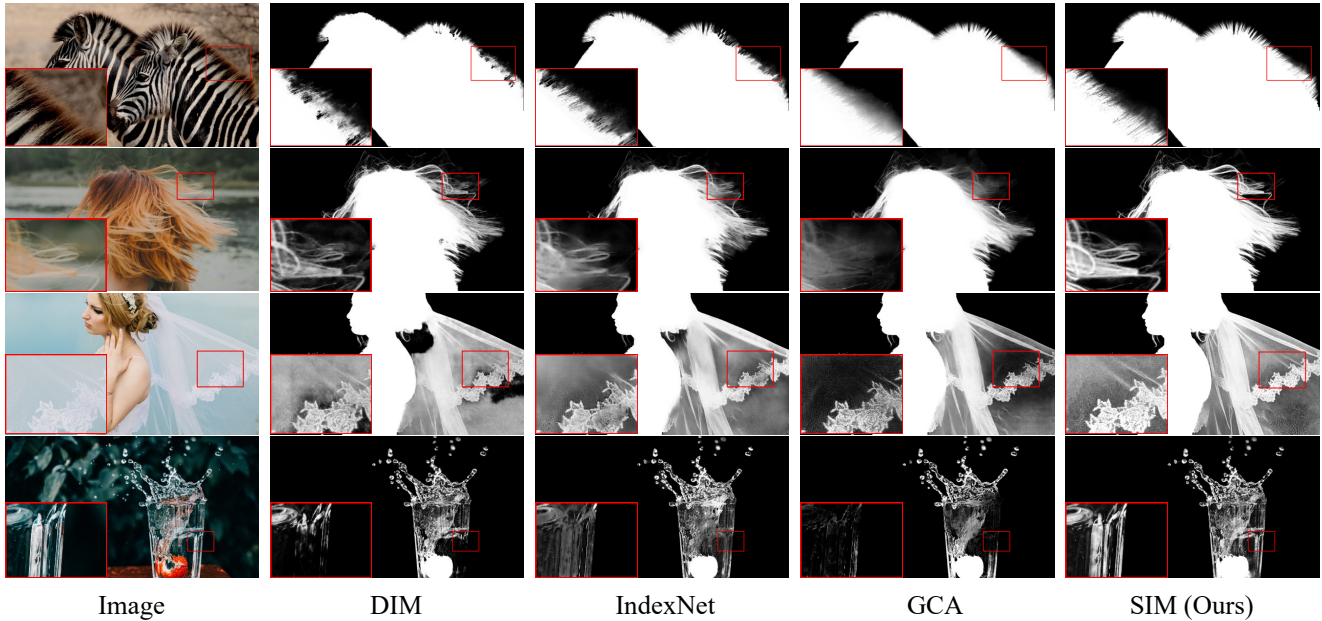
| Image | DIM | IndexNet | GCA | SIM (Ours) |

Figure 9. Qualitative results on the real-world images.

Methods	SAD	MSE(10^3)	Grad	Conn
Closed-Form [24]	168.1	91.0	126.9	167.9
KNN-Matting [10]	175.4	103.0	124.1	176.4
DCNN-Matting [11]	161.4	87.0	115.1	161.9
Learning-Based [47]	113.9	48.0	91.6	122.2
Information-flow [3]	75.4	66.0	63.0	-
DIM [43]	50.4	14.0	31.0	50.8
AlphaGan [31]	52.4	30.0	38.0	-
IndexNet [30]	45.8	13.0	25.9	43.7
AdaMatting [6]	41.7	10.0	16.9	-
SampleNet [40]	40.4	9.9	-	-
Context-Aware [20]	35.8	8.2	17.3	33.2
GCA [26]	35.3	9.1	16.9	32.5
SIM (Ours)	**28.0**	**5.8**	**10.8**	**24.8**

Table 2. Quantitative results on Composition-1K [43].

5.2. Main Results

Results on Semantic Image Matting Dataset. Table 1 tabulates the quantitative comparisons with other methods on our Semantic Image Matting Dataset. Our method outperforms them and achieves state-of-the-art performance. Table 3 further lists the quantitative results on each matting class. Our method also performs better than other methods on all matting classes. Figure 8 show the qualitative comparisons of some classes on the test set.

Results on Composition-1K. We also evaluate our model on the Composition-1K dataset proposed by Xu et al. [43], which has 50 unique foreground objects and 1000 testing samples. Table 2 shows the quantitative results. Our method also achieves the most competitive state-of-the-art performance on this dataset.

Results on alphamatting.com. Table 4 shows the evaluation results on alphamatting.com [1] benchmark proposed in [34]. Our method outperforms other methods and ranks first on three metrics. Detailed comparisons are provided in supplementary materials.

Results on Real-World Images. To evaluate the generalization ability of our framework, we collect from the Internet a number of free real-world images of different matting classes with user-labeled trimaps. Figure 9 shows the qualitative results which further demonstrates the effectiveness and generalization ability of our framework.

5.3. Analysis

Analysis of Semantic Trimap. Figure 10 visualizes an example semantic trimap, which is the combination of a conventional trimap and an automatically generated 20-channel score maps. The conventional trimap is divided into foreground, background, and unknown regions, which does not provide any semantics of the unknown pixels. Although deep neural networks are capable of encoding the implicit semantics of alpha patterns into high-level features, many fail cases still cannot be avoided due to restricted receptive fields and complex background, especially when the conventional trimap is too coarse to provide enough clues for extracting the relevant foreground object. The semantics score maps on the other hand provide the much-needed prior knowledge for indicating the potential class of unknown pixels, which guides the network to generate more accurate predictions.

Analysis of Multi-Class Discriminator. The arguably most widely used criterion in matting optimization is pixel-to-pixel distance which unfortunately does not take any dis-

Classes	defocus	fire	fur	glass_ice	hair_easy	hair_hard	insect	motion	net	flower
DIM [43]	25.91	60.53	9.88	91.36	11.23	13.01	111.21	6.78	87.09	65.40
IndexNet [30]	22.86	97.85	9.99	91.95	8.33	13.24	130.52	6.68	91.43	59.60
GCA [26]	18.33	46.29	8.12	76.20	8.24	11.31	99.11	6.08	83.71	44.86
SIM (Ours)	**13.49**	**35.44**	**5.90**	**49.19**	**5.68**	**7.72**	**96.85**	**4.04**	**50.35**	**37.10**

Classes	leaf	tree	plastic_bag	sharp	smoke_cloud	spider_web	lace	silk	water_drop	water_spray
DIM [43]	45.43	91.71	65.44	2.96	48.21	145.57	101.78	51.89	32.48	41.96
IndexNet [30]	43.85	99.26	89.70	3.32	35.31	145.62	114.47	62.81	33.90	34.92
GCA [26]	41.12	87.61	47.40	3.35	41.18	107.14	80.51	51.93	25.83	31.12
SIM (Ours)	**20.98**	**34.14**	**36.70**	**1.39**	**27.42**	**63.79**	**51.08**	**41.78**	**16.94**	**20.53**

Table 3. Quantitative results of 20 classes on Semantic Image Matting Dataset.

Methods	SAD				MSE	Grad
	Overall	S	L	U	Overall	Overall
AdaMatting [6]	7.6	6.9	6.5	9.4	8.5	8.1
SampleNet [40]	8.2	6.5	7.6	10.5	9.2	9.5
Background [37]	7.9	5.9	5.4	12.4	7.4	6.9
GCA [26]	9	10	6.4	10.8	9.9	8.2
SIM (Ours)	**2.5**	**2.6**	**1.8**	**3**	**2.9**	**3.1**

Table 4. Quantitative results of our method and several representative state-of-the-art methods on alphamatting.com [1] benchmark. "S", "L", "U" denote three trimap sizes and scores denote average rank across 8 test samples. Best results are shown in bold.

Methods	SAD	MSE(10^3)	Grad	Conn
Basic	32.04	5.9	12.05	26.20
Basic + S	30.24	5.4	11.60	23.83
Basic + S + D	29.84	5.3	12.37	23.33
Basic + S + D + G	27.87	4.7	11.57	20.83

Table 5. Ablation studies. The basic model is trained with reconstruction losses. "S", "D", "G" denotes semantic trimap , multi-class discriminator and gradient-related losses respectively.

tribution of a region into consideration. Thus, we propose a multi-class discriminator to make the model learn the distribution of different classes by taking into consideration class-specific statistics of alpha mattes.

Previous works [41, 32] had exploited natural image statistics and utilized them to boost performance in image reconstruction or translation. As for matting, if the predicted mattes are satisfactory, such predictions are supposed to have similar distributions to those of the groundtruth. The classification and feature reconstruction losses provided by the multi-class discriminator enforces the model to preserve the statistics of different patterns and consequently improves the performance.

Analysis of Content-Sensitive Weights. During the training stage, we introduce the gradient constraints with learnable weights to regularize the model on different classes. Figure 11 visualizes sample learnable weights. From this example, we observe that for the semi-transparent smooth textures such as silk, the alpha matte and the background contribute most to the image gradient. With the learnable content-sensitive weights, the performance is improved by
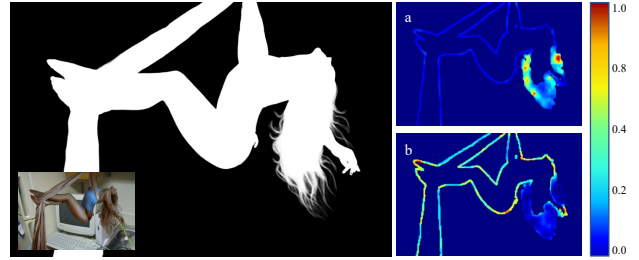

Figure 10. An example semantic trimap. We visualize two channels of the score maps: $a.\ hair\_hard$; $b.\ sharp$.

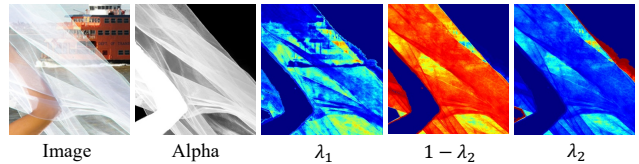| Image | Alpha | λ_1 | $1 - \lambda_2$ | λ_2 |

Figure 11. Visualization of learnable weights.

1.97 on SAD as shown in Table 5.

6. Conclusion

In this paper, we make use of the semantics underlying different alpha mattes, and develop a method incorporating semantic classification of matting regions for extracting better alpha matte. Specifically, we first cluster 20 classes according to regional alpha patterns, which cover a wide range of matting scenarios. Based on the classes, we extend traditional trimap to semantic trimap which is automatically extracted by a classifier to provide valuable semantic information for predictions. To further preserve alpha statistics for each class, a novel multi-class discriminator is designed to regularize the model according to class-specific distributions. Finally, we introduce an alpha gradient constraint with learnable content-sensitive weights as a new regularization to achieve better optimization of different classes. We conduct extensive experiments on Semantic Image Matting Dataset, Composition-1K dataset, and real-world images. Quantitative and qualitative results demonstrate the clear advantages of our method over existing methods.

References

[1] alphamatting. http://www.alphamatting.com/.

[2] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics*, 37(4):1–13, 2018.

[3] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *CVPR*, 2017.

[4] Benish Amin, Muhammad Mohsin Riaz, and Abdul Ghafoor. A hybrid defocused region segmentation approach using image matting. *Multidimensional Systems and Signal Processing*, 30(2):561–569, 2019.

[5] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*, 2007.

[6] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, 2019.

[7] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. Tom-net: Learning transparent object matting from a single image. In *CVPR*, 2018.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *ACM MM*, 2018.

[10] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. In *CVPR*, 2012.

[11] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *ECCV*, 2016.

[12] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *CVPR*, 2001.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[14] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *ECCV*, 2016.

[15] Marco Forte and François Pitié. F, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020.

[16] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. *Computer Graphics Forum*, 29(2):575–584, 2010.

[17] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, 2005.

[18] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR*, 2011.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[22] Rolf Köhler, Michael Hirsch, Bernhard Schölkopf, and Stefan Harmeling. Improving alpha matting and motion blurred foreground estimation. In *ICIP*, 2013.

[23] Xiang-Nan Kong, Michael K Ng, and Zhi-Hua Zhou. Transductive multi-label learning via alpha matting. *Unpublished manuscript*, 2009.

[24] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. In *CVPR*, 2006.

[25] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. In *CVPR*, 2007.

[26] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI*, 2020.

[27] Hai Ting Lin, Yu-Wing Tai, and Michael S Brown. Motion regularization for matting motion blurred objects. *TPAMI*, 33(11):2329–2336, 2011.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[29] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-Sheng Hua. Boosting semantic human matting with coarse annotations. In *CVPR*, 2020.

[30] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019.

[31] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. In *BMVC*, 2018.

[32] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018.

[33] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, 2020.

[34] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *CVPR*, 2009.

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

[36] Mark A Ruzon and Carlo Tomasi. Alpha estimation in natural images. In *CVPR*, 2000.

[37] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020.

[38] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, 2016.

[39] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Soft color segmentation and its applications. *TPAMI*, 29(9):1520–1537, 2007.

[40] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *CVPR*, 2019.

[41] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.

[42] Jue Wang, Michael F Cohen, et al. Image and video matting: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(2):97–175, 2007.

[43] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017.

[44] Sai Kit Yeung, Chi-Keung Tang, Michael S. Brown, and Sing Bing Kang. Matting and compositing of transparent and refractive objects. *ACM Transactions on Graphics*, 30(1):1–13, 2011.

[45] Xuaner Cecilia Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018.

[46] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, 2019.

[47] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *ICCV*, 2009.

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[49] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *ACM MM*, 2017.