Semantics-Adding Flaw-Erasing Network for Semantic Human Matting

Jiayu Sun[1,2]
jiayusun666@gmail.com
Zhanghan Ke[2]
kezhanghan@outlook.com
Ke Xu[2]
kkangwing@gmail.com
Fan Shao[3]
shaofan075@126.com
Lihe Zhang[1]
zhanglihe@dlut.edu.cn
Huchuan Lu[1,4]
lhchuan@dlut.edu.cn
Rynson W.H. Lau[2]
rynson.lau@cityu.edu.hk

[1] School of Information and Communication Engineering Dalian University of Technology Dalian, China

[2] Department of Computer Science City University of Hong Kong, Hong Kong SAR, China

[3] Wonxing Technology, Inc. Beijing, China

[4] Peng Cheng Laboratory Shenzhen, China

Abstract

Addressing human image matting without trimap is very challenging. The latest methods rely on estimating a segmentation map or a pseudo trimap to constrain the matting process. However, their matting accuracy typically affects by the errors in these auxiliary maps. Motivated by recent flaw correction approaches, we propose a novel neural approach to address this problem: We first train a model to directly compute an initial matte, of which the errors are further detected by a flaw detector and corrected by a refinement process. Our method, named Semantics-Adding Flaw-Erasing network (SAFE-Net), has two novel modules: a Semantic Addition module (SAM) to enrich matting features with human semantics via an attention mechanism and a Flaw Elimination module (FEM) to correct errors in the defective matte regions. To facilitate the learning process, we have further constructed a large human matting dataset containing 4,729 unique foregrounds with fine annotations. Extensive experiments demonstrate that SAFE-Net outperforms existing trimap-free human image matting methods.

1 Introduction

Semantic human matting aims to accurately separate the humans from a given image . It is a task with many practical applications, *e.g.*, image composition. Given an RGB image as input, human matting can be divided into two subtasks: foreground prediction and

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

(a) Input image (b) Pseudo trimap (c) Matte of (b) (d) Ground truth (e) Our initial matte (f) Our final matte

Figure 1: Limitations of the pseudo trimap based human matting approach. Given an input image (a), a state-of-the-art pseudo trimap based method [21] would predict a pseudo trimap first (b) and then an alpha matte (c). However, errors appearing in the pseudo trimap deteriorate the alpha matte prediction. Our approach first predicts an initial alpha matte (e), to which we then apply a flaw detector for detecting and correcting errors to produce our matte (f).

transparency (or foreground probability) estimation. Modeling these two subtasks simultaneously can easily result in incomplete foregrounds and inaccurate transparency due to the interference between the semantic information used for foreground partitioning and the appearance cues used for transparency estimation. Previous methods typically require the user to provide an additional trimap as a priori to help obtain the matte from the input RGB image. These trimap-based methods simplify the matting problem to estimating only the transparency within the unknown area of the trimap. However, the requirement of creating a trimap limits the matting task to non-realtime applications only. Besides, it may not be easy for a non-expert user to create an accurate trimap.

Some recent methods [6, 21] propose to predict the alpha matte without pre-defined trimaps. Instead, they first predict a psuedo trimap, and then perform the trimap-based matting procedure. However, there are two major problems of using the pseudo trimap. First, it is challenging to generate a high-quality pseudo trimap for two reasons: (1) A direct mapping between the transparency regions and the semantics of the surrounding foreground/background pixels does not exist; (2) The optimized parameters of morphological processing (*i.e.*, erosion and dilation, which are used to generate the extra annotations for the pseudo trimap prediction), should vary according to the proportion of the foreground region in the image but are fixed in existing methods. Second, erroneous labelling in the prior pseudo trimap can greatly affect the quality of the predicted alpha matte (as shown in Figure 1(b, c)),as a mistaken partition in a certain pseudo trimap region may misguide the alpha prediction process in the corresponding matte region.

Recently, a flaw detection method for semi-supervised learning is proposed in [16] . It minimizes the pixel-wise flaw of the feedforward result to achieve a better performance. Inspired by this work, we propose in this paper an iterative semantics-adding flaw-erasing network (SAFE-Net) for semantic human matting. SAFE-Net predicts an initial alpha matte, which is then refined by an iterative refinement process under the guidance of a flaw detector. Specifically, we apply a single encoder-decoder network to predict the initial alpha matte. We then iteratively refine the alpha matte to obtain the final alpha matte (as shown in Figure 1(e, f)) through a novel semantic addition module (SAM) and a novel flaw elimination module (FEM). During refinement, SAM uses attention mechanisms to constrain the shallow details-rich features to converge to a common semantic space for partitioning image regions, while FEM aims to correct error predictions (detected by the flaw detector) in the alpha matte.

The lack of large-scale datasets with high-quality annotations seriously impedes the development of the human matting task. Existing public datasets are either small-scale [24, 33] or coarsely labeled [29]. Although two large-scale human matting datasets have been cre-

ated recently [6, 21], they are unavailable to the public. To address this limitation, we have constructed a new dataset including 4,729 unique foregrounds with fine annotations.

In summary, the main contributions of this paper are as follows: (1) We propose a novel method (*SAFE-Net*) for human image matting. SAFE-Net first predicts an initial alpha matte and then iteratively detects and corrects its errors. (2) SAFE-Net contains two novel modules: SAM and FEM. While SAM enriches matting features with human semantics, FEM corrects the prediction errors based on a flaw detector and the enriched matting features. (3) We build a large human matting dataset that contains 4,729 unique foregrounds with high-quality annotations, to facilitate the learning of matte representations. (4) We conduct extensive experiments to verify the effectiveness of the proposed SAFE-Net, demonstrating its advantages over the existing trimap-free human matting methods.

2 Related Works

Natural Image Matting estimates the opacity of objects in natural images. Previous methods typically take an auxiliary trimap as a semantic priori and focus only on transparency estimation. These methods can be classified into two categories: sampling-based methods [7, 8, 9, 11, 13, 14, 26] and affinity-based methods [1, 2, 3, 5, 10, 18, 19]. Both kinds of methods only consider low-level pixel properties, such as the color, texture and intensity, without considering high-level semantic information. Driven by the rapid development of CNNs, many learning-based methods are proposed that introduce high-level semantic context into the natural image matting task. Xu *et al.* [33] propose a fully convolutional framework with a large-scale dataset for image matting, which is a milestone work at the time. Cai *et al.* [4] and Hou and Liu [12] treat matting as a multi-task learning problem. Lu *et al.* [22] design a network to generate indices to guide pooling and upsampling, which avoids the detail loss caused by pooling. Sun *et al.* [30] propose to first generate a semantic trimap from the input trimap, and then exploit the semantics for matting. Some methods propose to replace the trimap with other inputs to constrain the matting process, *i.e.,* segmentation mask [35] and background image without matting target [20, 28]. However, all these methods require additional inputs, which may limits them in real-time applications. To address this limitation, some recent methods [15, 24, 34, 36] directly predict the alpha matte without using trimaps but they often produce inaccurate mattes.

Semantic Human Matting has attracted extensive research interests recently. Unlike natural image matting that covers diverse object categories, human matting only focuses on extracting humans from the input images. It generates an alpha matte for the humans in the foreground of the image, especially for their fine-grained details, such as hairs. Several learning-based methods have been proposed for semantic human matting in recent years. Shen *et al.* [29] propose a semi-deep portrait matting method. However, due to the use of a shape prior, their model can only apply to half-length portraits. Chen *et al.* [6] utilize a segmentation network to generate low-resolution segmentation map to guide the matting process. Liu *et al.* [21] propose a coupled pipeline with three encoder-decoder networks. However, the coarse annotation produces coarse prediction, which misleads the alpha matte prediction in the last network.

Different from the above methods, our SAFE-Net does not apply a segmentation-matting pipeline. Instead, we first predict an initial alpha matte directly from the input image. Since this initial matte may contain errors, we then iteratively correct its errors to produce the output matte.

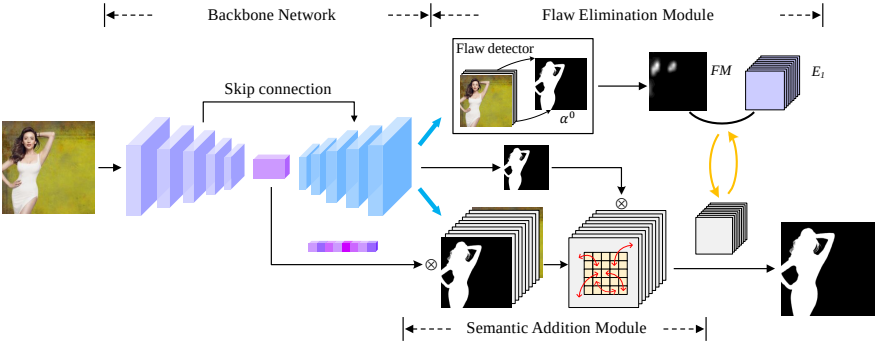

Figure 2: Architecture of the proposed method. Our network consists of a backbone network, a semantic addition module (SAM), and a flaw elimination module (FEM). The network predicts an initial alpha matte α^0 using the backbone network, and it then predicts a first refined alpha matte which corrects the erroneous regions in the initial alpha matte based on human semantic context learning and the guidance of the flaw detector. We can then feed the first refined alpha matte to SAM and FEM multiple times, so that the matting performance will further improve.

3 Our Method

As shown in Figure 2, our human matting pipeline consists of three components: a backbone network, a semantic addition module (SAM), and a flaw elimination module (FEM). Given an input image, the backbone network first produces an initial alpha matte. This initial matte is then fed into SAM to enhance the semantics. Next, FEM takes the initial matte and the semantic-enhanced features to predict the first refined alpha matte. The first refined alpha can then iteratively feed into SAM and FEM, and finally we have an output alpha matte. We explain our method in details below.

3.1 Backbone Network

Our backbone network is based on an auto-encoder architecture. We adopt a lightweight MobileNetV2 [27] model as the encoder E. Given an input image I, the encoder outputs a group of feature maps after each convolutional block. We integrate them as multi-scale feature maps and denote them by $(E_1, E_2, E_3, E_4, E_5)$, where the subscripts represent the indices of the convolution blocks. The resolution of these feature maps gradually decreases, *i.e.*, E_1 has the largest spatial resolution while E_5 has the smallest one. The decoder D consists of several convolutional layers and upsampling operations. We denote the side outputs of five decoder layers as $(D_1, D_2, D_3, D_4, D_5)$. To reuse the appearance details from E, we apply skip connections to transfer its feature maps to D. The backbone network produces an initial alpha matte. However, both semantics and details in α^0 are unsatisfactory, since it is difficult for a single backbone network to handle both localization information and appearance cues well simultaneously. To improve α^0, we propose to correct its flaw semantics as below.

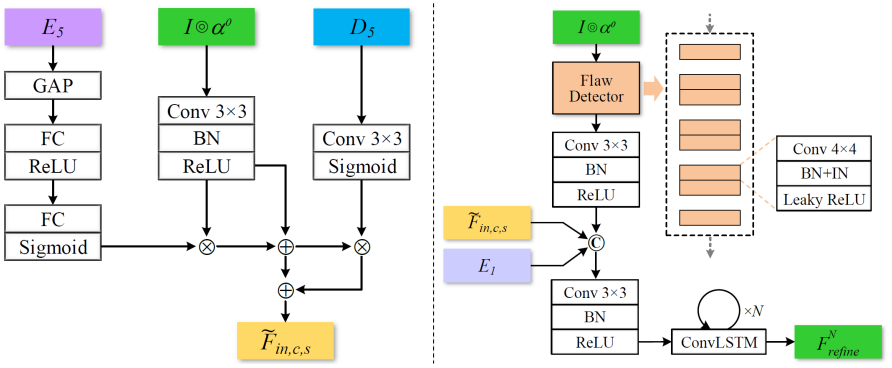

Figure 3: Semantic addition module (left) and Flaw Elimination Module (right).

3.2 Semantic Addition Module (SAM)

As analysed in [37], the feature map of each channel at the last convolutional layer E_5 usually highlights the discriminative class-specific regions for category identification. A global average pooling on E_5 and further usage of fully-connected layers can squeeze global spatial information into a vector called importance descriptor, of which each element represents the confidence of belonging to a specific class. In the human matting task, the foreground only contains the people category. Thus, this descriptor can embody a strong inherent pattern, that is, the corresponding elements of people have the highest responses. Therefore, we use the importance descriptor [37] as a channel-wise attention vector $Att_c \in R^{1\times1\times C}$ to constrain the semantically ambiguous F_{in}, which is obtained by operating a convolution (with kernel size of 3) on the concatenation of α^0 and I. Benefited from applying Att_c, F_{in} will converge to a state where the human-related class-specific channels have the most significant contributions.

Unlike the sharp foreground boundaries (with pixel values of either 0 or 1) predicted by semantic segmentation, the output of human matting requires a smooth transition (with pixel values ranging between 0 to 1) on human boundaries. Hence, we reinforce the effect of F_{in} to ensure the spatial smoothness of the features and avoid losing details as much as possible. This procedure is defined as:

$$\begin{aligned} Att_c &= \sigma(W_*(\delta(W_*(AvgPooling(E_5))))), \\ \widetilde{F}_{in,c} &= W_*(Att_c \otimes F_{in} + F_{in}), \end{aligned} \tag{1}$$

where $W_*(\cdot)$ denotes a convolution layer, $\delta(\cdot)$ refers to the ReLU function, $\sigma(\cdot)$ indicates a sigmoid function, and $\otimes$ refers to the element-wise multiplication.

In order to completely extract the humans, we regulate the spatial distribution of the features. We compute an alpha-guided spatial attention map $Att_s = \sigma(W_*(D_5)) \in R^{W\times H\times1}$, which is supervised by the ground truth alpha matte. Att_s and $\widetilde{F}_{in,c}$ are then multiplied in an element-wise manner to weight the features spatially. For the reason mentioned above, we also adopt a residual structure here to enhance the appearance information and maintain the global semantic context at the same time. This operation can be define as:

$$\widetilde{F}_{in,c,s} = W_*(Att_s \otimes \widetilde{F}_{in,c} + \widetilde{F}_{in,c}). \tag{2}$$

Figure 3 (left) shows the overall structure of SAM.

3.3 Flaw Elimination Module (FEM)

Through explicitly emphasizing the human-related semantics in the low-level features via SAM, we improve the accuracy of figure extraction. However, some important details, such as hairs, may still not be subtly characterized due to the coarseness of the semantic-biased representation. Hence, we propose to recover these important details by using a flaw map $\mathcal{FM}$ to indicates the inferior pixels in α^0.

The discriminator-like flaw detector [16] can predict pixel-wise flaw probabilities as $\mathcal{FM}$. The flaw detector was to repair errors through minimizing the values in the flaw map, *i.e.*, the map of flaw probabilities. In our pipeline, we modify the flaw detector for a different function. Instead of minimizing the flaw map, we use it as an indicator of inferior pixels. We first concatenate the input image I and the initial alpha α^0 as the input to the flaw detector to produce $\mathcal{FM}$. We then apply a convolutional layer (with kernel size of 3) on $\mathcal{FM}$ to obtain $F_{\mathcal{FM}}$, and concatenate $F_{\mathcal{FM}}$ with the side-output of the first convolutional layer E_1 and $\widetilde{F}_{in,c,s}$ as:

$$\widetilde{F}_{in,c,s,d} = W_*(F_{\mathcal{FM}} \circledcirc E_1 \circledcirc \widetilde{F}_{in,c,s}), \tag{3}$$

where $\circledcirc$ denotes concatenation. With this operation, the detail of the inferior pixels indicated by $\mathcal{FM}$ is purposefully enhanced. Finally, we feed the refined features $\widetilde{F}_{in,c,s,d}$ to a ConvLSTM [32] to further learn the spatial correlation between the features in different time steps t. The detail of the inferior pixels indicated by $\mathcal{FM}$ can be enhanced via the operations will provide in supplementary.

3.4 Loss Function

We leverage two losses to train our pipeline. One is the alpha prediction loss L_α^n. We use L_α^n to measure the absolute difference between the ground truth alpha matte α_g and the predicted alpha matte α_p, as:

$$L_\alpha^n = \gamma\sqrt{\left(\alpha_g - \alpha_p^n\right)^2 + \varepsilon^2}, \tag{4}$$

where ε is a small constant value. γ is a binary mask that is set to 4 in the unknown area and 1 otherwise. The purpose of γ is to let the alpha prediction loss pay more attention to the fine details, *e.g.,* the hair.

Both outputs of our network and the backbone network are supervised by an alpha prediction loss, as:

$$L_\alpha = L_\alpha^0 + L_\alpha^1, \tag{5}$$

where L_α^0 is the alpha prediction loss for the initial alpha matte α^0 predicted by the backbone network, and L_α^1 is the alpha prediction loss for the first refined alpha matte α^1. This loss can improve the quality of the initial alpha matte by repairing holes in the foreground and reducing artifacts in the background.

In addition, in order to detect the flaw regions in the predicted alpha mattes, we train the flaw detector by:

$$L_{fd} = \frac{1}{2}\left(\mathcal{FM} - \mathcal{FR}\right)^2, \tag{6}$$

where $\mathcal{FR}$ is the ground truth flaw map calculated according to [16]. We train the flaw detector and the other parts alternately.

Figure 4: Some examples in our dataset. The examples from the left to right are the example of single frontal subset, single pose-varied subset and multiple subset, respectively

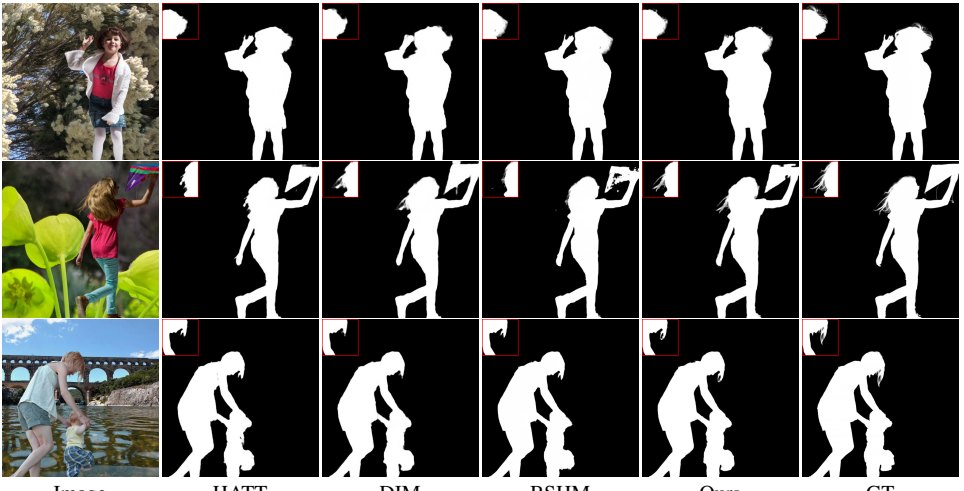

Figure 5: Qualitative comparison on the proposed dataset.

4 Human Matting Dataset

The lack of data is still a significant challenge for the semantic human matting task. Publicly available datasets typically lack diversity of the foreground humans, *e.g.,* the pose and the number of people. Although some large-scale datasets have been proposed, they are publicly available. Specifically, the most widely used natural image matting dataset [33] contains only 202 human images for training and 11 images for evaluation. And the humans-only subset of Distinctions646 [24] contains 362 training and 11 test samples. Due to their small size, these datasets cannot be used by itself for semantic human matting but as a supplement of other datasets. Shen *et al.* [29] propose the first human matting dataset which contains 2,000 upper-body portraits with limited poses. Moreover, their annotations are coarsely labeled, as they are generated by using traditional algorithms [5, 18], which have bias inevitably. Later, Chen *et al.* [6] propose a human fashion dataset with 34K images for matting. Liu *et al.* [21] propose a hybrid annotated dataset with 9,526 finely-annotated images and 10,597 coarsely annotated images. However, these two datasets are not publicly available. Hence, we propose a dataset with a training set of 4,494 and a test set of 235 unique foregrounds to address the data shortage problem in the area of semantic human matting. Supplementary Table 1 will compare these datasets.

Compared with the existing datasets, our dataset has richer postures or headcounts in each image. Based on the categories of posture and headcount, we split the dataset into three subsets, namely, single frontal subset, single pose-varied subset, and multiple subset. Supplementary Table 2 will show details of the three subsets of our dataset. The single

frontal subset contains the portraits of a single person facing towards the camera, while the single pose-varied subset contains the portraits of a single person with diverse types of human postures, such as figures in profile or figures viewed from behind. As for the multiple subset, it is the collection of some group photos. While the single frontal subsect is the largest of the three subsets and contains a total of $2,283$ images, the single pose-varied subset contains a total of $2,011$ images and the multiple subset contains a total of 435 images. We compose the above foregrounds and fine annotated alpha mattes with background images. We use OpenImages [17] and the SUN Database [31] as our background dataset for the composition. OpenImages is a dataset of 9 million images with more than 600 object classes. The SUN Database is a scene recognition benchmark with 899 categories and $130,519$ images. Each foreground image is composed with 10 randomly selected background images. However, we skip the background images that contain human to avoid confusion. Some examples of the three subsets are shown in Fig. 4.

5 Experiments

5.1 Experimental Setup

Data Augmentation. Most existing matting methods crop a high-resolution input image into low-resolution patches. However, this causes the inputs to have incomplete semantics during the training phrase, which severely affects the accuracy of feature extraction. Therefore, we propose a data augmentation strategy to fully exploit the deep semantic information of the input image. Our data augmentation strategy is as follows. We first properly scale the image and its corresponding ground truth with a random scaling factor. We then localize the bounding box of the human foreground using its corresponding ground truth. In order to guarantee that the image patch contains human head and part of the body, we vertically shift the center of the bounding box upward from the center of the image to a quarter of the image. In addition, we set the top of the cropped patch higher than that of the bounding box of the human foreground. To balance between semantic completeness and computational cost, we crop one 512×512 patch from each image. Finally, we randomly flip the patch and feed it to our network for training.

Evaluation Metrics. We adopt four widely used evaluation metrics to verify the performance of various models, including mean square error (MSE), sum of the absolution difference (SAD), Gradient error and Connectivity error. The first two metrics are objective indicators, and the last two metrics [25] are perceptual metrics to represent human subjective judgements. Lower values of these metrics correspond to a higher quality predicted alpha matte. Given a alpha matte with continuous values normalized to the range of 0 and 1, we compute these evaluation metrics with its corresponding ground truth. MSE, Gradient error and Connectivity error are averaged over the whole image.

Implementation Details. Our method is implemented using PyTorch toolbox [23], and trained on a PC with 2 NVIDIA TITAN RTX GPU. The flaw detector is trained using the Adam optimizer with learning rate initially set to 0.0001, which gradually decreases by a polynomial decay with power of 0.9. The other part of our model is trained in an end-to-end manner using the SGD optimizer with a fixed learning rate of 0.1. We trained our model 50 epochs with a batch size of 16. The flaw detector and the the other part of our model are trained alternately. During testing, we only input images and perform feed-forward inference

Method	Addtional Input	MSE↓	SAD↓	Grad↓	Conn↓
CF [18]	Trimap	72.51	42.81	33.45	14.77
DIM [33]	Trimap	28.36	14.23	15.85	4.916
SHM [6]	-	68.58	31.98	25.15	10.37
HATT [24]	-	33.85	15.08	15.18	5.203
BSHM [21]	-	53.18	19.50	15.57	6.812
Backbone	-	41.18	18.42	18.03	6.782
Ours	-	28.81	12.06	11.90	4.280

Table 1: Quantitative results of different methods on the proposed dataset.

to output alpha matte predictions. The iteration procedure is stopped at *iter t* if the SAD metric cannot be decreased further.

5.2 Performance Comparison

Since our method is trimap-free based, we compare it to two state-of-the-art trimap-free deep semantic human matting methods, *SHM* [6] and *BSHM* [21], and one trimap-free deep natural image matting method *HATT* [24]. We attach the performance of one trimap-based deep natural image matting method *DIM* [33], and one trimap-based non-deep natural image matting method *CF* [18] for reference. For a fair comparison, all deep learning based methods are retrained on our dataset. We follow DIM to generate trimaps from the ground truth alpha matte via dilation and erosion.

Quantitative Evaluation. Table 1 shows the evaluations on the proposed test set with $2,350$ images. For the sake of readability, we use $1e^{-4}$, $1e^{-4}$, $1e^{-3}$ and $1e^{-3}$ to scale mean square error (MSE), sum absolution difference (SAD), spatial gradient error (Grad), and connectivity error (Conn), respectively. All the metrics are the lower the better.

We can see that the proposed method consistently outperforms all trimap-free methods, *i.e.,* BSHM [21], SHM [6], and HATT [24] on all four evaluation metrics. This is because BSHM [21] and SHM [6] rely on the segmentation-matting pipeline, where errors in the predicted segmentation mask or pseudo trimap can significantly affect the subsequent matte prediction. The HATT [24] exploits adversarial learning to fit the image-to-matte mapping in a single stage, so it cannot correct the matte errors. Note that although DIM [29] (which is a trimap-based method) performs better in terms of MSE, our method can produce comparable results in terms of SAD, Connectivity error and Gradient error.

Method	MSE↓	SAD↓
HATT [24]	32.19	9.93
BSHM [21]	23.34	8.22
Ours	11.93	5.00

Table 2: Quantitative results of different methods on D646.

iter	MSE↓	SAD↓
1	12.46	5.17
2	11.98	5.01
3	11.93	5.00

Table 3: Quantitative results of different iteration on D646 dataset.

Table 2 shows the evaluations on the test set of D646 [24] dataset. Here, we use $1e^{-3}$ and $1e^{-3}$ to scale MSE and SAD respectively. It shows that our method outperforms HATT [24] and BSHM [21] on the D646 dataset.

	Bonebone	SAM	FEM	FEM*	MSE↓	SAD↓
val	✓				41.18	18.42
	✓	✓			37.64	15.25
	✓	✓	✓		28.81	12.06
	✓	✓		✓	37.29	15.07

Table 4: Ablation study on the test set of our proposed dataset.

Ablation Studies. We perform ablation analysis over the main components of the SAFE-Net and further investigate their importance and contributions. Table 4 shows the performance improvements contributed by different structures in terms of MSE and SAD. We can see that SAM and FEM can help significantly improve the performances. In addition, we use FEM* to represent a variant of FEM, which shares the same structure as FEM but uses zeros map instead of flaw map for guidance. We compare FEM with FEM* to validate the effectiveness of the flaw map. Moreover, Table 3 shows that iterative refinement can improve matting performance. *iter t* represents the result of the the t th refined alpha matte. We can see the matting performance improves after each iteration of refinement.

Visual Comparison. In Fig. 5, we visualize some alpha mattes produced by different matting approaches to qualitatively evaluate their performances. The examples embody various scenarios. Among the predicted alpha mattes, BSHM [5] (which is a non-trimap based method) shows an incomplete shape of human, caused by an inferior prediction of the trimap substitute, and HATT [24] shows unsatisfactory detail prediction. In contrast, our method achieves accurate figure extraction and detail prediction, similar to those of the trimap-based DIM [33].

In conclusion, these results indicate that our method can produce accurate alpha mattes even without the trimap as additional input.

6 Conclusion

In this paper, we have proposed a novel semantics-adding flaw-erasing Network (SAFE-Net) for semantic human matting. It first predicts an initial alpha matte with a single model. It then strengthens human semantics and erases erroneous regions in the initial alpha matte. In SAFE-Net, a backbone network is used to predict the initial alpha matte, and two novel modules (semantic addition module (SAM) and flaw elimination module (FEM)) are exquisitely designed to predict a finer alpha matte from the initial matte. Experimental results have shown that the proposed method performs favourably against the state-of-the-art methods. In addition, we have also built a fine-annotated dataset for semantic human matting, which contains 4,729 unique foregrounds.

Acknowledgements

This work was supported by the National Natural Science Foundation of China #62276046, and the Liaoning Natural Science Foundation #2021-KF-12-10.

References

[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In CVPR, pages 29–37, 2017.

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

[3] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In ICCV, pages 1–8, 2007.

[4] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In ICCV, pages 8819–8828, 2019.

[5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. IEEE TPAMI, 35(9): 2175–2188, 2013.

[6] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In ACM MM, pages 618–626, 2018.

[7] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In CVPR, volume 2, pages II–II, 2001.

[8] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In ECCV, pages 204–219. Springer, 2016.

[9] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In Computer Graphics Forum, volume 29, pages 575–584. Wiley Online Library, 2010.

[10] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In Proceedings of VIIP, volume 2005, pages 423–429, 2005.

[11] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In CVPR, pages 2049–2056, 2011.

[12] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In ICCV, pages 4130–4139, 2019.

[13] Jubin Johnson, Ehsan Shahrian Varnousfaderani, Hisham Cholakkal, and Deepu Rajan. Sparse coding for alpha matting. IEEE TIP, 25(7):3032–3043, 2016.

[14] Levent Karacan, Aykut Erdem, and Erkut Erdem. Image matting with kl-divergence based sparse sampling. In ICCV, pages 424–432, 2015.

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

[16] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In ECCV, August 2020.

[17] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html, 2017.

[18] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. IEEE TPAMI, 30(2):228–242, 2007.

[19] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. IEEE TPAMI, 30 (10):1699–1712, 2008.

[20] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In CVPR, pages 8762–8771, 2021.

[21] 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, pages 8563–8572, 2020.

[22] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In ICCV, pages 3266–3275, 2019.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, NeurIPS, pages 8024–8035. 2019.

[24] 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, pages 13676–13685, 2020.

[25] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In CVPR, pages 1826–1833, 2009.

[26] Mark A Ruzon and Carlo Tomasi. Alpha estimation in natural images. In CVPR, volume 1, pages 18–25, 2000.

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, pages 4510–4520, 2018.

[28] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In CVPR, pages 2291–2300, 2020.

[29] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In ECCV, pages 92–107. Springer, 2016.

[30] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In CVPR, pages 11120–11129, 2021.

[31] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, pages 3485–3492, 2010.

[32] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In NeurIPS, pages 802–810, 2015.

[33] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In CVPR, pages 2970–2979, 2017.

[34] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. In NeurIPS, 2018.

[35] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In CVPR, pages 1154–1163, 2021.

[36] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In CVPR, pages 7469–7478, 2019.

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, pages 2921–2929, 2016.