Boosting Semantic Human Matting with Coarse Annotations

Jinlin Liu[1,2] Yuan Yao[1] Wendi Hou[1] Miaomiao Cui[1]
Xuansong Xie[1] Changshui Zhang[2] Xian-sheng Hua[1]
[1]Alibaba Group, [2] Department of Automation, Tsinghua University
{ljl191782, ryan.yy, wendi.hwd, miaomiao.cmm}@alibaba-inc.com xingtong.xxs@taobao.com
zcs@mail.tsinghua.edu.cn xiansheng.hxs@alibaba-inc.com

Abstract

Semantic human matting aims to estimate the per-pixel opacity of the foreground human regions. It is quite challenging and usually requires user interactive trimaps and plenty of high quality annotated data. Annotating such kind of data is labor intensive and requires great skills beyond normal users, especially considering the very detailed hair part of humans. In contrast, coarse annotated human dataset is much easier to acquire and collect from the public dataset. In this paper, we propose to use coarse annotated data coupled with fine annotated data to boost end-to-end semantic human matting without trimaps as extra input. Specifically, we train a mask prediction network to estimate the coarse semantic mask using the hybrid data, and then propose a quality unification network to unify the quality of the previous coarse mask outputs. A matting refinement network takes in the unified mask and the input image to predict the final alpha matte. The collected coarse annotated dataset enriches our dataset significantly, allows generating high quality alpha matte for real images. Experimental results show that the proposed method performs comparably against state-of-the-art methods. Moreover, the proposed method can be used for refining coarse annotated public dataset, as well as semantic segmentation methods, which reduces the cost of annotating high quality human data to a great extent.

1. Introduction

Human matting is an important image editing task which enables accurate separation of humans from their backgrounds. It aims to estimate the per-pixel opacity of the foreground regions, making it valuable to use the extracted human image in some recomposition scenarios, including digital image and video production. One may refer this task as semantic segmentation problem [4, 7, 24], which achieves fine-grained inference for enclosing objects. However, segmentation techniques focus on pixel-wise binary classification towards scene understanding, although semantic information is well labelled, it could not catch complicated semantic details like human hair.

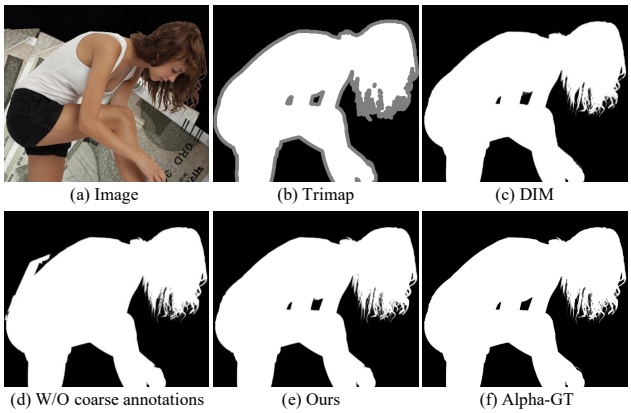


Figure 1. The user interactive method could catch precise semantics and details under the guidance of trimaps. Without the trimap and enough training dataset, one may get inaccurate semantic estimation, which inevitably leads to wrong matting results. Our methods achieve comparable matting results by leveraging coarse annotated data while do not need trimaps as inputs.

The matting problem can be formulated in a general manner. Given an input image I, matting is modeled as the weighted combination of foreground image F and background image B as follows [30]:

$$I_z = \alpha_z F_z + (1 - \alpha_z) B_z \,, \quad \alpha_z \in [0, 1] \,. \tag{1}$$

where z represents any pixel in image I. The known information in Eq. 1 are the three dimensional RGB color I_z, while the RGB color F_z and B_z, and the alpha matte estimation α_z are unknown. Matting is thus to solve the 7 unknown variables from 3 known values, which is highly under-constrained. Therefore, most existing matting methods take a carefully specified trimap as constraint to reduce

the solution space. However, a dilemma in terms of quality and efficiency for trimaps still exists.

The key factor that affecting the performance of matting algorithm is the accuracy of trimap. The trimap divides the image into three regions, including the definite foreground, the definite background and the unknown region. Intuitively, the smaller regions around foreground boundary that the trimap contains, the less unknown variables would be estimated, leading to a more precise alpha matte result. However, designing such an accurate trimap requires a lot of human efforts with low efficiency. The labeling quality should be unified among all the data, either large or small size of unknown regions will degrade the final alpha matte effects. One possible solution to solve the dilemma is adaptively learn a trimap from coarse to fine [28, 6]. In contrast, another solution discards the trimap from the input and employs it as an implicit constraint to a deep matting network[8, 33]. However, these methods still rely on the quality of the generated trimap, unable to retain both the semantic information and high quality details when implicit trimap is inaccurate.

Another limitation comes from the data for human matting. It is important to have high quality annotation data for image matting task. Since humans in natural images possess a variety of colors, poses, head positions, clothes, accessories, etc. The semantically meaningful structure around the foreground like human hair, furs are the challenging regions for human matting. Annotating such accurate alpha matte is labor intensive and requires great skills beyond normal users. Shen *et al*. [28] proposed a human portrait dataset with 2000 images, but it has strict constraint on position of human upper body. The widely used DIM dataset [32] is limited in human data, with only 213 human images. Although Chen *et al*. [8] created a large human matting dataset, it is only for commercial use. Unfortunately, collecting the dataset in [8] with 35,311 images takes more than 1,200 hours, which is undesirable in practice. Therefore, we argue that there is a solution by combining the limited fine annotated image with easily collected coarse annotated image for human matting.

To address the aforementioned problems, we propose a novel framework to utilize both coarse and fine annotated data for human matting. Our method could predict accurate alpha matte with high quality details and sufficient semantic information without trimap as constraint, as shown in Figure 1. We achieve this goal by proposing a coupled pipeline with three subnetworks. The mask prediction network (MPN) aims to predict low resolution coarse mask, which contains semantic human information. MPN is trained using both fine and coarse annotated data for better performance on various real images. However, the output of MPN may vary and are not consistent with respect to different input images. Therefore, a quality unification network (QUN) trained on hybrid annotated data is introduced to rectify the quality level of MPN output to the same level. A matting refinement network (MRN) is proposed to predict the final accurate alpha matte, taking in both the origin image and its unified coarse mask as input. Different with MPN and QUN, the matting refinement network is trained using only the fine annotated data.

We also constructed a hybrid annotated dataset for human matting task. The dataset consists of both high quality (fine) annotated human images and low quality (coarse) annotated human images. We first collect 9526 images/alpha pairs with fine annotations. In comparison with previous dataset, we diversity the distribution of human images with carefully annotated alpha matte [28, 32], within a labor rational volume size [8]. We further collect 10597 coarse annotated data to better capture accurate semantics within our framework. We follow [32] to composite both data onto 10 background images in MS COCO [23] and Pascal VOC [12] to form our dataset. Comprehensive experiments have been conducted on this dataset to demonstrate the effectiveness of our method, and our model is able to refine coarse annotated public dataset as well as semantic segmentation methods, which further verifies the generalization of our method. The main contributions of this work are:

- To our best knowledge, this is the first method that uses coarse annotated data to enhance the performance of end-to-end human matting. Previous methods either take trimap as constraint or use sufficient fine annotated dataset only.
- We propose a quality unification network to rectify the mask quality during the training process so as to utilize both coarse and fine annotations, allowing accurate semantic information as well as structural details.
- The proposed method can be used to refine coarse annotated public dataset as well as semantic segmentation methods, which makes it easy to create fine annotated data from coarse masks.

2. Related Work

Natural Image Matting. Natural image matting tries to estimate the the unknown area with known foreground and background in the trimap.

The traditional methods can be summarized to sampling based methods and affinity based methods [30]. The sampling based methods [11, 14, 15, 17, 19, 20, 27] leverage the nearby known foreground and background colors to infer the alpha values of the pixels in the undefined area. Assuming that alpha values for two pixels have strong correlations if the corresponding colors are similar. Following the assumption, various sampling methods are proposed including Bayesian matting [11], sparse coding [14, 19], global sampling [17] and KL-divergence approaches [20]. Com-

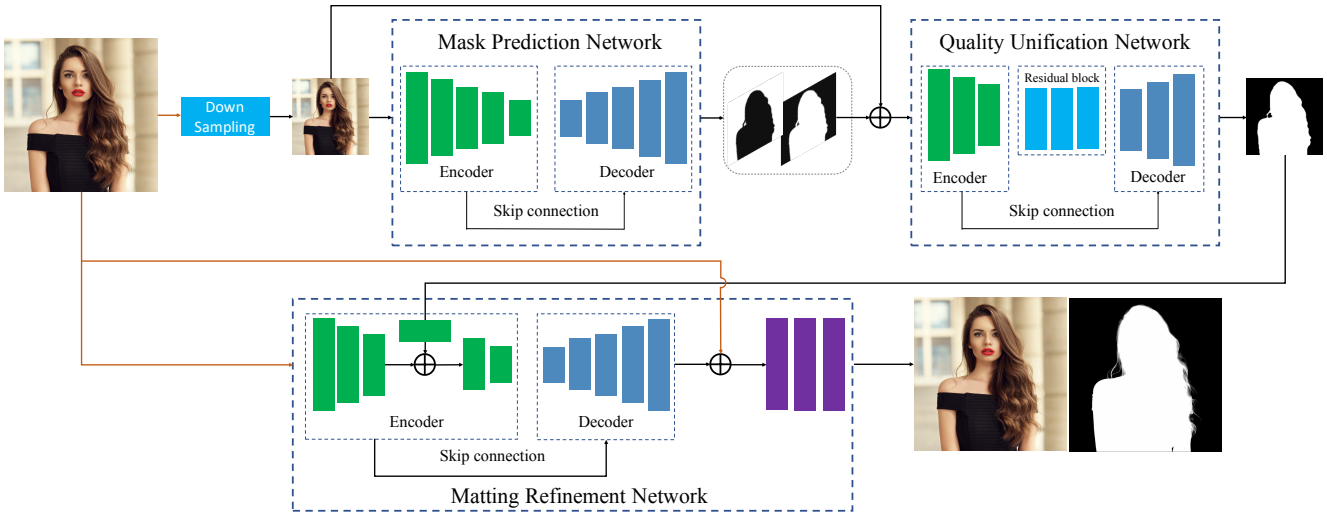


Figure 2. An overview of our network architecture. The proposed method is composed of three parts. The first part is mask prediction network (MPN), to predict low resolution coarse semantic mask. MPN is trained using both coarse and fine data. The second part is quality unification network (QUN). QUN aims to rectify the quality of the output from the mask prediction network to the same level. The rectified coarse mask is then unified and enables consistent input for training the following alpha matte prediction stage. The third part is matting refinement network (MRN), taking in the input image and the unified coarse mask to predict the final accurate alpha matte.

pared with sampling based methods, Affinity based methods [2, 3, 5, 9, 16, 21, 22, 29] define different affinities between neighboring pixels, trying to model the matte gradient instead of the per-pixel alpha value.

Deep learning based method is able to learn a mapping between the image and corresponding alpha matte in an end-to-end manner. Cho *et al*. [10] take the advantage of close-form matting [21] and KNN matting [9] for alpha mattes reconstruction. Xu *et al*. [32] integrate the encoder-decoder structure with a following refinement network to predict alpha matte. Lutz *et al*. [25] further employ the generative adversarial network for image matting task. Cai *et al*. [6] argue the limitation of directly estimating the alpha matte from a coarse trimap, and propose to disentangle the matting into trimap adaptation and alpha estimation tasks. Compared with the above methods, our method simply use RGB images as input without the constraint of designated trimaps.

Human image Matting. As a specific type of image matting, human matting aims to estimate the accurate alpha matte corresponding to the human in the input image, which involves semantically meaningful structures like hair. Recently, several deep learning based human matting methods [8, 28, 34] have been proposed. Shen *et al*. [28] propose a deep neural network to generate the trimap of a portrait image and add a matting layer[21] for network optimization using the forward and backward propagation strategy. Zhu *et al*. [34] use a similar pipeline and design a light dense network for portrait segmentation and a feature block to learn the guided filter [18] for alpha matte prediction. Chen *et al*. [8] introduce an automatic human matting algorithm without feeding trimaps. It combines a segmentation module with a matting module for end-to-end matting. The late fusion CNN structure in [33] integrates the foreground and background classification presents its capacity for human image matting. However, these models require carefully collected image/alpha pairs, which may also suffer from inaccurate semantics due to lack of fine annotated human data.

3. Proposed Approach

We develop three subnetworks as a sequential pipeline. The first one is mask prediction network (MPN), to predict coarse semantic masks using data at different annotation quality level. The second one is quality unification network (QUN). QUN rectifies the quality of the output coarse mask from MPN to the same level. The third part is matting refinement network (MRN), to predict the final accurate alpha matte. The flowchart and the network structure is displayed in Figure 2.

3.1. Mask Prediction Network

As no trimap is required as input, the first stage of the proposed method is to predict a coarse semantic mask. The network we use is encoder-decoder structure with skip connection, and we predict the foreground mask and the background mask at the same time. At this stage, we aim to estimate a coarse mask, and therefore the network is not trained at a high resolution. We resize all training data to resolution 192×160 so as to train the mask prediction network (MPN) efficiently. In addition, the mask predic-

tion network is trained using all training data, including low quality and high quality annotated data. The loss function to train LRPN is L_1 loss,

$$\mathcal{L}_{MPN} = \lambda_L |\alpha_p^c - \alpha_g^c|_1 + (1 - \lambda_L)|\beta_p^c - \beta_g^c|_1 \,, \quad (2)$$

where the output is a 2-channel mask, α_p^c denotes the first channel of the output, i.e., the predicted foreground mask, α_g^c denotes the ground truth foreground mask, β_p^c denotes the second channel of the output, i.e., the predicted background mask, and β_g^c denotes the ground truth background mask. We set $\lambda_L = 0.5$ in experiments.

3.2. Quality Unification Network

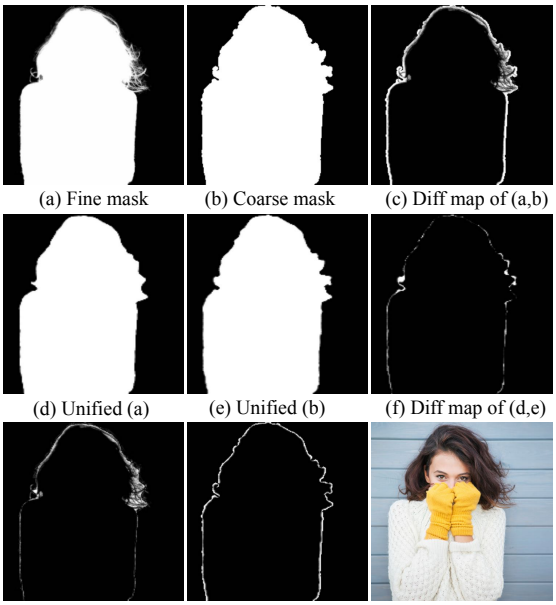
Due to the high cost of annotating high quality matting data, we propose to use hybrid data from different data source. Some of the data is annotated at high quality , even hairs are very well separated from the background (Figure 3(a)). Whereas, majority of other data are annotated at a relatively low quality (Figure 3(b)). Mask prediction network is trained with both fine annotated data and coarse annotated data. Thus, the quality of the predicted mask may vary significantly. As the alpha matte prediction network can only be trained on the high quality annotated data, the variation of the coarse mask quality will inevitably lead to inconsistent matting results during the inference stage. As illustrated in Figure 6(c), if the coarse mask is relatively accurate, the refinement network will work well to output accurate alpha matte. On the contrary, the refinement network will fail if the coarse mask lacks important details.

We proposed to eliminate the data bias for training matting refinement network by introducing a quality unification network (QUN). The quality unification network aims to rectify the output quality of the mask prediction network to the same level, by improving the quality of coarse masks and lowering the quality of fine masks simultaneously. The output of the mask prediction network and the original image are feed into the quality unification network to unify the quality level. The rectified coarse mask is unified and enables consistent input for training the following accurate alpha matte prediction stage.

The loss function of training QUN network contains two parts, identity loss and consistence loss. Identity loss forces the output of QUN not to change much from the original input,

$$\mathcal{L}_{identity} = |Q(x) - x|_1 + |Q(x') - x'|_1 \,, \quad (3)$$

where $Q(\cdot)$ represent the quality unification network. x denotes the concatenation of the input image and the accurate mask, x' denotes the concatenation of the input image and the inaccurate mask. The second part is consistence loss. Consistence loss forces the output of QUN corresponding to accurate mask and inaccurate mask to be close.



(a) Fine mask (b) Coarse mask (c) Diff map of (a,b)
(d) Unified (a) (e) Unified (b) (f) Diff map of (d,e)
(g) Diff map of (a,d) (h) Diff map of (b,e) (i) Input image

Figure 3. Different quality of masks are unified by QUN. (a) High quality mask. (b) low quality mask. (c) Difference map of high and low quality mask. (d) Unified result of high quality mask by QUN. (e) Unified result of low quality mask by QUN. (f) Difference map of the unified high quality mask and the low quality mask. (g) Difference map of the unified high quality mask and the original high quality mask. (h) Difference map of the unified low quality mask and the original low quality mask. (i) Input image.

$$\mathcal{L}_{consist} = |Q(x) - Q(x')| \,. \quad (4)$$

Thus, the loss function of training QUN is the weighted sum of identity loss and consistence loss,

$$\mathcal{L}_{QUN} = \lambda_1 \mathcal{L}_{identity} + \lambda_2 \mathcal{L}_{consist} \,. \quad (5)$$

During the training, we set $\lambda_1 = 0.25$ and $\lambda_2 = 0.5$.

In Figure 3, we illustrate the results of QUN. Fine mask (Figure 3(a)) and coarse mask (Figure 3(b)) are unified by QUN to Figure 3(d) and (e) respectively. The difference maps are also calculated. We can observe that the unified high quality mask become relatively coarser and low quality mask becomes relatively finer. As a result, the unified masks are much closer to each other than the original fine and coarse masks.

3.3. Matting Refinement Network

Matting refinement network (MRN) aims to predict accurate alpha matte. Therefore, we train MRN at a higher resolution ($768 * 640$ in all experiments). Note that the coarse mask from MPN and QUN is at low resolution (192×160). The coarse mask is integrated to MRN as external input feature maps, where the input is downscaled 4 times after several convolution operations. The output of

(a) Coarse annotated dataset

(b) Fine annotated dataset

Figure 4. Input images and the corresponding annotations in our dataset. Our dataset consists of both coarse annotated images (a) and fine annotated images (b).

MRN are 4-channel maps, including three foreground RGB channels and one alpha matte channel. Predicting the foreground RGB channels coupled with alpha matte is able to increase the robustness, which plays a similar role of the compositional loss used in [32, 8]. The loss function we used to train MRN is L_1 loss,

$$\mathcal{L}_{MRN} = \lambda_H |RGB_p - RGB_g|_1 + (1-\lambda_H)|\alpha_p - \alpha_g|_1 \,, \tag{6}$$

where RGB_p and RGB_g denote the predicted RGB foreground channels and ground truth foreground channels respectively. α_p and α_g denote the predicted alpha matte and ground truth alpha matte respectively. We set $\lambda_H = 0.5$ in experiments.

3.4. Implementation details

We implement our method with Tensorflow [1] framework. We perform training for our three networks sequentially. Before feeding into the mask prediction network, we conduct a down-sampling operation on images at 192×160 resolution, including both fine and coarse annotated data. Flipping is performed randomly on each training pair. We first train the mask prediction network for 20 epochs and fix the parameters. Then we concatenate the low resolution image and the output foreground mask as input to train quality unification network. When training QUN, random filters(filter size set as 3 or 5), binarization and morphology operations(dilate and erode) are exerted to fine annotated data to generate paired high and low quality mask data. After training quality unification network, all parameters are fixed. We finally train the matting refinement network with only the fine annotated data. The entire data pairs (image, alpha matte) are randomly cropped to 768×640. The learning rate for training all networks is $1e-3$. MPN and QUN are trained using batch size 16 and MRN is trained using batch size 1, as MRN is trained using only high resolution data.

When testing, a feed-forward pass of our pipeline is performed to output the alpha matte prediction with only the image as input. The average testing time on multiple 800×800 images is 0.08 seconds.

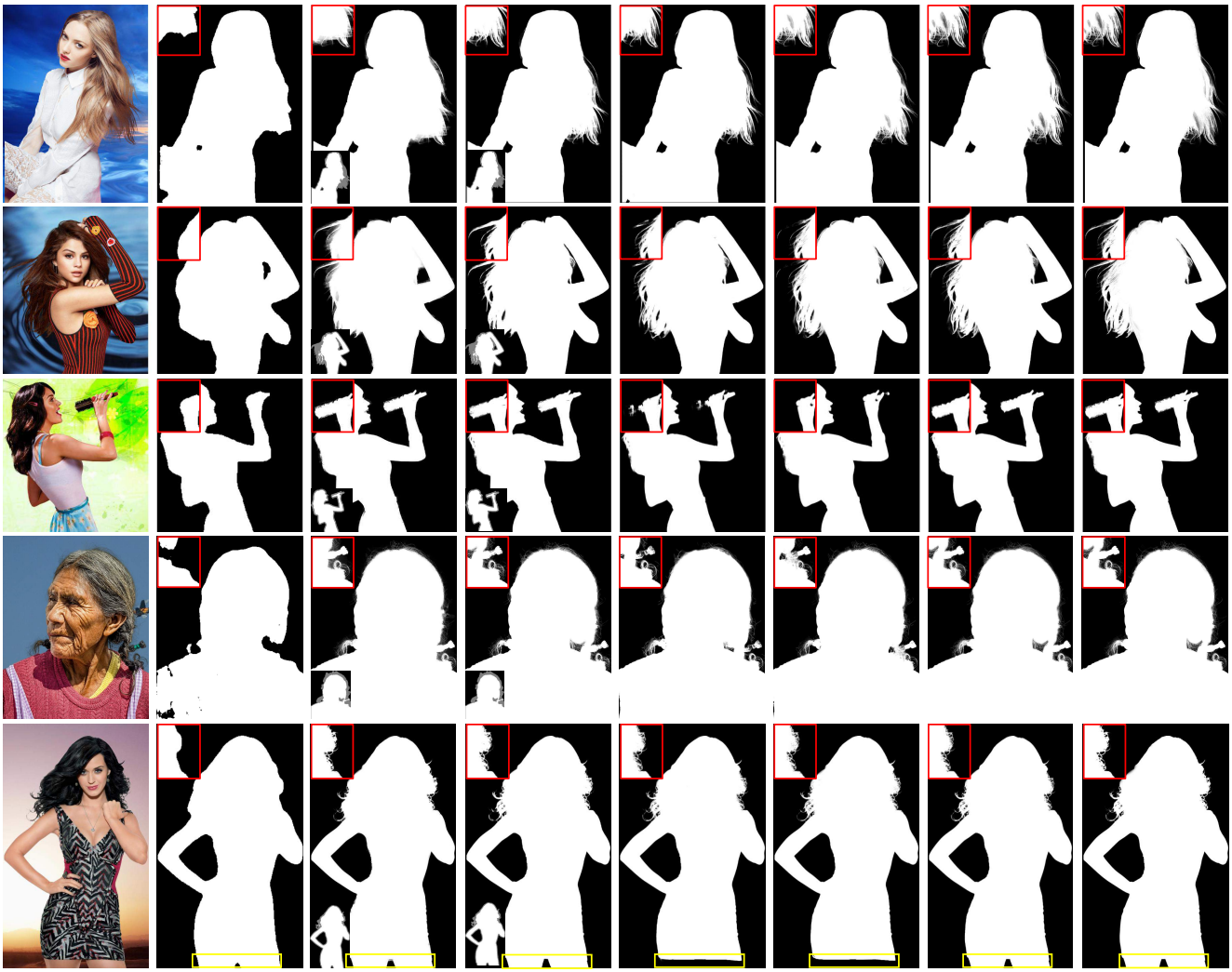
Table 1. The configurations of human matting datasets.

Dataset	Train Set		Test Set	
	Human	image	Human	image
Shen *et al*. [28]	1700	1700	300	300
TrimapDIM [32]	202	20200	11	220
SHM [8]	34493	34493	1020	1020
Ours(coarse)	10597	105970	125 (+11)	1360
Ours(fine)	9324(+202)	95260		

4. Human matting dataset

A main challenge for human matting is the lack of data. Xu *et al*. [32] proposed a general matting dataset by compositing foreground objects from natural images to differents backgrounds, which has been widely used in the following matting works[6, 25, 33]. However, the diversity of human images is severely limited, including only 202 human images in training set and 11 human images in testing set. For human matting dataset, Shen *et al*. [28] collected a portrait dataset with 2000 images, it assumes that the upper body appears at similar positions in human images and the images are annotated by Closed From [21], KNN [9] methods, which can be inevitably biased. Although a large human fashion dataset is created by [8] for matting, it is only for commercial use. To this end, we create a human matting dataset with high-quality for research. We carefully collected 9449 diverse human images with simple background from the Internet (i.e., white or transparent background in PNG format), each human image acquires a well annotated alpha matte after simple processing. The human images are split to training/testing set, with 9324 and 125 respectively. Following Xu *et al*. [32], we first add the human images in DIM dataset[32] into our training/testing set, forming a total of 9526 and 136 human foregrounds respectively. We then randomly sample 10 background images in MS COCO [23] and Pascal VOC [12] and composite the human images onto those background images. During composition, we ensure that the background images are not containing humans.

Another issue should be addressed for human matting dataset is the quality of annotations. Image matting task requires user designated annotations for objects, i.e., the high quality alpha matte. Besides, the user interaction methods require carefully prepared trimaps and scribbles as constraints, which is labor intensive and less scalable. Method without user provided trimaps is to predict the alpha matte by first generating implicit trimaps for further guidance, thus lead to some artifacts as well as losing some semantics



(a) Image (b) DeepLab (c) Closed-form (d) DIM (e) SHM (f) Ours (fine data) (g) Ours (h) Alpha-GT

Figure 5. The qualitative comparison on our proposed dataset. The first column and the last column show the input image and the ground truth alpha matte, and the rest columns present the estimation results by DeepLab [7], Closed-form matting [21], DIM [32], SHM [8], our method trained using fine annotated data only and our method trained using hybrid annotated data.

for complex structures. We integrate the coarse annotation data to tackle this problem as they are much easier to obtain. We collect another 10597 human data from [31] and Supervisely Person Dataset, and follow the above setup to generate 105970 image with coarse annotations.

Table 1 shows the configuration of the existing human matting dataset. Our dataset consists of both fine and coarse annotated data, with nearly the same amount. Compared with user interactive methods [28, 32], our dataset covers diverse high quality human images, making it more robust for human matting models. Although sacrifice the number of high quality annotations than automatic method [8], we introduce coarse annotated data to enhance the capacity for extracting both semantic and matting details at a lower cost. The data for both annotations are shown in Figure 4.

5. Experiments

5.1. Evaluation results.

Evaluation metrics. We adopt four widely used metrics for matting evaluations following the previous works [32, 8]. The metrics are MSE (mean square error), SAD (sum of the absolution difference), the gradient error and the connectivity error. The gradient error and connectivity error proposed in [26] are used to reflect the human perception towards visual quality of the alpha matte. Lower values of these metrics correspond to better estimated alpha matte. We normalize the estimated alpha matte and true alpha matte to [0, 1] to calculate these evaluation metrics. Since no trimap is required, we calculate over the entire images and average by the pixel number.

Table 2. The quantitative results.

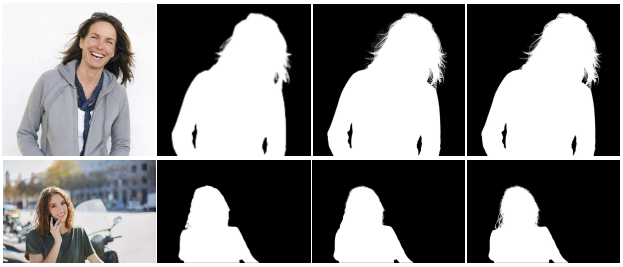
Method	SAD	MSE	Gradient	Connectivity
DeepLab [7]	0.028	0.023	0.012	0.028
Trimap+CF [21]	0.0083	0.0049	0.0035	0.080
Trimap+DIM [32]	**0.0045**	**0.0017**	**0.0013**	**0.0043**
SHM [8]	0.011	0.0078	0.0032	0.011
ours(w/o coarse data)	0.0099	0.0067	0.0029	0.0095
ours(w/o QUN)	0.0076	0.0042	0.0024	0.0072
ours	**0.0058**	**0.0026**	**0.0016**	**0.0054**

Baselines. We select the most typical method from semantic segmentation methods, traditional matting methods, user interactive methods and automatic methods respectively as our baselines. These methods are DeepLab [7], Closed-form matting [21], DIM [3] and SHM [8]. Note that the Closed-form matting and DIM need extra trimap as input. DIM and SHM can only be trained using the fine annotated data. DeepLab and the proposed method are trained using the proposed hybrid annotated dataset.

Performance comparison. In Table 2, we list the quantitative results over 1360 testing images. The semantic segmentation method DeepLab [7] only predict coarse mask and lack fine details (Figure 5(b)), resulting in the worst quantitative metrics. SHM [8] does not perform well as the volume of our high quality training dataset is limited, and fails to predict accurate semantic information for some images (Figure 5(d)). In contrast, the interactive method closeform matting [21] and DIM [32] performs well, benefiting from the input semantic information provided by trimaps. These two methods only need to estimate the uncertain part in trimaps. The proposed method using hybrid training dataset outperforms most methods and is comparable with state-of-the-art methods. DIM [32] is slightly better than the proposed method. Note that the proposed method only take in input images, DIM requires high informative trimaps as extra input. Even though, the visual quality of the proposed method (Figure 5(g)) and DIM (Figure 5(d)) looks very close.

Self-comparisons. Our method can achieve high quality alpha matte estimation by incorporating coarse annotated human data. Coarse annotated data promote the proposed network to estimate semantic information accurately. To verify the importance of the these data, we separately train the same network with fine annotated dataset only. The quantitative results are listed in Table 1. Without using the coarse data, the performance is obviously worse. From Figure 5(f) and (g), we can also observe that method trained only with fine annotated data suffers from inaccurate semantic estimation and presents incomplete alpha matte.

The mask quality unification network make it possible for the final matting refinement network to adapt to different kinds of coarse mask input. Without QUN, inputs to the matting refinement network may vary significantly, which is hard to deal with at inference stage. We list the quantitative metrics without QUN being used in Table 1. Both fine and coarse annotated dataset are used in this comparison. The results are obviously worse when QUN is removed. For a better visual comparison, we display the results in Figure 6. The predicted alpha matte is fine if the coarse mask is relatively accurate. When the coarse mask lacks most hair details, the estimated alpha matte is not good. With QUN, the mask quality is unified before feeding into MRN. The estimated alpha matte is more accurate and robust to different kinds of coarse masks.



(a) Input image (b) Predicted coarse mask (c) Estimated alpha matte without QUN (d) Estimated alpha matte with QUN

Figure 6. Self-comparisons. Without quality unification network (QUN), the quality of coarse mask sent to the matting refinement network (MRN) may vary significantly. When the coarse mask is relatively accurate, MRN predicts alpha matte well. When the coarse mask lacks most hair details, the estimated alpha matte is accurate. Equipped with QUN, the mask quality is unified before feeding into MRN. The estimated alpha matte is more consistent against different kinds of coarse masks.

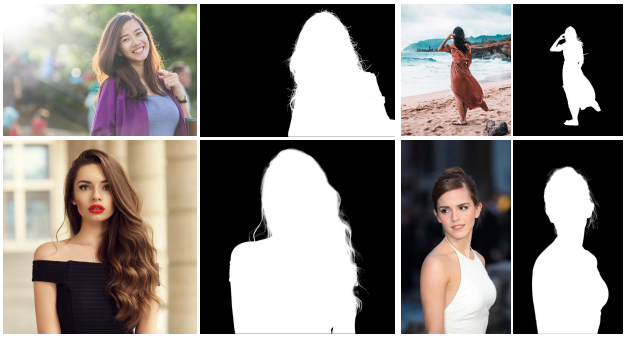
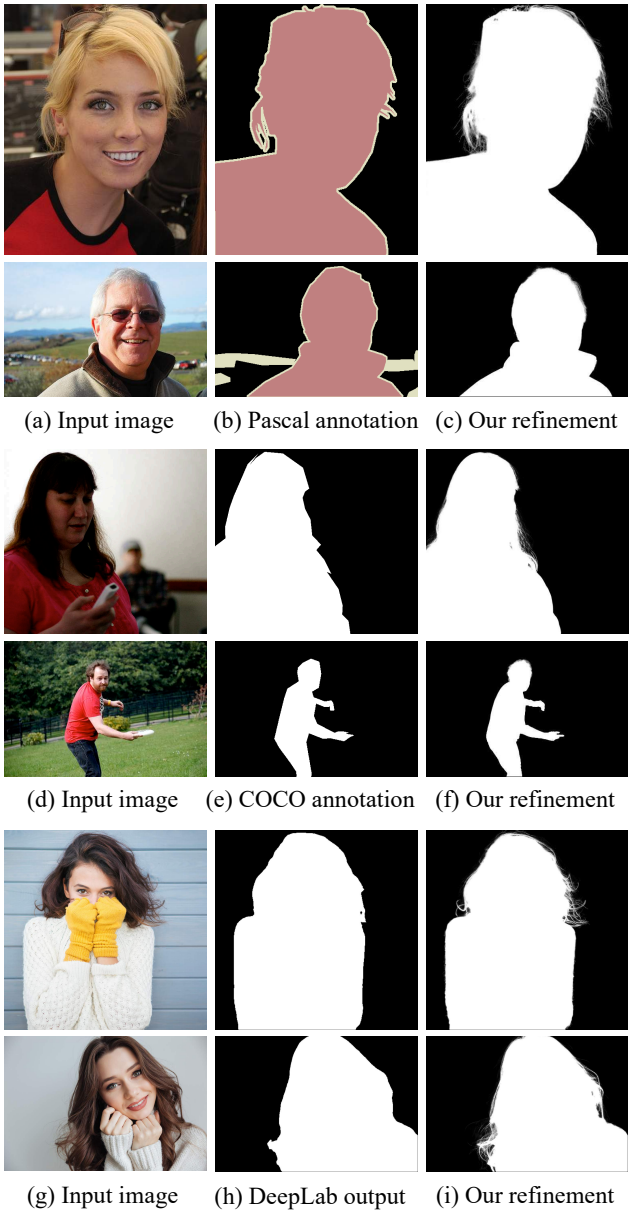


Figure 7. Real image matting results. The collected coarse annotated dataset enriches our dataset significantly and enables the proposed method to capture the semantic information well and predicts accurate alpha matte for different kinds of input images.

5.2. Applying to real images

We further apply the proposed method to real images from the Internet. Matting on real images is challenging as the foreground is smoothly fused with the background. In Figure 7, we display our testing results on real images. Benefiting from the sufficient training on our hybrid dataset, the



(a) Input image (b) Pascal annotation (c) Our refinement

(d) Input image (e) COCO annotation (f) Our refinement

(g) Input image (h) DeepLab output (i) Our refinement

Figure 8. Using the proposed method to refine coarse human mask from public dataset annotations or semantic segmentation methods. Feed the coarse human mask from Pascal (b) or Coco (e) dataset annotation or DeepLab (h) to our quality unification network, and then use the matting refinement network to generate the accurate human alpha matte.

proposed method captures the semantic information very well for different kinds of input images and predicts accurate alpha matte at a detailed level.

6. Applications

The mask prediction network in the proposed method aims to capture coarse semantic information requiring by the subsequent networks. The semantic mask from this network can be coarse or accurate. The following quality unification network will unify the mask quality for the final matting refinement network. Therefore, if the semantic mask is arrange in some way, the proposed method is still able to work seamlessly and generate accurate alpha matte.

Thus we can apply our framework to refine coarse annotated public dataset, such as the PASCAL [13] (Figure 8(a-c)) and COCO dataset [23] (Figure 8(d-f)). The annotated human mask are resized and used as input for our QUN and MRN. Even though the annotations are not accurate, especially the annotations from COCO dataset, the proposed method manages to generate accurate refinement results.

We can also use the proposed method to refine semantic segmentation methods (Figure 8(g-i)). Semantic segmentation methods are usually trained on coarse annotated public dataset, and the output mask is not precise. We feed the coarse mask obtained from DeepLab [7] to our QUN and MRN. The proposed method generates surprisingly good alpha matte. Details that are missing from the coarse mask are well recovered, even for the very detailed hair parts.

7. Conclusion

In this paper, we propose to use coarse annotated data coupled with fine annotated data to enhance the performance of end-to-end semantic human matting. We propose to use MPN to estimate coarse semantic masks using the hybrid annotated dataset, and then use QUN to unify the quality of the coarse masks. The unified mask and the input images are fed into MRN to predict the final alpha matte. The collected coarse annotated dataset enriches our dataset significantly, and makes it possible to generate high quality alpha matte for real images. Experimental results show that the proposed method performs comparably against state-of-the-art methods. In addition, the proposed method can be used for refining coarse annotated public dataset, as well as semantic segmentation methods, which potentially brings a new method to annotate high quality human data with much less effort.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016. 5

[2] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4):72, 2018. 3

[3] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29–37. IEEE, 2017. 3, 7

[4] 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 (TPAMI)*, 39(12):2481–2495, 2017. 1

[5] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 3

[6] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2, 3, 5

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 6, 7, 8

[8] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626. ACM, 2018. 2, 3, 5, 6, 7

[9] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(9):2175–2188, 2013. 3, 5

[10] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *The European Conference on Computer Vision (ECCV)*, pages 626–643. Springer, 2016. 3

[11] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–271. IEEE, 2001. 2

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 8

[14] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *The European Conference on Computer Vision (ECCV)*, pages 204–219. Springer, 2016. 2

[15] 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. 2

[16] 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. 3

[17] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2056. IEEE, 2011. 2

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *The European Conference on Computer Vision (ECCV)*, pages 1–14. Springer, 2010. 3

[19] Jubin Johnson, Ehsan Shahrian Varnousfaderani, Hisham Cholakkal, and Deepu Rajan. Sparse coding for alpha matting. *IEEE Transactions on Image Processing (TIP)*, 25(7):3032–3043, 2016. 2

[20] Levent Karacan, Aykut Erdem, and Erkut Erdem. Image matting with kl-divergence based sparse sampling. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 424–432. IEEE, 2015. 2

[21] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):228–242, 2007. 3, 5, 6, 7

[22] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1699–1712, 2008. 3

[23] 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 *The European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2, 5, 8

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440. IEEE, 2015. 1

[25] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018. 3, 5

[26] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1826–1833. IEEE, 2009. 6

[27] Mark A Ruzon and Carlo Tomasi. Alpha estimation in natural images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 18–25. IEEE, 2000. 2

[28] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *The European Conference on Computer Vision (ECCV)*, pages 92–107. Springer, 2016. 2, 3, 5, 6

[29] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 315–321. ACM, 2004. 3

[30] 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, 2008. 1, 2

[31] Zifeng Wu, Yongzhen Huang, Yinan Yu, Liang Wang, and Tieniu Tan. Early hierarchical contexts learned by convolutional networks for image segmentation. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, pages 1538–1543. IEEE, 2014. 6

[32] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2970–2979. IEEE, 2017. 2, 3, 5, 6, 7

[33] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 2, 3, 5

[34] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305. ACM, 2017. 3