Boosting Semantic Human Matting with Coarse Annotations: Supplementary Materials

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

In this supplementary material we elaborate on implementation details of our network architecture, as well as show extended results for our human matting method and its applications.

1. Network Architecture

The proposed method is composed of three subnetworks. Mask prediction network estimates the coarse semantic mask, quality unification network unifies the predicted coarse mask, and matting refinement network predicts the final alpha matte. The network structure details corresponding to these three networks are listed in Table 1, Table 2 and Table 3 respectively.

2. Dataset with Hybrid Annotations

In Fig. 1, we present more examples to illustrate our hybrid annotated dataset. The coarse annotated dataset is not well labelled, with only rough outlines marked, especially at the human hair regions. In contrast, the fine annotated dataset is annotated at very detailed level.

3. Experimental Results

3.1. More Qualitative Results

In Fig. 2, we demonstrate more results on our human matting testing dataset. The semantic segmentation method DeepLab [1] only predict coarse mask and lack fine details. SHM [2] is trained using high quality data only, and suffers from inaccurate semantic information estimation when data is insufficient. We also train the proposed method with fine annotated dataset only and observe similar phenomenon. The close-form matting [4] takes in extral trimap input and performs well for most images, but presents inaccurate hair details. The proposed method and DIM [6] perform best. The visual quality looks very close. Note that our proposed method only takes in input images, while DIM requires high informative trimaps as extra input.

3.2. More Real Image Results

In Fig. 3, we show more results on real images. Benefiting from the sufficient training on our hybrid dataset, the proposed method captures the semantic information well for different kinds of input images and predicts accurate alpha matte at a detailed level.

4. More Refinement Applications

We display more application results in Fig. 4. We utilize the proposed method to refine coarse human annotations from COCO [5], Pascal [3] dataset, as well as DeepLab [1] segmentations. After refinement, the quality of masks improves significantly with regard to the initial ones.

References

[1] 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

[2] 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. 1

[3] 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. 1, 6

[4] 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. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Mask Prediction Network	
Operation	Output size
input	$192 \times 160 \times 3$
conv+BN+ReLU	$192 \times 160 \times 32$
conv+BN+ReLU+Maxpooling	$96 \times 80 \times 32$
conv+BN+ReLU	$96 \times 80 \times 64$
conv+BN+ReLU+Maxpooling	$48 \times 40 \times 64$
...	...
conv+BN+ReLU	$12 \times 10 \times 256$
conv+BN+ReLU+Maxpooling	$6 \times 5 \times 256$
conv transpose+BN+ReLU	$12 \times 10 \times 256$
conv+BN+ReLU	$12 \times 10 \times 256$
conv transpose+BN+ReLU	$24 \times 20 \times 256$
conv+BN+ReLU	$24 \times 20 \times 128$
...	...
conv transpose+BN+ReLU	$192 \times 160 \times 32$
conv+BN+ReLU	$192 \times 160 \times 32$
conv	$192 \times 160 \times 1$

Table 1: Network details of mask prediction network.

Quality Unification Network	
Operation	Output size
input	$192 \times 160 \times 4$
conv+BN+ReLU	$192 \times 160 \times 32$
conv+BN+ReLU+Maxpooling	$96 \times 80 \times 32$
conv+BN+ReLU	$96 \times 80 \times 64$
conv+BN+ReLU+Maxpooling	$48 \times 40 \times 64$
Residual block	$48 \times 40 \times 64$
Residual block	$48 \times 40 \times 64$
Residual block	$48 \times 40 \times 64$
conv transpose+BN+ReLU	$96 \times 80 \times 64$
conv+BN+ReLU	$96 \times 80 \times 32$
conv transpose+BN+ReLU	$192 \times 160 \times 32$
conv+BN+ReLU	$192 \times 160 \times 32$
conv	$192 \times 160 \times 1$

Table 2: Network details of quality unification network.

Zitnick. Microsoft coco: Common objects in context. In *The European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1, 6

[6] 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. 1

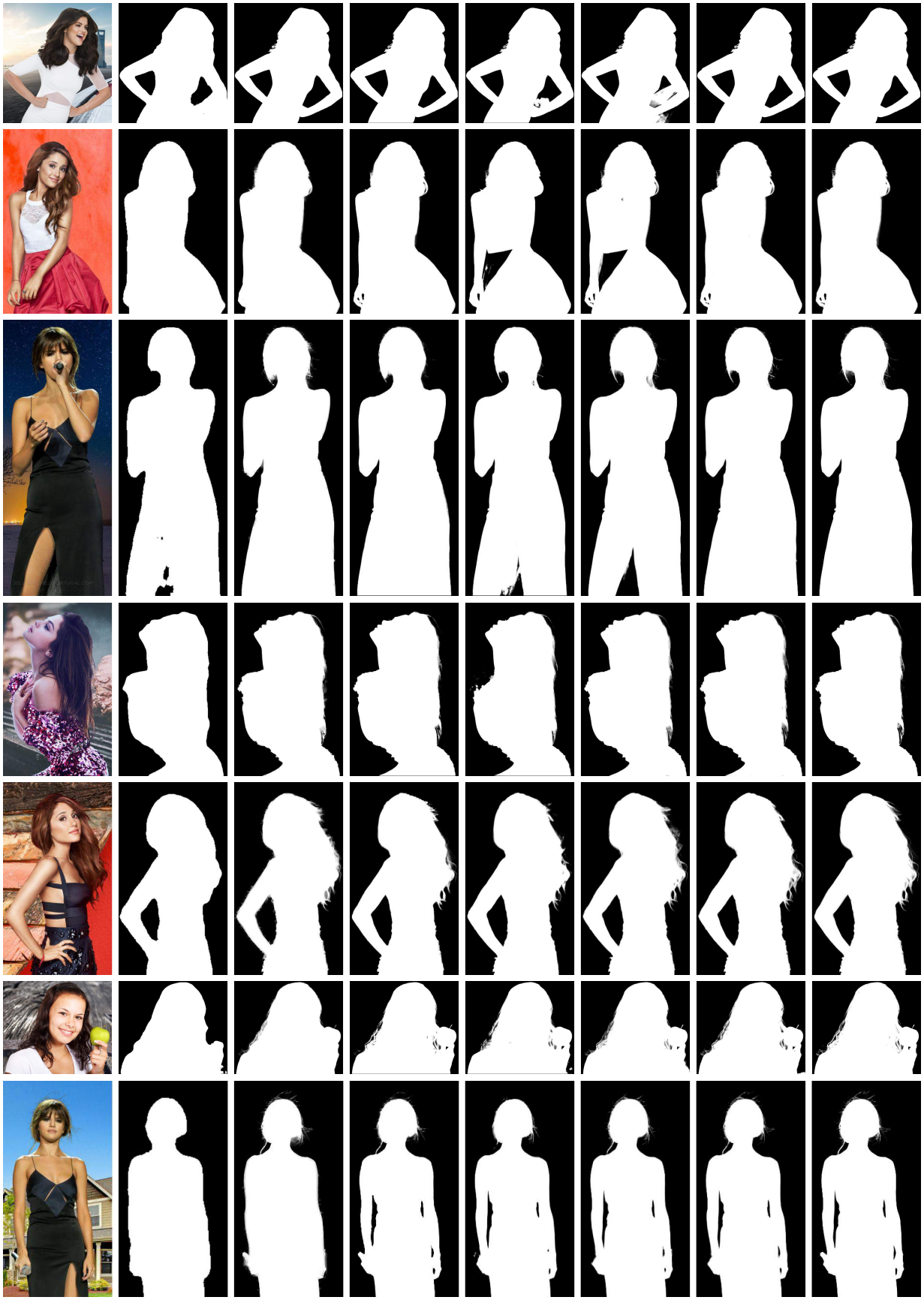
Matting Refinement Network	
Operation	Output size
input	$768 \times 640 \times 3$
conv+ReLU	$768 \times 640 \times 32$
conv+ReLU+Maxpooling	$384 \times 320 \times 64$
conv+ReLU	$384 \times 320 \times 64$
conv+ReLU+Maxpooling	$192 \times 160 \times 128$
coarse alpha+conv+ReLU+concat	$192 \times 160 \times 192$
conv+ReLU	$192 \times 160 \times 256$
conv+ReLU+Maxpooling	$96 \times 80 \times 256$
...	...
conv+ReLU	$48 \times 40 \times 256$
conv+ReLU+Maxpooling	$24 \times 20 \times 256$
conv transpose+ReLU	$48 \times 40 \times 256$
conv+ReLU	$48 \times 40 \times 256$
conv transpose+ReLU	$96 \times 80 \times 256$
conv+ReLU	$96 \times 80 \times 128$
...	...
conv transpose+ReLU	$768 \times 640 \times 32$
conv+ReLU	$768 \times 640 \times 32$
conv	$768 \times 640 \times 4$

Table 3: Network details of matting refinement network.

(a) Coarse annotated dataset

(b) Fine annotated dataset

Figure 1: More images of our hybrid annotated dataset.



(a) Image (b) Deeplab (c) Close-form (d) DIM (e) SHM (g) Ours(fine data) (h) Ours (i) Alpha-GT

Figure 2: More visual comparison results on testing dataset (Best viewed in PDF with zoom).

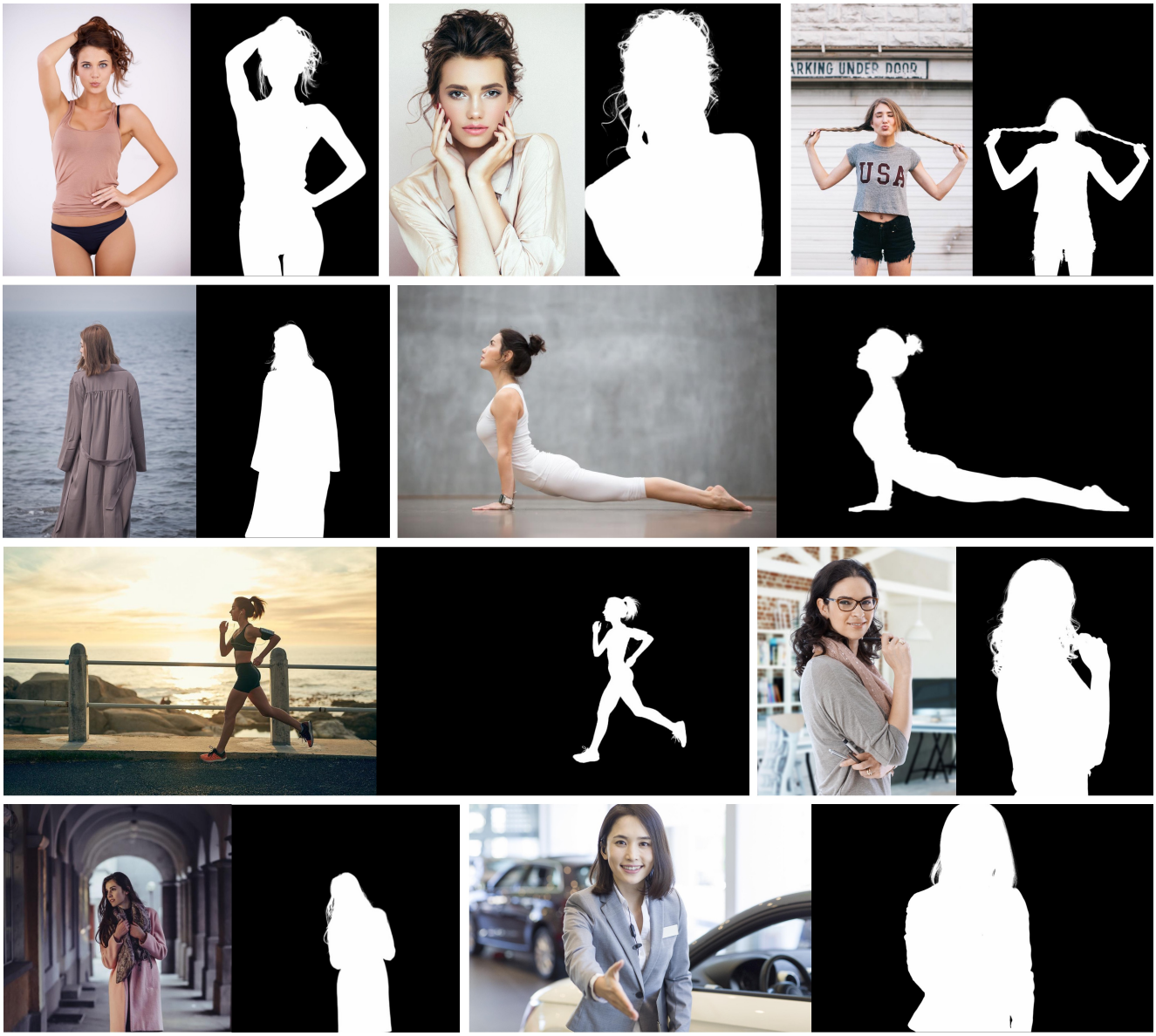


Figure 3: More results on real images (Best viewed in PDF with zoom).

Figure 4: Using the proposed method to refine coarse human mask from COCO [5] dataset annotations, Pascal [3] dataset annotations and DeepLab [1] output (Best viewed in PDF with zoom).