High Quality Segmentation for Ultra High-resolution Images
Supplementary Material

Tiancheng Shen[1] Yuechen Zhang[1] Lu Qi[1] Jason Kuen[2]
Xingyu Xie[3] Jianlong Wu[4] Zhe Lin[2] Jiaya Jia[1,5]
[1]The Chinese University of Hong Kong [2]Adobe Research [3]Peking University
[4]Shandong University [5]SmartMore

In the supplementary material, we first illustrate some technical details about the the weights in Eq. (3), the position information P, and some experiment issues. Then, we proof the Eq. (6) in the paper. Third, experiments on different resolutions of latent feature, pure model capacity, and refining transformer-based segmentation results help understand CRM. Finally, additional visualizations show CRM's refinement performance.

1. Technical details

Owing to space limitations, we do not explain some details on paper. Here, we will illustrate the weights in Eq. (3) and the position information P in detail.

For the weights $\mathrm{w}_{z_k}, k \in \{1,2,3,4\}$ in Eq. (3), we present its details in Fig. 1. After finding the red supporting points $z_k, k \in \{1,2,3,4\}$ of blue queried point x, we calculate the area value a_k between z_k and x. Then swap the symmetric area values a_k about the x point to be weights w_k. The final prediction is the weighted average of the predictions of the supporting points.

For position information P, in the paper, it consists of the refinement target position C_{t}, the relative target coordinate offset C_{r}, and the ratio r between feature and target [3]. We also normalize the feature map coordinate C_{f} and the refinement target coordinate C_{t} to align. In Fig. 2, all components are drawn from left to right, top to down in a simple example. The upper row is the normalized C_{f}, normalized C_{t}, and the relative offset C_{r} between them. Each item of C_{r} is the offset vector of blue point on C_{t} from corresponding red supporting point on C_{f}. The offset is showed in "details of C_{r}" in the lower row. And the way to find the red points of blue point is illustrated in Sec. 3.3. What's more, the rightest part of lower row is a more complex but common example of C_{r}.

The coarse masks, from FCN [11], Deeplavbv3+ [2], RefineNet [10], and PSPNet [15], are all trained on Pascal VOC [8]. And we utilize MaskFormer [4] and SegFormer [12]'s pretrained weights on ADE20K from their Github repositories.

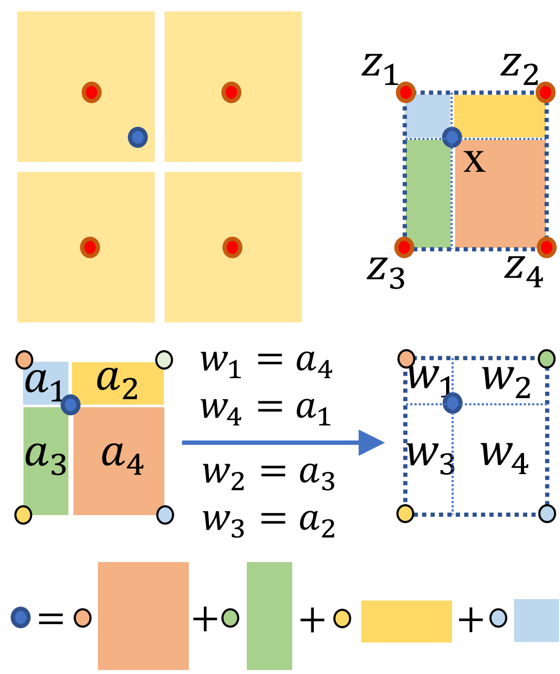

Figure 1. Details of the w_{z_k} in Eq. (3).

In quantitative comparison, we train the SegFix [14] on the same mask perturbing dataset and keep other settings consistent. We use MGMatting pretrained on RWP [13]. To erase the performance degradation caused by the mask-insensitive matting setting, we update cases that have at least 0.80 IoU with coarse inputs after inference.

2. Proof of Eq. (6)

For convenience, we suppose that each entry of the matrix $A \in \mathbb{R}^{2\times m}$ is sampled from $\mathcal{N}(0, 2/m)$. We now define the fixed feature space $\mathbb{F} \subset \mathbb{R}^m$:

$$-1 \le f_i^\top f_j \le 1, \quad \|f_i\|_2 = \|f_j\|_2 = 1, \quad \forall f_i, f_j \in \mathbb{F}.$$

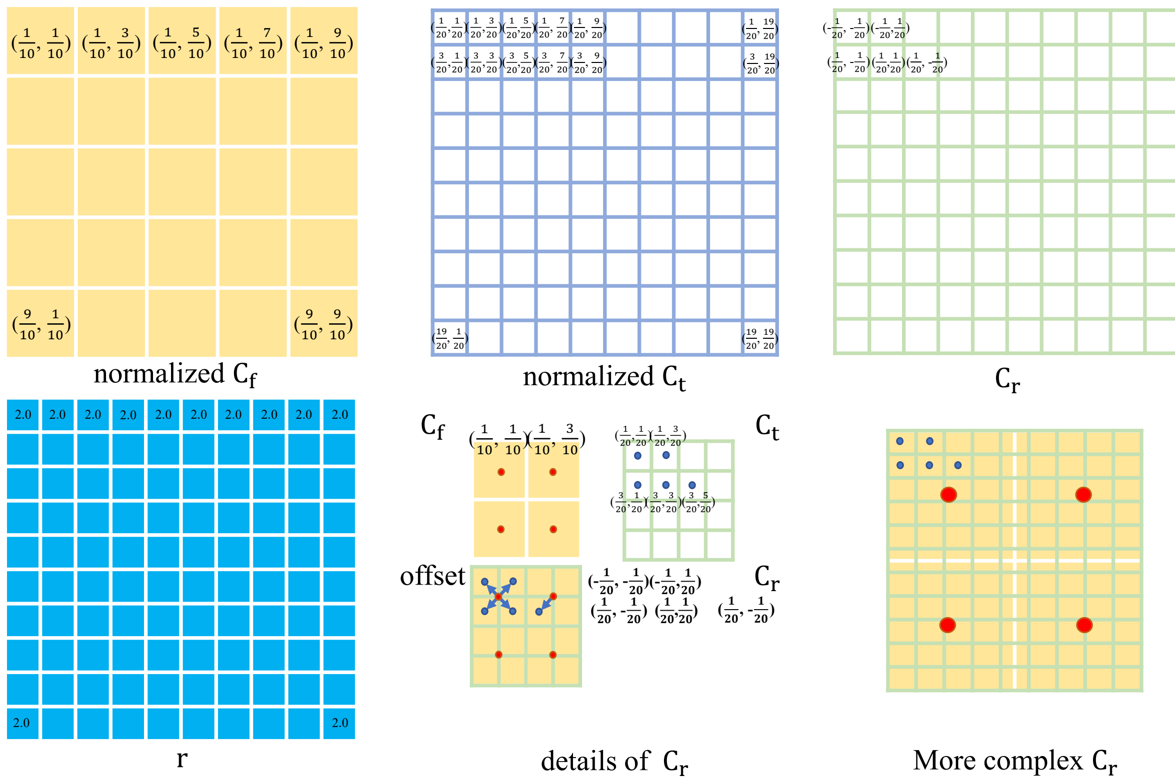

Figure 2. A simple example of position information.

Show that

$$\dim(\mathrm{conv}\,(\phi(Af)) \geq \dim(\phi(Af)), \quad f \in \mathbb{F},$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the ReLU activation, $\mathrm{conv}(\cdot)$ is the convex hull of a bounded set, and $\dim$ is the covering number of a compact set[1].

Proof. Before proving the main results, we first present two useful lemmas.

Lemma 1. *Let $x_1,\ x_2 \in \mathbb{R}^m$, $\|x_1\|_2 = \|x_2\|_2 = 1$, and $x_1^\top x_2 = z$. We have*

$$\mathbb{E}_w\big(\phi(w^\top x_1)\phi(w^\top x_2)\big) = \frac{\sqrt{1-z^2} + z(\pi - \arccos z)}{\pi},$$

where $w \sim \mathcal{N}(0, 2)$ and $\phi(\cdot)$ is the ReLU function.

This lemma is a direct corollary of the results in [7] (see Table 1 therein). We then present the norm preserving property [1].

Lemma 2. *If $A_{i,j} \sim \mathcal{N}(0, 2/m)$ is the random Gaussian matrix and $\phi(\cdot)$ is the ReLU function, then for fixed feature $x \in \mathbb{R}^m$:*

$$\mathbb{P}_A(\|\phi(Ax)\|_2 \in (1 \pm \epsilon)\|x\|_2) \geq 1 - \exp\{-\epsilon^2 m/100\}.$$

First of all, for any $f, f_1, f_2 \in \mathbb{F}$, it is easy to show that the random variable $\big(\phi(w^\top f_1)\phi(w^\top f_2)\big)$ is sub-Exponential since $\big(\phi(w^\top f)\big)$ is sub-Gaussian. Hence, with probability at least $(1 - 2\exp\{-\Omega(m\epsilon)\})$, we have:

$$\big|\phi(Af_1)^\top \phi(Af_2) - \mathbb{E}_w\big(\phi(w^\top f_1)\phi(w^\top x_2)\big)\big| \leq \epsilon.$$

Note that the function $\frac{\sqrt{1-z^2}+z(\pi - \arccos z)}{\pi} : [-1, 1] \to [0, 1]$ is bijective and maps the values from $[-1, 1] \to [0, 1]$. Namely, with probability at least $(1 - 2\exp\{-\Omega(m\epsilon)\})$:

$$0 \leq \phi(Af_1)^\top \phi(Af_2) \leq 1 + \epsilon.$$

Combine with Lemma 2, we have:

$$1 - \epsilon \leq \|\phi(Af)\|_2 \leq 1 + \epsilon.$$

Therefore, with probability $\big(1 - 2\exp\{-\Omega(\epsilon^2 m)\}\big)$, $\phi(Af), \forall f \in \mathcal{F}$ belongs to the *banded* area of a positive half axis semicircle in $\mathbb{R}^2$ with the width being ϵ. Then the ϵ-covering number[2] of $\phi(Af)$ is:

$$\dim\left(\phi(Af)\right) = \Theta(\frac{1}{\epsilon}), \quad f \in \mathbb{F}.$$

[1]Covering number can be seen as a finer measure of the dimensions of a compact set.

[2]The smallest possible cardinality of an ϵ-net of a given set.

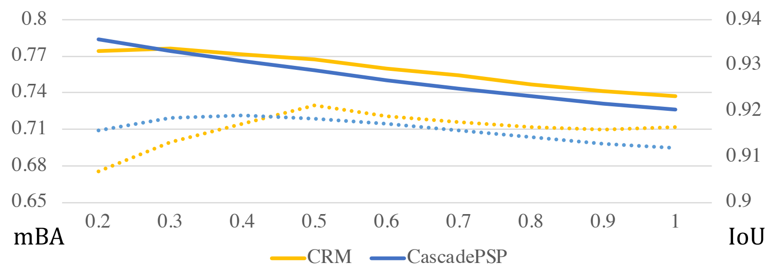

Figure 3. Performance changing with resolution in single-forward inference. The dashed lines refer to the mean Boundary Accuracy (mBA) and the solid for Intersection over Union (IoU). Each color represents one method.

Res. of F_{latent}	IoU	mBA
conv2_x	94.18	76.09
conv3_x	94.02	75.84
conv4_x	93.50	72.96

Table 1. The influence of different resolutions of latent feature. Res. denotes resolution.

In contrast, in the worst case, it is impossible to cover the set $\phi(Af)$ with only $\Theta(\frac{1}{\epsilon})$ ϵ-balls and we have:

$$\dim\left(\mathrm{conv}(\phi(Af))\right) = \Theta(\frac{1}{\epsilon^2}), \quad f \in \mathbb{F}.$$

We now finish the proof. □

Discussion In general, the representation ability of a ball and a sphere is almost the same in high dimensional space. However, in low dimensional space, it is a completely different scene. This is also the reason that why some common methods which work well for image data are needed to be re-designed for 3D point cloud. It is more meaningful to expand the representation ability of the model in low dimensional space than to only increase the parameters of the model.

3. Quantitative results on different resolutions of latent feature

We additionally conduct the experiment on the different resolutions of feature fusion. In detail, using ResNet-50 [9] without conv5_x as the backbone, we choose to fuse the feature from conv2_x, conv3_x, and conv4_x. The fused feature is the concatenation of the resized feature from these three layers. The resolution of fused feature is an important parameter, which influences F_{latent} and the r in position feature P. From the Tab. 1, we find larger resolution of the fused feature is better for performance. It may be because segmenting ultra high-resolution images needs finer feature, corresponding to higher resolution.

IoU/mBA	w/o CRM	w CRM
MaskFormer [4]	82.38/62.52	85.24/76.17
SegFormer [12]	74.87/56.63	80.25/74.52

Table 2. Refine transformer-based segmentation results.

4. Results of pure model with single forward

To verify performance of the pure model, we compare our method and CascadePSP [5] in single forward for different resolution inputs. In detail, CRM directly inferences on different resolution inputs, and CascadePSP [5] refines the inputs patch by patch. The inputs are from PSPNet [15] on the BIG dataset. Fig. 3 shows the performance changing with the image resolution through a single forward of model. CRM is not good at very low resolution. However, when the resolution increases, the performance of CRM is better.

5. Refine transformer-based segmentation

Since more and more transformer-based segmentation methods emerge, we also apply our CRM on their segmentation results. From the Tab. 2, we find CRM can also increase the performance of transformer-based methods on the BIG dataset.

In comparison, [4, 12] only release the pretrained weights on Cityscape [6] and ADE20K [16], but the BIG [5]'s annotation follows Pascal VOC's guideline. We choose the union of ADE20K and Pascal VOC's categories to evaluate.

6. Additional visualization

We also provide additional visualization results of CRM in Figs. 4, 5, 6 and 7. Masked areas are put on a green background for easy distinguishing. Although the images are resized, their original resolution is very large (2K$\sim$6K), where the details are more obvious. These visualizations are better viewed on the screen.

Figure 4. Visualization of the refinement on FCN [11]'s output. **Better viewed on the screen.**

Figure 5. Visualization of the refinement on DeepLabV3+ [2]'s output. **Better viewed on the screen.**

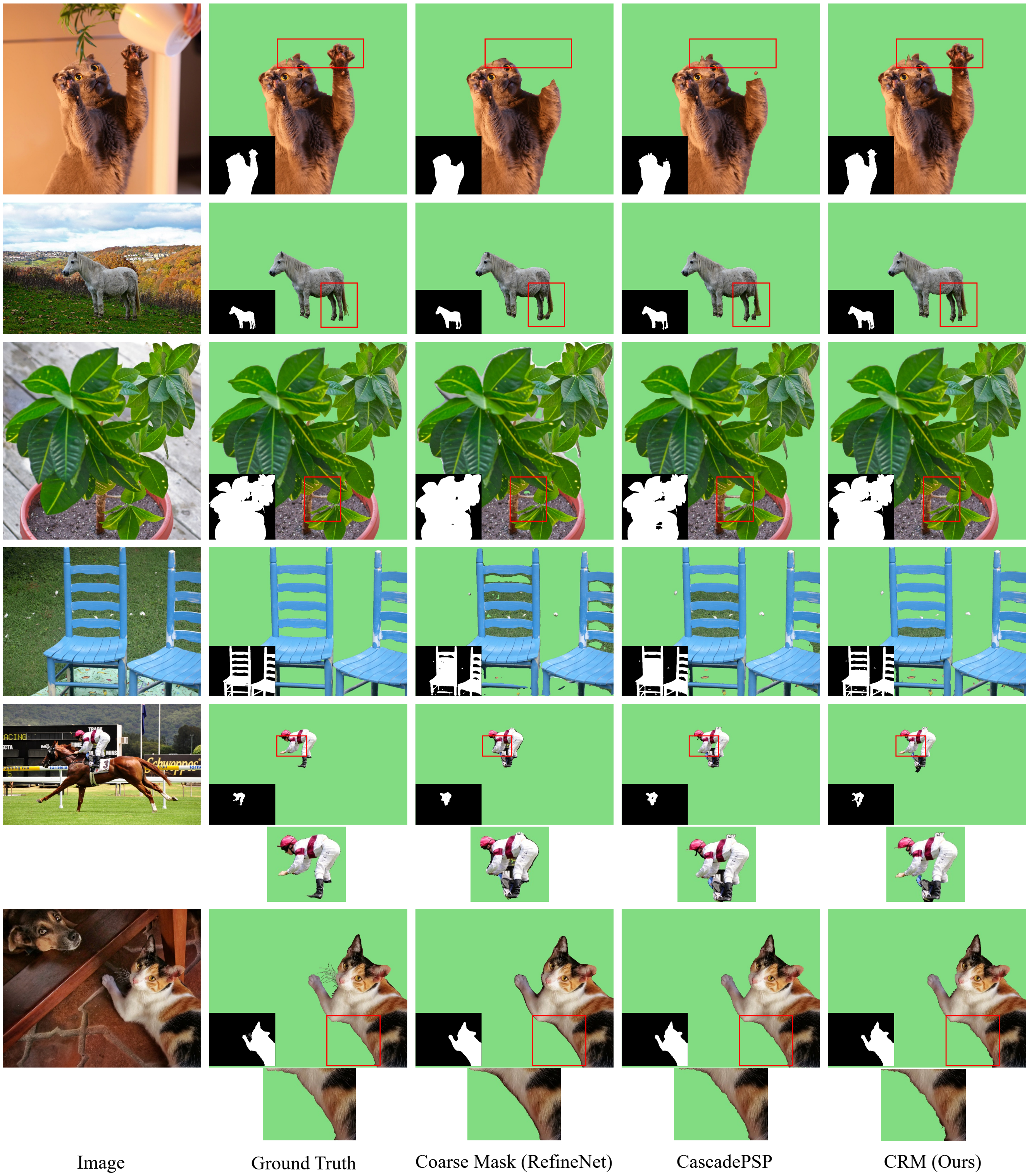

Figure 6. Visualization of the refinement on RefineNet [10]'s output. **Better viewed on the screen.**

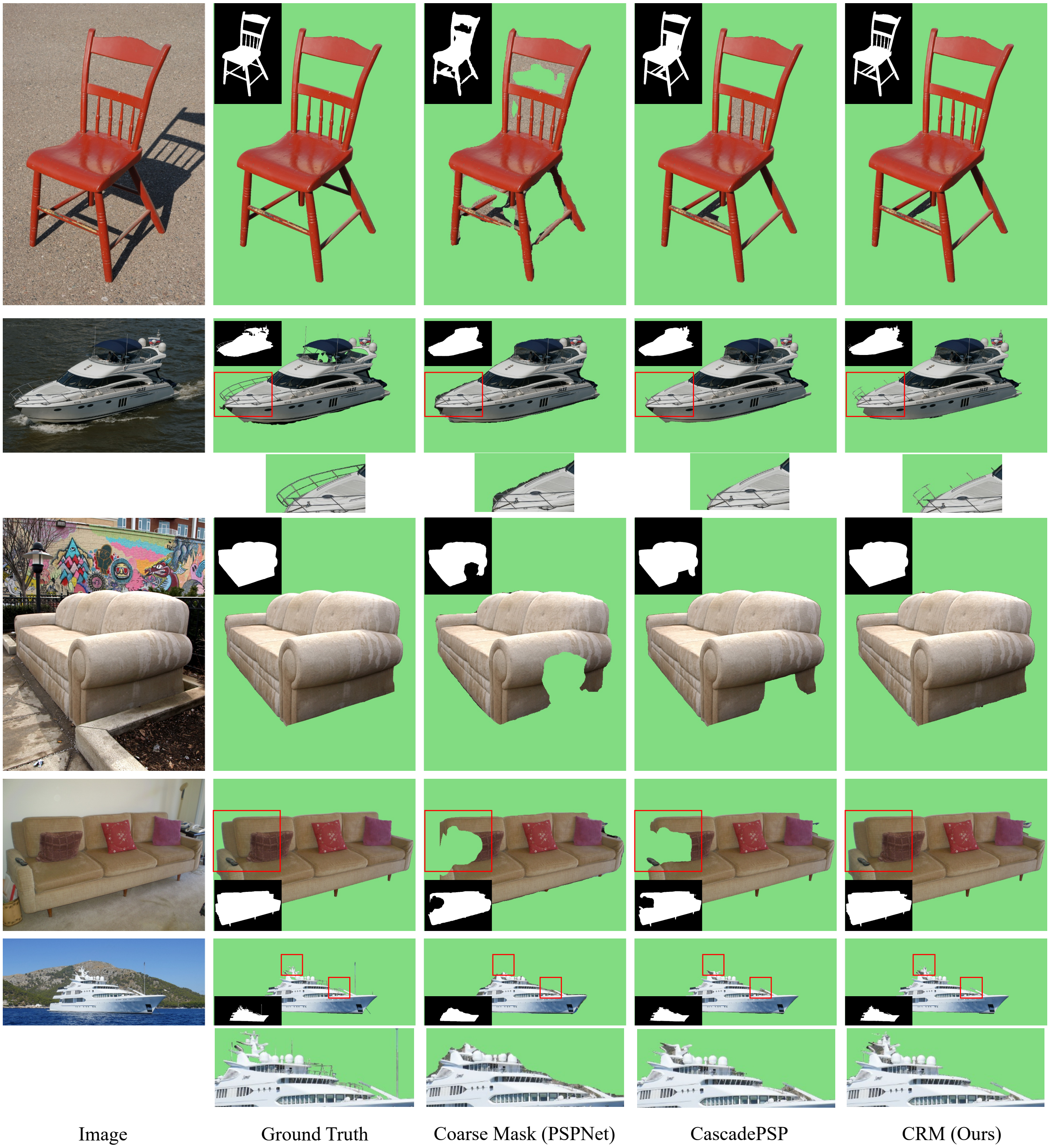

Figure 7. Visualization of the refinement on PSPNet [15]'s output. **Better viewed on the screen.**

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 1, 5

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 1

[4] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 1, 3

[5] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, pages 8890–8899, 2020. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 3

[7] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *NeurIPS*, pages 2253–2261, 2016. 2

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[10] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 1, 6

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 4

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 1, 3

[13] 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, June 2021. 1

[14] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, pages 489–506, 2020. 1

[15] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1, 3, 7

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 3