

# High Quality Segmentation for Ultra High-resolution Images

Tiancheng Shen<sup>1</sup> Yuechen Zhang<sup>1</sup> Lu Qi<sup>1</sup> Jason Kuen<sup>2</sup>  
 Xingyu Xie<sup>3</sup> Jianlong Wu<sup>4</sup> Zhe Lin<sup>2</sup> Jiaya Jia<sup>1,5</sup>

<sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>Adobe Research <sup>3</sup>Peking University  
<sup>4</sup>Shandong University <sup>5</sup>SmartMore

## Abstract

To segment 4K or 6K ultra high-resolution images needs extra computation consideration in image segmentation. Common strategies, such as down-sampling, patch cropping, and cascade model, cannot address well the balance issue between accuracy and computation cost. Motivated by the fact that humans distinguish among objects continuously from coarse to precise levels, we propose the Continuous Refinement Model (CRM) for the ultra high-resolution segmentation refinement task. CRM continuously aligns the feature map with the refinement target and aggregates features to reconstruct these image details. Besides, our CRM shows its significant generalization ability to fill the resolution gap between low-resolution training images and ultra high-resolution testing ones. We present quantitative performance evaluation and visualization to show that our proposed method is fast and effective on image segmentation refinement. Code is available at <https://github.com/dvlab-research/Entity/tree/main/CRM>.

## 1. Introduction

With the rapid development of camera and display equipment, the resolution of images is getting higher and higher, where 4K and 6K resolutions become common. It gives different chances in portrait photo post-processing, industrial defect detection, medical diagnose, etc. However, ultra high-resolution images also bring challenges to the classical image segmentation methods. First, the significant number of input pixels is computationally expensive and GPU memory-hungry. Second, most existing methods up-sample the final prediction for 4 to 8 times through interpolation [5, 49, 52, 55, 56], without building fine-grained details on output masks.

Previous segmentation refinement methods include those of [18, 22, 27, 53]. They still target at images with 1K~2K resolutions. Work of [9, 47] handles ultra high-resolution refinement based on low-resolution masks generated from classic segmentation algorithms. They utilize cascade-

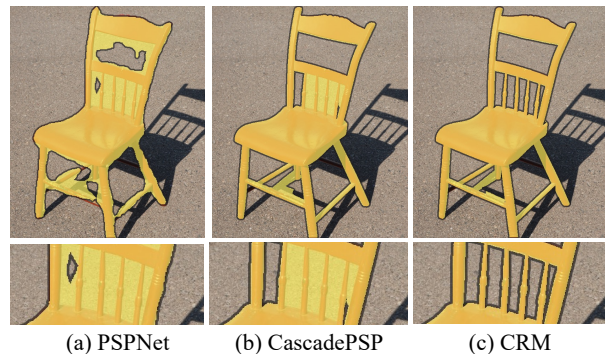


Figure 1. Coarse mask refinement results. (a) Coarse mask from PSP [55], (b) refined mask of state-of-the-art [9], and (c) refined mask of our proposed CRM. The image is from BIG (2K~6K res).

scheme in decoder to upsample intermediate refinement results in several resolution stages until reaching the target resolution. They are still time-consuming due to working in discrete style on pre-defined resolution stages of decoder. We instead consider continuity to make the decoding more efficient and more friendly to the learning of up-sampling resolution. We propose the Continuous Refinement Model (CRM) to exploit continuity.

The coarse mask is from low-resolution segmentation. In order to expand it further, the problem is similar to a classical super-resolution (SR) task. Other than classical SR methods, constructing continuous local representation is proposed [7]. We note that utilizing implicit function [31] to handle high-resolution segmentation refinement is not trivial. First, the resolution of the training image in our task is around 500, while the training image for SR is with 2K resolution. The training strategy to down-sample the input to SR would make our input mask tiny and meaningless. Second, more multi-level semantic features are needed compared with super-resolution configuration. Third, there exists a resolution gap between training on low-resolution and testing on ultra high-resolution. Therefore, this task needs specific designs.

To realize the continuity in ultra high-resolution segmen-

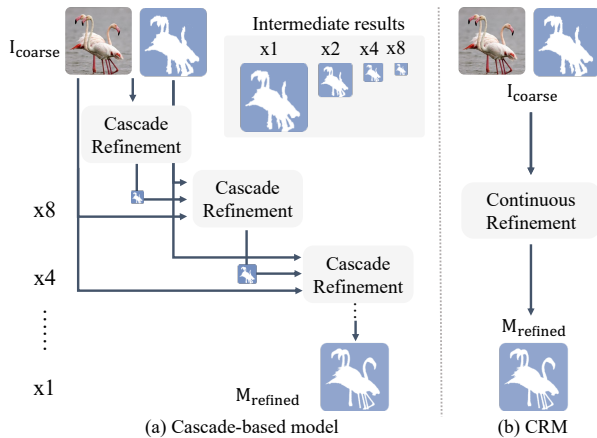


Figure 2. Structure difference between (a) Cascade-based decoder in model [9] and (b) our CRM. We can see CRM is much simpler, which is the base of our speed advantage.

tation refinement, we first propose Continuous Alignment Module (CAM) to align the feature and refinement target continuously (different from utilizing the cascade scheme in decoder). In CAM, the coordinates of feature and refinement target are transferred into a continuous space. We then align position and feature based on the continuous coordinate. An implicit function combines position information and aligned latent image feature to predict the segmentation label for the queried pixel on images. Here, the pixel-wise implicit function models the relationship between continuous position and prediction and realizes image-aware refinement by latent feature. Overall, this design is simpler and lighter than the cascade-based decoder, but generates more precise refinement mask as Fig. 1.

In addition, there is a resolution gap between low-resolution training images and ultra high-resolution testing ones. In cascade-decoder-based methods [9, 47], convolution always covers a fixed size neighbor patch under the training resolution, which reduces its generalization to other testing resolutions. In contrast, implicit function in CRM is in pixel-wise extracted feature without this bias. Also, in our multi-resolution inference strategy, low-resolution input is inferred first. Then we increase the input resolution to generate more details in the refined mask. Working with a multi-resolution inference strategy, CRM realizes stronger generalization ability than previous methods [9] with much faster inference speed.

Our main contribution is the following.

- We propose a general Continuous Refinement Model (CRM). It introduces an implicit function that utilizes continuous position information and continuously aligns latent image feature in ultra high-resolution segmentation refinement. Without a cascade-based decoder, we effectively reduce

computation cost and yet reconstruct more details.

- CRM with multi-resolution inference is suitable for using low-resolution training images and ultra high-resolution testing images. Due to the simple design, even with refining from low to high-resolution, the total inference time is less than half of CascadePSP [9].
- In experiments, CRM yields the best segmentation results on ultra high-resolution images. It also helps boost the performance of state-of-the-art panoptic segmentation models without fine-tuning.

## 2. Related Work

### 2.1. Semantic Segmentation

Semantic segmentation is to assign a class label to each pixel for an image. FCN [30] introduces the deep convolution network into semantic segmentation and achieved remarkable progress, and deep convolution networks are the dominant solution in this area. Later work includes PSP-Net [55], DeepLab series methods [2–5], and other outstanding work [13–15, 19, 24, 26, 33, 38, 44–46, 49, 52, 59, 60].

Among these methods, output stride (or down-sample ratio) is one point that cannot be ignored. In most semantic segmentation methods, it is set to  $4\times$  [49, 52] or  $8\times$  [5, 55, 56], which reduces precision. Directly interpolating prediction logits to target-size results in jagged edge and fewer details. In contrast, our proposed CRM continuously aligns features to arbitrary target refinement resolution, which is more natural for visual instinct and friendly to detail reconstruction.

### 2.2. Segmentation Refinement

The segmentation refinement technique is proposed to improve the quality of image segmentation. In this track, recent work can be categorized into two classes according to the image size of high-resolution (1K~2K) or ultra high-resolution (4K~6K).

For the refinement techniques of images around 1K resolution, they greatly improve the segmentation quality. The remaining drawbacks include graphical models adhering to low-level color boundaries [2, 57], propagation-based approaches facing computational and memory constraints [29], and large models prone to overfitting while shallow refinement networks with limited refinement capability [18, 22, 27].

This paper focuses on ultra high-resolution image segmentation refinement on, e.g., 4K images. Due to this resolution setting, the above methods would face resource and effectiveness difficulties. Cascade-in-decoder methods [6, 9] achieve the state-of-the-art refinement performance on ultra high-resolution images due to its cascade

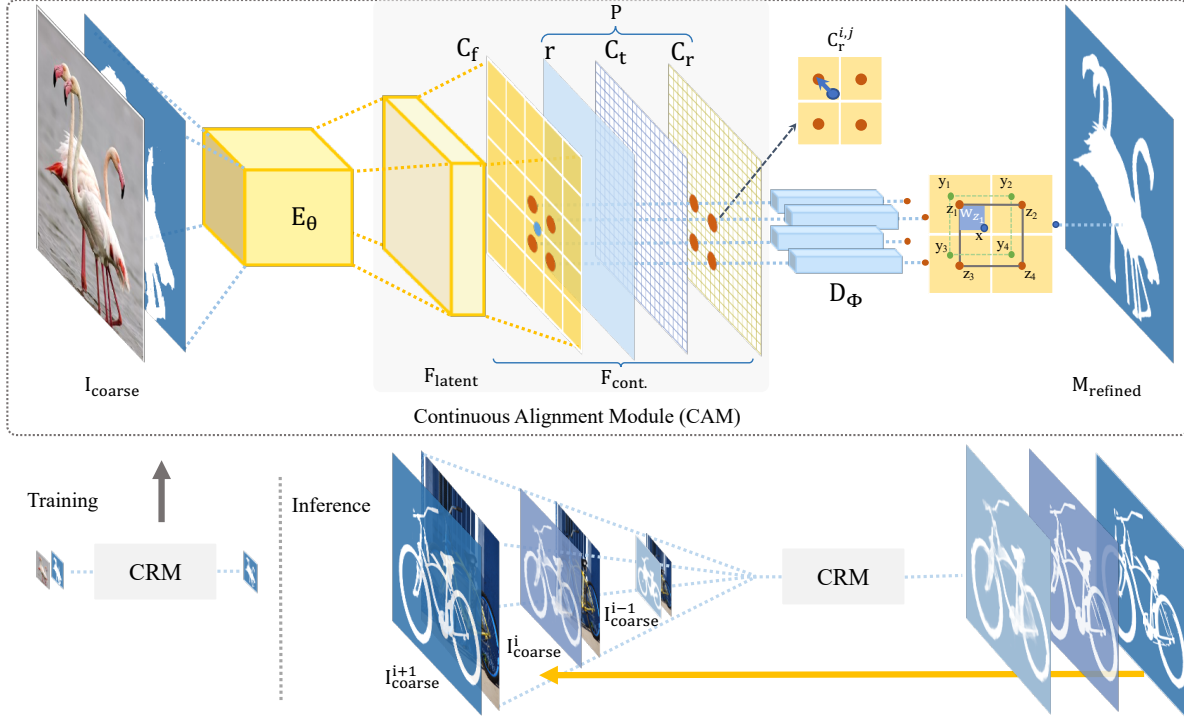


Figure 3. The general framework of CRM. The upper part is the structure of the model. The lower part is the training and testing process of CRM. From the lower part, we can also see the resolution gap between low-resolution training and high-resolution testing.

network structure [16, 37, 40, 43, 54] and a global-local patch-based refining pipeline.

However, the heavy cascade structure in the decoder needs down-sampling and cropping patches during inference, which increases cost, loses details, and destroys global context. To solve these problems in ultra high-resolution image segmentation, we propose CRM. Through CAM in CRM, we continuously align the feature map with refinement target simply and elegantly. The structure difference between cascade-based model [9] and our CRM is presented in Fig. 2.

### 2.3. Implicit Function for Representation

In the beginning, implicit function is designed to represent an object or a scene in a neural network (by usually multi-layer perceptron), which maps continuous coordinates and the features around to the label at the coordinates. For example, NeRF [31] maps the 3D coordinate and 2D view angle into RGB and transparency of certain positions from specific views. PixelNeRF [50] introduces an architecture that conditions a NeRF [31] on image input in a fully convolutional manner, which realizes scene-aware modeling. In addition, its “relative camera poses” idea also inspires research to use relative position information.

As another extension, Semantic-NeRF [58] extends neural radiance fields to encode semantics with appearance

and geometry jointly. The intrinsic multi-view consistency and implicit function’s smoothness benefit segmentation by enabling efficient propagation on sparse and noisy labels. There are works utilizing implicit functions in 2D image [7, 8, 11, 39, 42]. We use implicit function to continuously upsample feature map to final mask.

## 3. Proposed Method

This section first describes the general framework for the Continuous Refinement Model (CRM), then illustrates the Continuous Alignment Module (CAM) and the following implicit function. Finally, we introduce the corresponding inference strategies to exploit continuity in ultra high-resolution. And for the details of Sec. 3.1 and Sec. 3.2, please refer to the supplementary material.

### 3.1. General Framework

As illustrated in Fig. 3, following the setting of CascadePSP [9], our proposed CRM takes an image  $I \in \mathbb{R}^{3 \times H \times W}$  and a coarse segmentation mask  $M_{\text{coarse}} \in \mathbb{R}^{1 \times H \times W}$  as input. First,  $I$  and  $M_{\text{coarse}}$  are concatenated as  $I_{\text{coarse}} \in \mathbb{R}^{4 \times H \times W}$  and are represented as latent embedding  $F_{\text{latent}} \in \mathbb{R}^{C \times h \times w}$  by an encoder  $E_{\theta}$  as Eq. (1), where  $\theta$  denotes the parameters.

$$F_{\text{latent}} = E_{\theta}(I_{\text{coarse}}). \quad (1)$$

Second,  $F_{\text{latent}}$  and position information  $P$  are continuously aligned to be the target size feature  $F_{\text{cont.}} \in \mathbb{R}^{(C+6) \times H \times W}$  through CAM without explicit up-sampling as Eq. (2), where  $[\cdot, \cdot]$  denoted concatenation.

$$F_{\text{cont.}} = \text{CAM}([P, F_{\text{latent}}]). \quad (2)$$

Finally,  $F_{\text{cont.}}$  passes an implicit-function-based decoder [7]  $D_\phi$  and feature aggregation step, making refined mask  $M_{\text{refined}}$  generated as below:

$$M_{\text{refined}}(x) = \sum_{z_k \in N(x)} \frac{w_{z_k}}{\sum w_{z_k}} D_\phi(F_{\text{cont.}}(z_k)), \quad (3)$$

where  $x$  is an aligned point,  $N(x)$  denotes the set of  $x$ 's supporting points  $z_k, k \in \{1, 2, 3, 4\}$ ,  $w_{z_k}$  is the aggregation weights (swap the area value of the box between  $x$  and  $z_k \in N(x)$  symmetrically with  $x$  as the center), and  $F_{\text{cont.}}(z_k)$  is the feature vector of  $z_k$  on  $F_{\text{cont.}}$ .

### 3.2. Continuous Alignment Module

**Motivation** After passing the image encoder, the size of the encoded feature is smaller than the refinement target. Intermediate feature or refined results need to be up-sampled to later stages progressively. In previous work [9, 47] on ultra high-resolution image segmentation, cascade scheme seems an indispensable part of the decoder. Although novel designs alleviate information damage after up-sampling in a specific resolution, the overall process is hard to restore more details.

We note that the discrete manner in cascade-based decoder with predefined up-sampling ratios can be regarded as constraints to up-sampling, limiting the further improvement and reducing generality. In addition, it increases the complexity of the whole framework, illustrated in Fig. 2. Our proposed Continuous Alignment Module (CAM) utilizes position information and feature alignment to model the continuous deep feature  $F_{\text{cont.}}$ .

**Position Information  $P$**  Referring to NeRF-Series [31, 50, 58], the position information is the essential input to the implicit function. Coordinate of refinement target  $C_t$  is projected to feature map coordinate  $C_f$ . This operation creates continuous coordinates for pixels on different resolution feature maps and various desired inference resolutions, shown in Sec. 3.4.

The absolute coordinate may vary with the image and feature size. To make our CRM universal for images of arbitrary sizes, the  $C_t$  and  $C_f$  are normalized to certain range  $[-1, 1]$ . After projection, the offset between the points on  $C_t$  and their corresponding nearest points on  $C_f$  is denoted as  $C_r$ . In Fig. 3, the  $C_r^{i,j}$  represents the offset (blue arrow) on position  $(i, j)$ . The relative target coordinate offset  $C_r$ , the ratio  $r$  between feature and target [7], and the refinement target position  $C_t$  form the position information  $P$  as

$$P = \{C_r, r, C_t\}. \quad (4)$$

The continuous position information is the basis of continuity in CRM.

**Continuous Feature Alignment** Compared with continuous resolution conversion in SR [7],  $F_{\text{latent}}$  from  $E_\theta$  in the Eq. (1) needs to enhance by fusing global-local information for the segmentation refinement task. For simplicity,  $F_{\text{latent}}$  includes the enhancement. The refinement target position  $C_t$  can also be regarded as a global feature. Then, same as that for the position information, we align each pixel in refinement target to  $F_{\text{latent}}$ . The continuous feature  $F_{\text{cont.}}$  is established by concatenating the position information  $P$  and the aligned  $F_{\text{latent}}$  as shown in Eq. (2).

Therefore, compared with discrete resolution conversion, CAM up-samples feature in a continuous manner. The discrete predefined up-sampling ratios reduce the learning difficulty but constrain the up-sampling process. Our CAM has a greater degree of freedom in this respect, which means a larger space to optimize and higher performance potential. The multi-resolution inference in Sec. 3.4 gives full play to the advantage of continuity of CAM.

### 3.3. Implicit Function in CRM

After CAM, implicit-function  $D_\phi$  takes  $F_{\text{cont.}}$  as input. The reason to utilize implicit function is its impressive ability to process continuous coordinates and reconstructing details [7, 31, 50, 58].

A queried point (blue point on Fig. 3) on target refinement mask could be denoted as  $x(i, j)$ , in which  $(i, j)$  is its unnormalized position. First, we find its neighbor points  $y_k, k \in \{1, 2, 3, 4\}$  (green points on Fig. 3) on target refinement mask, whose positions are  $(i \pm 1, j \pm 1)$ . Next, the nearest points of  $y_k$ , denoted as  $z_k$  (red points on Fig. 3), are selected on the aligned feature map. And  $z_k$  are utilized as the supporting points of  $x$ , represented as  $N(x)$ . We then input  $z_k$ 's feature vector  $F_{\text{cont.}}(z_k)$  to implicit function  $D_\phi$  (a 5-layer MLP mapping the 6+256 channel coordination and feature into 1 channel mask). Finally, we aggregate the implicit function's output. The aggregation weights, i.e., area value  $w_{z_k}$ , are calculated from relative coordinate offsets  $C_r$  in Eq. (3). The aggregated output is the final prediction result on  $(i, j)$ .

**Analysis** It is well-known that the forward process of CNNs (e.g., CascadePSP [9]) and MLPs (e.g., CRM) can be regarded as a sequence of operations built on matrix-vector multiplications and nonlinear activation. At initialization, all the weights are sampled from well-scaled Gaussian. Hence, each layers' feature shares almost the same Euclidean norm with high probability (see Cor. A.10 in [1]). Namely, for some constant  $c$ , with probability at least

$1 - 2 \exp(-c\varepsilon^2 m)$ , we have:

$$\|\phi(AF_{\text{cont.}})\|_2 \in (1 \pm \varepsilon)(\|F_{\text{cont.}}\|_2), \quad (5)$$

where each entry of the matrix  $A \in \mathbb{R}^{d \times m}$  is sampled from  $\mathcal{N}(0, \frac{1}{m})$ ,  $F_{\text{cont.}}$  is the fixed feature (same as  $F_{\text{cont.}}$  in Eq. (2)),  $\varepsilon \in [0, 1]$ ,  $\|\cdot\|_2$  is  $\ell_2$ -norm, and  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  is the ReLU activation.

The norm is almost preserved after going through one layer. However, if we further append one operation of weighted average on  $\phi(AF_{\text{cont.}})$ , things become interesting. The appending weighted average can always help to improve the representation ability of model, i.e.,

$$\dim\left(\sum_{z_k \in N(x)} \frac{w_{z_k}}{\sum w_{z_k}} \phi(AF_{\text{cont.}}(z_k))\right) \geq \dim(\phi(AF_{\text{cont.}})), \quad (6)$$

where  $\dim$  is the dimension of space.

A toy example is that, when  $F_{\text{cont.}}$  is the  $m$ -dimensional sphere  $\mathcal{S}(m)$ ,  $\phi(AF_{\text{cont.}})$  will concentrate around the sphere  $\mathcal{S}(d)$  by the norm-preserving property. However, after combining with the weighted average operator, we can get any points in the  $d$ -dimensional ball  $\mathcal{B}(d)$ . Generally,  $\dim(\mathcal{B}(d)) > \dim(\mathcal{S}(d))$ .

Back to the section, the main difference between CRM and CascadePSP [9] is the decoder part. Take four neighboring points as an example. CRM utilizes MLP and area-based average instead of  $2 \times 2$  convolution. Therefore, the dimension of CRM's feature space is larger. If the four points all belong to the same class, the influence is not very large. Still, for boundary region, where 4 points belonging to different classes, larger feature space always provides more distinguishable feature to classified. From this view, we can give some hints about CRM having stronger boundary region representation and predicting better details.

### 3.4. Training and Inference Strategy

**Training without Cascade** LIIF [7] proposes an elegant solution for SR with the implicit function. It has 2K images as ground truth and generates any low-resolution images as input. However, ultra high-resolution images with precise segmentation annotations are too few to train. In addition, high-resolution training is directly limited by the constraint of GPU memory and batch size.

With these challenges, we follow the training setting of CascadePSP [9] to use low-resolution images in their initial resolution.  $M_{\text{coarse}}$  is generated by morphological perturbations on the provided ground truth mask  $M_{\text{gt}}$ . We design the training loss in a simple way on the final prediction  $M_{\text{refined}}$  without different loss functions on different resolution stages [9]. Our loss term  $L(\theta, \phi)$  is calculated on the refinement target as

$$L(\theta, \phi) = \sum_{i=1}^4 w_i \cdot L_i(M_{\text{refined}}, M_{\text{gt}}), \quad (7)$$

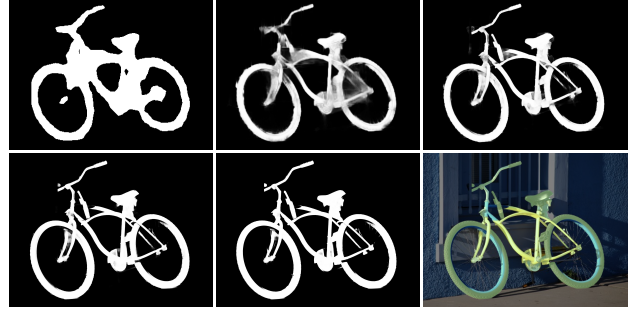


Figure 4. Visualization of refinement steps in our inference strategy. From left to right, top to down:  $M_{\text{coarse}}$ , refined mask  $M_{\text{refined}}^i$ ,  $i \in \{1, 2, 3, 4\}$  (The rescale ratios are 0.125, 0.25, 0.5, and 1.0 here.), and overlay  $M_{\text{refined}}^4$  on the original image.

where  $L_i$ ,  $i \in [1, 2, 3, 4]$  denote cross-entropy loss, L1 loss, L2 loss, and gradient loss, respectively.  $w_i$  are their corresponding weights.  $(\theta, \phi)$  are the parameters of encoder  $E_\theta$  and decoder  $D_\phi$ .  $M_{\text{gt}}$  denotes the ground truth mask.

Although we train on the low resolution, multi-resolution inference strategy exploits the continuity potential and narrows the training and testing resolution gap.

**Inference Strategy** For the resolution gap between low-resolution in training (300~1K) and ultra high-resolution (2K~6K) in testing, we propose multi-resolution inference to exploit CRM's continuous  $P$  and aligned  $F_{\text{cont.}}$  fully. The lower part of Fig. 3 shows the resolution contrast. Due to the continuous property of CAM, for one image, we can generate outputs of the same target ultra high-resolution  $M_{\text{refined}}^i$  from multi-resolution input  $R^i(I_{\text{coarse}}^i)$ .

In the beginning, inference is around the resolution of training images, and gradually increases input's resolution along the continuous ratio axis  $Rs$  (with infinite different rescale ratios) as illustrated in Fig. 3. In particular, we concatenate the original ultra high-resolution image  $I$  and the coarse mask  $M_{\text{coarse}}$  (initial stage) or refined mask  $M_{\text{refined}}^{i-1}$  in previous stage. We rescale it on rescale ratio by  $R^i \in Rs$  to be  $I_{\text{coarse}}^i$ . After refinement,  $M_{\text{refined}}^i$  is generated and used as  $M_{\text{coarse}}^{i+1}$  for the next rescale ratio stage. The progressive processing is illustrated as Eqs. (8) to (10):

$$I_{\text{coarse}}^0 = [I, M_{\text{coarse}}^0], \quad (8)$$

$$M_{\text{refined}}^i = D_\phi(\text{CAM}(E_\theta(R^i(I_{\text{coarse}}^i))))), \quad (9)$$

$$I_{\text{coarse}}^{i+1} = [I, M_{\text{refined}}^i], \quad (10)$$

where  $R^i$  is one rescale function of  $Rs$ ,  $i$  denotes the refinement stage as the upper right mark. For simplicity, Eq. (9) does not include aggregation. In practice, we select enough  $R^i$ 's as required regarding performance or by supporting resource. The relation between performance and the number of  $R^i$  is illustrated in Fig. 7. And Fig. 4 is an example.

IoU/mBA	Coarse Mask	SegFix [53]	MGMatting [51]	CascadePSP [9]	CRM(Ours)
FCN-8s [30]	72.39/53.63	72.69/55.21	72.31/57.32	77.87/67.04	<b>79.62/69.47</b>
DeepLabV3+ [5]	89.42/60.25	89.95/64.34	90.49/67.48	<b>92.23/74.59</b>	91.84/ <b>74.96</b>
RefineNet [27]	90.20/62.03	90.73/65.95	90.98/68.40	92.79/74.77	<b>92.89/75.50</b>
PSPNet [55]	90.49/59.63	91.01/63.25	91.62/66.73	93.93/75.32	<b>94.18/76.09</b>
Average Improve.	0.00/0.00	0.47/3.30	0.73/6.10	3.58/14.05	<b>4.01/15.12</b>

Table 1. IoU and mBA results on the BIG dataset comparing with other mask refinement methods. Coarse mask is from FCN, DeepLabV3+, RefineNet and PSPNet. Best results are noted with **bold**. Average Improve. represents average improvement based on coarse mask.

This strategy can also be regarded as a variant of coarse-to-fine operations, where methods [9, 47] realize it through cascade in decoder, and method of [18] through moving window size in range (256, 512, 1024, and 2048). They can also use this strategy to shrink the gap. Nevertheless, the relatively heavy cascade-based network and many forward times in inference design hinder their usage. Take CascadePSP [9] as example, CascadePSP [9] uses the whole ResNet-50 [17] as backbone, but CRM use it without conv5\_x. Then, the cascade-based decoder in CascadePSP [9] (three resolution up-samplings and the corresponding computation) is more costly than CRM’s CAM and  $D_\phi$ . Therefore, even with multi-resolution inference, the whole refinement process of CRM can be more than twice as fast as CascadePSP [9] in Tab. 2.

## 4. Experiments

In this section, we evaluate our CRM and compare it with other corresponding state-of-the-art methods on BIG [9], COCO [28] and relabeled PASCAL VOC 2012 [12]. We evaluate the Intersection over Union (IoU), mean Boundary Accuracy (mBA) [9], panoptic quality (PQ) [21] and average precision (AP) to measure the ability. Then, we present visualization along with ablation studies to understand the effectiveness of our CRM.

### 4.1. Datasets and Methods of Comparison

For training datasets, we follow the setting of CascadePSP [9]. MSRA-10K [10], DUT-OMRON [48], ECSSD [41], and FSS-1000 [23] are merged into the training datasets, consisting of 36,572 images with diverse semantic classes (>1,000 classes). For the testing datasets, CascadePSP [9] proposes an high-resolution image segmentation dataset, named BIG, for evaluation in ultra high-resolution. The image resolution in BIG ranges from 2K to 6K. To prove that our proposed model is general, we evaluate CRM as the extension of Panoptic Segmentation [25] and Entity Segmentation [36]. We also evaluate CRM on relabeled PASCAL VOC 2012, which is introduced in [9].

We choose CascadePSP [9] as the main comparison method on ultra high-resolution. MGMatting [51] is chosen as mask-guided matting method and Segfix [53] as a

Method (IoU/mBA)	Time(s)	FLOPs(G)	Params(M)
CasPSP (93.9/75.3) [9]	620	26518	67.62
CRM (94.2/76.1)	425	2536	9.27
CRM* (93.9/76.3)	259	1331	9.27

Table 2. Comparison of total inference time, FLOPs, and the number of parameters on the BIG dataset. CasPSP denotes CascadePSP and selects patches to compute. CRM computes on all pixels. CRM\* is a computational-friendly version by just computing the region of interest. Time is recorded on the whole BIG dataset. FLOPs are tested on the same image (2560\*1706).

high-resolution segmentation refinement method. PanopticFCN [25] and Entity Segmentor [36] make benchmark of panoptic and entity segmentation. Our proposed method performs better in terms of precision and speed in almost all experiments, especially on high-resolution images.

### 4.2. Implementation Details

We implement our model with PyTorch [32], and use ResNet-50 [17] without conv5\_x as our  $E_\theta$ . For training, we use Adam [20] with  $2.25 \times 10^{-4}$  learning rate. The learning rate is reduced to one-tenth at steps 22,500 and 37,500 in a total of 45,000 steps. The training input concatenates  $224 \times 224$  patches cropped from the original images and their corresponding perturbed masks. The perturbed masks are randomly perturbed on ground truth with a random IoU threshold between 0.8 and 1.0.

For evaluation, we select 4 rescale ratios from a continuous range to refine in experiments. The total inference time of CRM is still less than half of CascadePSP [9].

### 4.3. Quantitative Results

In Tab. 1 and Tab. 2, we show comparison among our CRM, CascadePSP [9], Segfix [53], and MGMatting [51]. (SegFix and MGMatting perform better on a rescaled image with a downsample ratio 0.5.) They prove that CRM’s performance is better, and it runs faster on high-resolution. All segmentation refinement models are trained on low-resolution images and tested on high-resolution images. Segfix and MGMatting’s refinement performances are not as good as other methods without a special design for ultra



Figure 5. Qualitative comparison between Segfix, CascadePSP and CRM on the coarse mask from FCN, DeepLabV3+, RefineNet and PSPNet. The images are from BIG (2K ~ 6K). And the black-white mask in bottom left part of first column is the coarse mask.

Method	PQ	Method	AP
PanopticFCN [25]	41.0	EntitySeg [36]	38.1
PanopticFCN+CRM	41.8	EntitySeg+CRM	38.9

Table 3. The performance after extending PanopticSeg and EntitySeg with our CRM without finetuning.

high-resolution images in BIG [9]. CascadePSP [9] gains more IoU after refinement. Moreover, our CRM produces the highest-quality refinement.

Besides, the inference time is essential for the ultra high-resolution task. Tab. 2 shows that CRM takes less than half inference time of CascadePSP [9] on the whole BIG dataset. FLOPs and parameters are also less. This advantage is due to the simplicity of CRM.

The experiments on panoptic segmentation and entity segmentation are illustrated in Tab. 3. After adding CRM to [25] and [36], their segmentation performance is enhanced.

We also report our performance on relabeled Pascal VOC 2012 in Tab. 4. Compared with CascadePSP [9] and Segfix [53], CRM runs better than Segfix [53] and is comparable with CascadePSP on IoU, but tends to emphasize more on details.

These quantitative results show CRM’s general effectiveness on ultra high-resolution images as well as low-resolution ones.

#### 4.4. Qualitative Results

We show comparison among CascadePSP [9], Segfix [53] and our proposed CRM in Fig. 5. There are more

IoU/mBA	CM	SF [53]	CasPSP [9]	CRM
FCN-8s [30]	68.85	70.02	72.70	73.74
	<u>54.05</u>	<u>57.63</u>	<u>65.36</u>	<u>67.17</u>
DeepLab	87.13	88.03	89.01	88.33
V3+ [5]	<u>61.68</u>	<u>66.35</u>	<u>72.10</u>	<u>72.25</u>
RefineNet [27]	86.21	86.71	87.48	87.18
	<u>62.61</u>	<u>66.15</u>	<u>71.34</u>	<u>71.54</u>
PSPNet [55]	90.92	91.98	92.86	92.52
	<u>60.51</u>	<u>66.03</u>	<u>72.24</u>	<u>72.48</u>

Table 4. Quantitative comparison on relabeled PASCAL VOC 2012. Due to the limited width, CM represent coarse mask, SF represents SegFix, and CasPSP denotes CascadePSP.

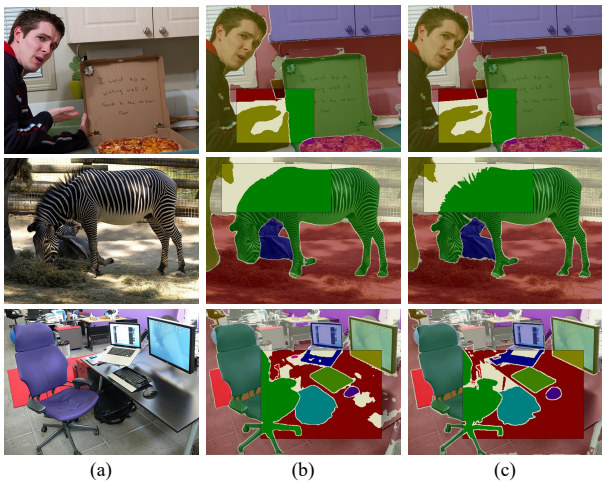


Figure 6. CRM applied in panoptic segmentation. (a) Input image, (b) coarse panoptic segmentation mask, (c) refined mask by our CRM. The images are from COCO.

details in our refinement results. It generates matting-style results with only semantic segmentation annotation in training—the matting benefits from continuous alpha-value supervision. Further, the missing part in coarse masks can be reconstructed better through CRM.

In addition, we show some visualization of applying CRM into panoptic segmentation in Fig. 6. We can see the mask details and overall segmentation are considerably improved. More results in supplement material further manifest the effectiveness of CRM and the continuous modeling.

#### 4.5. Ablation Study

**CRM and Inference Resolutions** CAM and implicit function are the key contributions of our work. The rows in Tab. 5 shows the existence of CRM and implicit function can enhance the performance on every resolution (the first column means the rescale ratios on  $I_{coarse}^i$ ).

For the inference strategy, we analyze the columns of Tab. 5. CRM refines a good general mask at low-resolution (IoU mainly increased in low resolution). As the resolution grows, more details are generated, and mBA increases.

IoU/mBA	w/o CAM&Impl.	w CAM&Impl.
0.125	92.68/63.70	93.07/65.61
0.25	93.49/69.23	93.88/71.41
0.5	93.85/73.43	94.15/74.95
1.0	93.94/75.42	94.18/76.09

Table 5. The effect of CRM and inference resolutions with PSPNet [55]’s output as coarse mask. Impl. denotes implicit function.

CAM	Impl.	IoU	mBA
×	×	93.94	75.42
✓	×	93.99	75.93
×	✓	93.96	75.55
✓	✓	94.18	76.09

Table 6. The ablation study about CAM and implicit function with PSPNet [55]’s output as coarse mask.

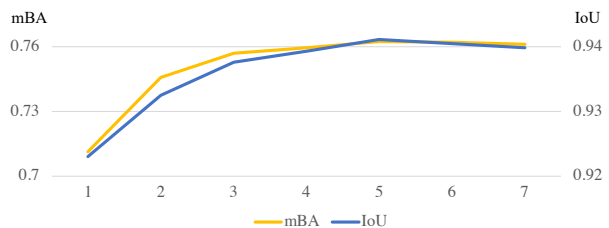


Figure 7. The effect of inference’s continuity. The horizontal axis represents the number of uniformly sampled points between 0 and 1. The sampled points are rescale ratios of input.

**CAM and Implicit Function** The results in Tab. 6 show CAM and implicit functions are all indispensable parts of CRM. Together, they achieve synergy effects.

**The effect of inference’s continuity** From Fig. 7, we can see the performance is growing with the number of sampled rescale ratios between 0 and 1. More numbers mean more continuity in the resolutions of inference, which helps improve performance until convergence. Different from the chosen rescale ratios in Fig. 4 and Tab. 5, the final performances are almost the same level as Fig. 4 and Tab. 5.

## 5. Conclusion

We have proposed CRM to refine segmentation on ultra high-resolution images. CRM continuously aligns the feature map with the refinement target, which helps aggregate features for reconstructing details on the high-resolution mask. Besides, our CRM shows its significant generalization potential regarding low-resolution training and ultra high-resolution testing. Experiments show that continuous modeling is promising in terms of performance and speed.

**Limitations** We use the configuration of “low-resolution training and ultra high-resolution testing” at present. Using ultra high-resolution images to train and test is still resource-consuming. Addressing this challenging problem will be our future work like the usage pretraining for segmentation [35] or low-resolution training and testing [34].



## 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. 4
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2
- [3] 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. 2
- [4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2
- [5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1, 2, 6, 8
- [6] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *CVPR*, pages 8924–8933, 2019. 2
- [7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 1, 3, 4, 5
- [8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 3
- [9] 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. 1, 2, 3, 4, 5, 6, 7, 8
- [10] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2014. 6
- [11] Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021. 3
- [12] 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. 6
- [13] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 2
- [14] Spyros Gidaris and Nikos Komodakis. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *CVPR*, pages 5248–5257, 2017. 2
- [15] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *CVPR*, pages 7519–7528, 2019. 2
- [16] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. In *CVPR*, pages 3828–3837, 2019. 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6
- [18] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *CVPR*, pages 16755–16764, 2021. 1, 2, 6
- [19] Md Amirul Islam, Shujon Naha, Mrigank Rochan, Neil Bruce, and Yang Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv preprint arXiv:1703.00551*, 2017. 2
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 6
- [22] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 1, 2
- [23] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, pages 2869–2878, 2020. 6
- [24] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *CVPR*, pages 8950–8959, 2020. 2
- [25] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *CVPR*, pages 214–223, 2021. 6, 7
- [26] Di Lin, Dingguo Shen, Siting Shen, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Zigzagnet: Fusing top-down and bottom-up context for object segmentation. In *CVPR*, pages 7490–7499, 2019. 2
- [27] 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, 2, 6, 8
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6
- [29] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *arXiv preprint arXiv:1710.01020*, 2017. 2
- [30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2, 6, 8
- [31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 1, 3, 4

- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019. 6
- [33] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, 2019. 2
- [34] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *CVPR*, 2021. 8
- [35] Lu Qi, Jason Kuen, Zhe Lin, Jiuxiang Gu, Fengyun Rao, Dian Li, Weidong Guo, Zhen Wen, and Jiaya Jia. Casp: Class-agnostic semi-supervised pretraining for detection and segmentation. *arXiv preprint arXiv:2112.04966*, 2021. 8
- [36] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. *arXiv preprint arXiv:2107.14228*, 2021. 6, 7
- [37] Lu Qi, Shu Liu, Jianping Shi, and Jiaya Jia. Sequential context encoding for duplicate removal. *NeurIPS*, 2018. 3
- [38] Lu Qi, Yi Wang, Yukang Chen, Ying-Cong Chen, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Pointins: Point-based instance segmentation. *TPAMI*, 2021. 2
- [39] Tamar Rott Shaham, Michaël Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. Spatially-adaptive pixel-wise networks for fast image translation. In *CVPR*, pages 14882–14891, 2021. 3
- [40] Tiancheng Shen, Xia Li, Zhisheng Zhong, Jianlong Wu, and Zhouchen Lin. R<sup>2</sup>-net: Recurrent and recursive network for sparse-view ct artifacts removal. In *MICCAI*, pages 319–327, 2019. 3
- [41] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 38(4):717–729, 2015. 6
- [42] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *NeurIPS*, 33, 2020. 3
- [43] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013. 3
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2
- [45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2
- [46] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, pages 1568–1576, 2017. 2
- [47] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zhe Lin, and Alan Yuille. Meticulous object segmentation. *arXiv preprint arXiv:2012.07181*, 2020. 1, 2, 4, 6
- [48] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 6
- [49] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018. 1, 2
- [50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 3, 4
- [51] 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. 6
- [52] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnets: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 1, 2
- [53] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, pages 489–506, 2020. 1, 6, 7, 8
- [54] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnets for real-time semantic segmentation on high-resolution images. In *ECCV*, pages 405–420, 2018. 3
- [55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1, 2, 6, 8
- [56] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, pages 267–283, 2018. 1, 2
- [57] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015. 2
- [58] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. *arXiv preprint arXiv:2103.15875*, 2021. 3, 4
- [59] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. Canet: Co-attention network for rgb-d semantic segmentation. *PR*, 2022. 2
- [60] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. Rgb-d co-attention network for semantic segmentation. In *ACCV*, 2020. 2