MaSS13K: A Matting-level Semantic Segmentation Benchmark

Chenxi Xie[1,2†] Minghan Li[1,2†] Hui Zeng[2] Jun Luo[2] Lei Zhang[1,2*]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute

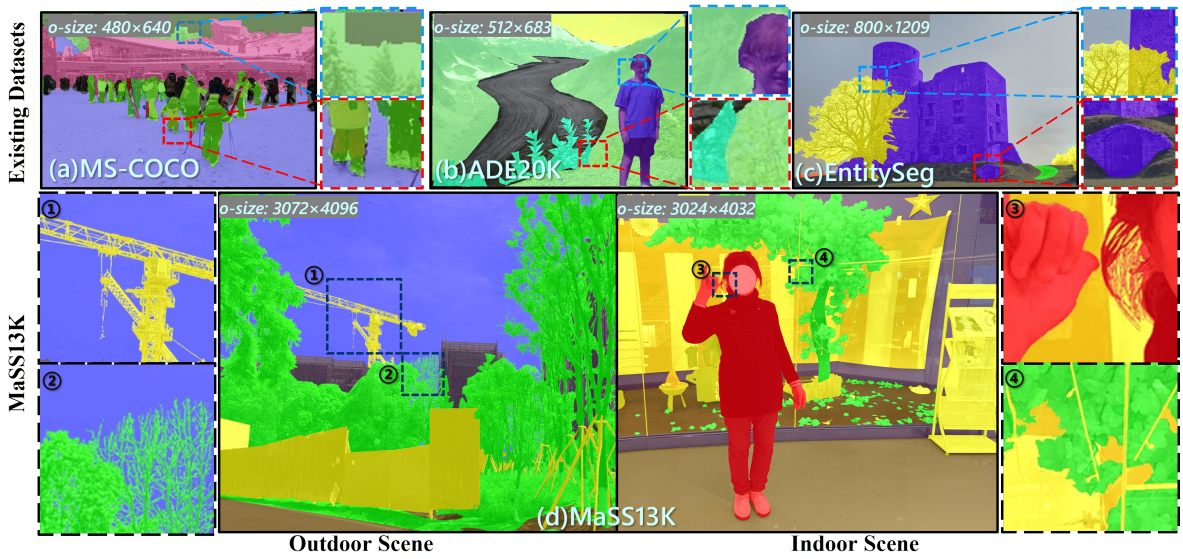

Figure 1. Comparisons on the resolution and annotation quality between existing semantic segmentation datasets (a) COCO, (b) ADE20K, (c) EntitySeg and (d) our established dataset MaSS13K. Note that the 'others' category, which is highlighted in yellow, in our MaSS13K actually contains a group of well-segmented objects without specific class names. Please zoom-in for a better view.

Abstract

High-resolution semantic segmentation is essential for applications such as image editing, bokeh imaging, AR/VR, etc. Unfortunately, existing datasets often have limited resolution and lack precise mask details and boundaries. In this work, we build a large-scale, matting-level semantic segmentation dataset, named MaSS13K, which consists of 13,348 real-world images, all at 4K resolution. MaSS13K provides high-quality mask annotations of a number of objects, which are categorized into seven categories: human, vegetation, ground, sky, water, building, and others. MaSS13K features precise masks, with an average mask complexity 20-50 times higher than existing semantic segmentation datasets. We consequently present a method specifically designed for high-resolution semantic segmentation, namely MaSSFormer, which employs an efficient pixel decoder that aggregates high-level semantic features and low-level texture features across three stages, aiming to produce high-resolution masks with minimal computational cost. Finally, we propose a new learning paradigm, which integrates the high-quality masks of the seven given categories with pseudo labels from new classes, enabling MaSSFormer to transfer its accurate segmentation capability to other classes of objects. Our proposed MaSSFormer is comprehensively evaluated on the MaSS13K benchmark together with 14 representative segmentation models. We expect that our meticulously annotated MaSS13K dataset and the MaSSFormer model can facilitate the research of high-resolution and high-quality semantic segmentation. Datasets and codes can be found at https://github.com/xiechenxi99/MaSS13K.

1. Introduction

Semantic segmentation, which assigns a category label to each pixel in an image, has made significant strides over the past decades, driven by the advancements of deep learning and the availability of large-scale benchmark datasets such as COCO-Stuff [30], and ADE20K [55], among others. Building upon these datasets, researchers have developed a variety of networks, including FCN-based models [34] such

*Corresponding author. †Equal contribution. This work is supported by the PolyU-OPPO Joint Innovative Research Center.

as the DeepLab series [5, 6]. More recently, with the introduction of transformer architectures [32, 42], transformer-based models [7, 9] have shifted the paradigm of semantic segmentation from pixel-level classification to mask-level classification, achieving new state-of-the-art performance.

The image resolution in COCO-Stuff, ADE20K, and Pascal VOC datasets, however, is commonly below 1000×1000. As the resolution of images encountered in our daily life has significantly risen, there is a growing demand for high-resolution semantic segmentation in applications such as image editing [13], bokeh imaging [23], image retouching [33], AR/VR [26], *etc*., where fine-scale mask details significantly impact user experience. Recognizing the limitations of low-resolution data, researchers have started to construct some higher-resolution datasets such as Mapillary Vistas [35] and EntitySeg [38]. However, these still fall short with resolutions below 2K and struggle to annotate objects with complex structures and details accurately, as illustrated in Fig. 1. Some datasets, like the DIS [39] and matting datasets [24, 27, 28], are meticulously annotated but they are designed for class-agnostic binary segmentation tasks, which can only be used to differentiate the foreground from the background, rather than parsing the entire scene into semantic categories. Therefore, there remains a demand for highly accurate semantic segmentation datasets to advance the development of high-resolution segmentation models.

To this end, we introduce the **Ma**tting-level **S**emantic **S**egmentation dataset, namely **MaSS13K**, which comprise 13,348 4K-resolution images captured from the real-world scenes. MaSS13K provides exceptionally high-quality annotations (please refer to Fig. 1) on a variety of objects, which are categorized into seven common semantic categories, including '`human`', '`vegetation`', '`ground`', '`sky`', '`water`', '`buildings`' and '`others`'. It should be noted that *'others' in MaSS13K is not the 'background' class in other datasets*. It actually refers to a group of well-segmented objects such as the 'crane' in Fig. 1. Using the mIPQ [37] score to assess mask complexity, MaSS13K's mIPQ is 20 to 50 times higher than that of existing semantic segmentation datasets and three times higher than the finely annotated DIS dataset [39]. With MaSS13K, we conduct a comprehensive analysis on 14 representative semantic segmentation methods [2, 6, 7, 9, 18, 20, 36, 41, 43, 46, 49–51, 53]. We show that while these methods achieve satisfactory overall accuracy, they struggle with capturing fine boundary details and impose significant computational and memory demands for high-resolution inputs. There is a high demand for developing new semantic segmentation methods, which can balance the computational cost and segmentation performance in high-resolution contexts.

To tackle the above-mentioned challenges, we propose a Transformer-based model, namely MaSSFormer, specifically designed for high-resolution semantic segmentation. We analyze the traditional FPN-based [31] pixel-decoder in the context of high-resolution input, and devise a lightweight pixel-decoder to balance the computational consumption while generating high-quality masks with fine boundaries. In specific, the decoding process is divided into two branches. In the high-level global semantic branch, we efficiently aggregate high-level semantic features across layers, providing robust global context. Moreover, we balance detail and semantic information by expanding the receptive field to capture multi-scale structures. In the low-level local structure branch, we extract edge-aware features to enhance detail segmentation accuracy. With the guidance of edge detection, we fuse the features in the two branches and generate the final feature with accurate details. Extensive experiments on MaSS13K dataset demonstrate that our proposed MaSSFormer outperforms existing semantic segmentation methods, particularly in boundary quality.

While the MaSS13K dataset contains seven classes, its highly detailed semantic annotations enable networks to learn the general ability to accurately segment regions based on class-agnostic boundaries. To validate this point, we develop a simple yet effective training pipeline that uses powerful pre-trained models to generate low-quality pseudo-labels for new classes, which are then combined with our finely annotated labels to train MaSSFormer. Our experiments show that MaSSFormer can produce higher-quality masks for these newly introduced classes, effectively transferring its fine segmentation abilities to novel categories. This highlights the significant potential of the meticulously annotated MaSS13K dataset for advancing future research in high-resolution and high-quality semantic segmentation.

Our contributions are summarized as follows:

- We introduce MaSS13K, a large-scale semantic segmentation dataset, containing 13K 4K-resolution images with ultra-high quality semantic annotations.
- We propose a simple yet effective baseline MaSSFormer for high-resolution semantic segmentation, which surpasses existing methods, particularly in boundary quality.
- We devise a pipeline to empower the model with the capacity to accurately segment novel classes of objects without extra human efforts, revealing the potential of the high-quality semantic annotations in MaSS13K for future high-resolution segmentation research.
- We conduct a comprehensive benchmarking on the MaSS13k dataset by evaluating MaSSFormer with 14 cutting-edge semantic segmentation methods.

2. Related Work

Image Segmentation Datasets. Numerous segmentation datasets have been developed recently, generally falling into two categories: multi-class and binary segmentation datasets. Multi-class datasets are primarily designed for

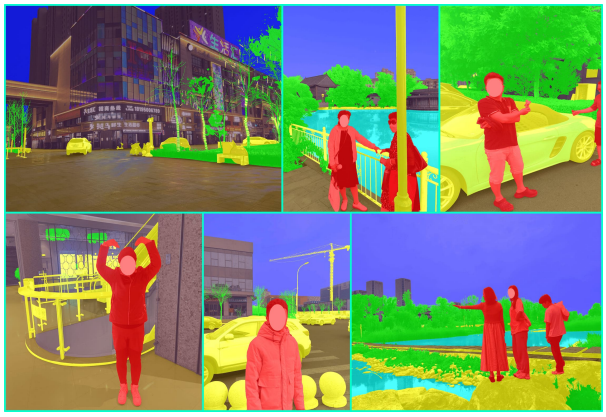

Figure 2. Some typical scenes in our MaSS13K dataset.

tasks such as semantic segmentation [19], instance segmentation [21], and panoptic segmentation [15]. They focus on high-level semantics within scenes and are typically large in scale, encompassing a variety of semantic categories, such as COCO [30] and ADE20K [55], EntitySeg [38], Cityscapes [12] and others [16, 35]. However, these datasets often suffer from lower resolution and poor annotation quality. On the other hand, binary segmentation datasets focus on tasks like salient object detection (SOD) [47, 52], camouflaged object detection (COD) [17, 48], and image matting [28]. They emphasize low-level structures of target objects with fine annotations but are class-agnostic and object-centric, making them unsuitable for full scene parsing. Our MaSS13K addresses these limitations and integrates the strengths of both types of datasets. It includes a range of semantic categories, providing comprehensive coverage of entire scenes. Additionally, it offers high-resolution and matting-level annotations, surpassing the accuracy of existing datasets.

Semantic Segmentation Methods. With the development of deep learning, semantic segmentation has made significant progress. Initially, FCN-based [34] methods dominate the field, with notable methods including DeepLab series [3–5], PSPNet [54], OCRNet [51] and some other efficient methods [49, 50]. These methods introduce various modules to aggregate multi-scale semantic information, classifying each pixel directly. With the rise of Transformers and their application in object detection [1], unified mask classification frameworks such as MaskFormer [2, 7, 9] have been proposed for multiple segmentation tasks. This approach decouples mask prediction from classification, achieving impressive performance. However, due to the limitations of existing semantic segmentation datasets, these methods are mostly designed for low-resolution data, resulting in challenges when applied to high-resolution datasets, such as the difficulties in global semantic awareness and poor boundary details. While some methods have considered edges [29, 44, 49], they focus on how to use edges to differentiate semantic regions rather than improving boundary and detail segmentation quality. Unlike previous methods, our MaSSFormer is designed to tackle the challenges of high-resolution semantic segmentation, ensuring semantic accuracy while producing segmentation maps with precise detail and sharp boundaries.

Table 1. Statistical comparison between MaSS13K and other segmentation datasets. 'SS' denotes semantic segmentation and 'BS' denotes binary segmentation. '#Size' indicates the diagonal length. The values are presented as mean ± standard deviation.

Type	Dataset	#Image	#Size	#Classes	#IPQ
SS	COCO-Stuff [30]	118K	762 ± 70	171	7 ± 4
	ADE20K [55]	20K	667 ± 227	150	17 ± 10
	Cityscapes [12]	3K	2290 ± 0	19	18 ± 7
	EntitySeg [38]	11K	2684 ± 2240	151	20 ± 29
BS	BIG [10]	0.3K	4655 ± 1312	1	21 ± 63
	HRSOD [52]	2K	4405 ± 1631	1	6 ± 13
	UHRSD [47]	6K	6185 ± 1332	1	7 ± 9
	DIS5K [39]	5K	4041 ± 1618	1	116 ± 452
SS	MaSS13K	13K	5144 ± 492	7	383 ± 818

3. MaSS13K Dataset

Data Collection. To address the shortage of high-resolution semantic segmentation data, we construct a large-scale dataset, which consists of 13,348 images of 4K resolution of real-world scenes. These images were captured by various smartphones, including iPhone 14 Pro, Huawei P50, Huawei Mate50, OPPO Find X5, OPPO Reno 8, Vivo X80, and Xiaomi 12, and under different lighting conditions, weather states, and times of day. The images encompass a diverse range of indoor and outdoor scenes, including urban areas, natural landscapes, street views, wilderness, parks, mall interiors, and other public spaces. The diverse sources of data collection not only enhance the dataset's scene diversity but also improve the segmentation model's adaptability to different environmental conditions. The snapshots of some typical image scenes are shown in Fig. 2. Note that all images have undergone rigorous ethical and privacy reviews to ensure privacy regulations, and a screening process to remove lower quality pictures.

Data Annotation. Each image is annotated with pixel-level precision to create accurate semantic segmentation masks for seven categories: '`human`', '`water`', '`vegetation`', '`building`', '`ground`', '`sky`' and '`others`'. An experienced team was invited to conduct this annotation using the editing tools in PhotoShop. For each of the seven categories, annotators were asked to perform matting-level labeling, focusing on accurate boundary delineation of fine features like hair, leaf edges, and intricate object contours. Afterward, each annotated image went through a multi-stage quality control process, including peer review and final check by senior annotators, to ensure the labeling accuracy and consistency.

It should be noted that the '`others`' category in our

MaSS13K dataset is different from the 'background' class in other semantic segmentation datasets [16, 30, 55]. In our dataset, the category '`others`' actually contains a group of accurately segmented objects but without specified class names. As shown by the yellow color in Fig. 1, the masks of the '`others`' regions are in rich details. Therefore, they are treated as a unique class in our training and evaluation.

Statistics Analysis. To better understand the statistics and advantages of our proposed MaSS13K dataset, we compare it against four popular semantic segmentation datasets, including COCO-Stuff [30], ADE20K [55], Cityscapes [12]) and EntitySeg [38], and four datasets used for high-quality binary segmentation (HQBS), including HRSOD [52], UHRSD [47], DIS5K [39] and BIG [10]. We comprehensively compare these datasets in several important attributes for high-resolution segmentation tasks, including dataset size, spatial size, number of categories and mask complexity. Inspired by [39], we also use the mean isoperimetric inequality quotient (mIPQ) [37] metric to measure the overall complexity of masks in an image: $\mathrm{mIPQ} = \frac{1}{4\pi n}\sum_{i=1}^{n}\frac{L_i^2}{A_i}$, where L^i and A_i denote the mask perimeter and the region area for the ith category, and n denotes the total number of categories in this image.

The statistics are presented in Tab. 1. We can see that our dataset has a much higher average diagonal length, which can well represent image resolution, than existing semantic segmentation datasets. Higher resolution allows clearer and more detailed edges to be discerned. While the HQBS datasets also have high resolution, our dataset surpasses them in the number of images and the number of object categories. In addition, MaSS13K has a very small variance in image resolution, indicating a more consistent image quality. In terms of the mIPQ metric, our dataset is 20 times more complex than the EntitySeg dataset, which is known for its high-quality annotations. Even compared to the most finely labeled HQBS dataset DIS5K, our MaSS13K still exceeds it by a significant margin.

4. High-Resolution Segmentation Model

4.1. Overview of MaSSFormer

Our segmentation model, namely MaSSFormer, is built upon the architecture of Mask2Former [7], as illustrated in Fig. 3. The network consists of three main components: a pixel encoder, a pixel decoder, and a transformer decoder. Specifically, for an input image $I \in \mathbb{R}^{3\times H\times W}$ with height H and width W, the pixel encoder generates pixel embeddings $\{F_i\}_{i=1}^{4}$, which are then fed into the pixel decoder to produce multi-scale mask features $\{D_i\}_{i=1}^{4}$. Multi-scale mask features are used to update learnable queries through the Transformer decoder. The updated queries are then passed through MLP layers to generate mask embeddings and class embeddings, which are combined with D_1 to produce the final semantic segmentation results. In this work, we focus on the design of pixel decoder to accommodate high-resolution inputs and generate high-quality mask prediction with precise details.

In our pixel decoder, we use SE block with channel-attention [22] to squeeze the dimension of input pixel embeddings $\{F_i\}_{i=1}^{4}$ and obtain squeezed features $\{S_i\}_{i=1}^{4}$ of resolution $\{\frac{H}{k}\times\frac{W}{k}, k=4,8,16,32\}$. There are two key challenges for high-resolution and accurate semantic segmentation. First, for high-resolution inputs, the receptive field becomes relatively small, making the *aggregation of high-level semantics difficult*. Second, *precise object edges and fine details* are difficult to extract accurately. To address these challenges, we divide $\{S_i\}_{i=1}^{4}$ and image I into two groups: $\{S_2, S_3, S_4\}$ and $\{S_1, I\}$, and aggregate these features with two parallel branches to obtain global semantics and local details, respectively. In the first branch, we use a Cross Semantic Transmission (CST) module to aggregate global context while generating features with larger spatial resolution. We further introduce a Receptive Field Broaden (RFB) module to expand the receptive field while capturing multi-scale information. In the second branch, we introduce a Low-level Structure Extraction (LSE) module to directly capture high-resolution features with details. Finally, an Edge Guided Fusion (EGF) module is employed to fuse the outputs of the two branches and efficiently produce high-resolution features for mask prediction.

4.2. Network Design Details

Global Semantic Branch. As shown in Fig. 3, we design a CST module to aggregate global semantic information. Inspired by [2, 54], we apply global average pooling and a 1×1 convolution layer on F_4, and add it to S_4 to enhance the global context. Then, the context-enhanced features are fed into a deformable convolution layer with batch normalization and ReLU activation to obtain feature D_4, which contains global semantics but with a small spatial size. In contrast, S_3 shares similar semantics with S_4 while offering higher spatial resolution. We use Cross-Attention to transfer the detailed information from S_3 to D_4, followed by a Feed-Forward Network (FFN) $\mathrm{FFN}(\cdot)$. To reduce the computational cost, attention is performed within non-overlap windows. After passing through FFN, Self-Attention is computed within the window to further enhance semantic information at higher resolutions. Finally, another FFN is applied, followed by a deformable convolution to expand the receptive field. The whole process can be expressed as:

$$\hat{D}_4 = \mathrm{FFN}(\mathrm{CAttn}(f_q(\mathrm{US}_2(D_4)), f_{k,v}(S_3))), \quad (1)$$

$$D_3 = \mathrm{DCN}(\mathrm{FFN}(\mathrm{SAttn}(f_{q,k,v}(\hat{D}_4)))), \quad (2)$$

where $f_{\{qkv\}}$ is linear projection and $\mathrm{US_n}(\cdot)$ means upsampling with factor n. By computing on low-resolution

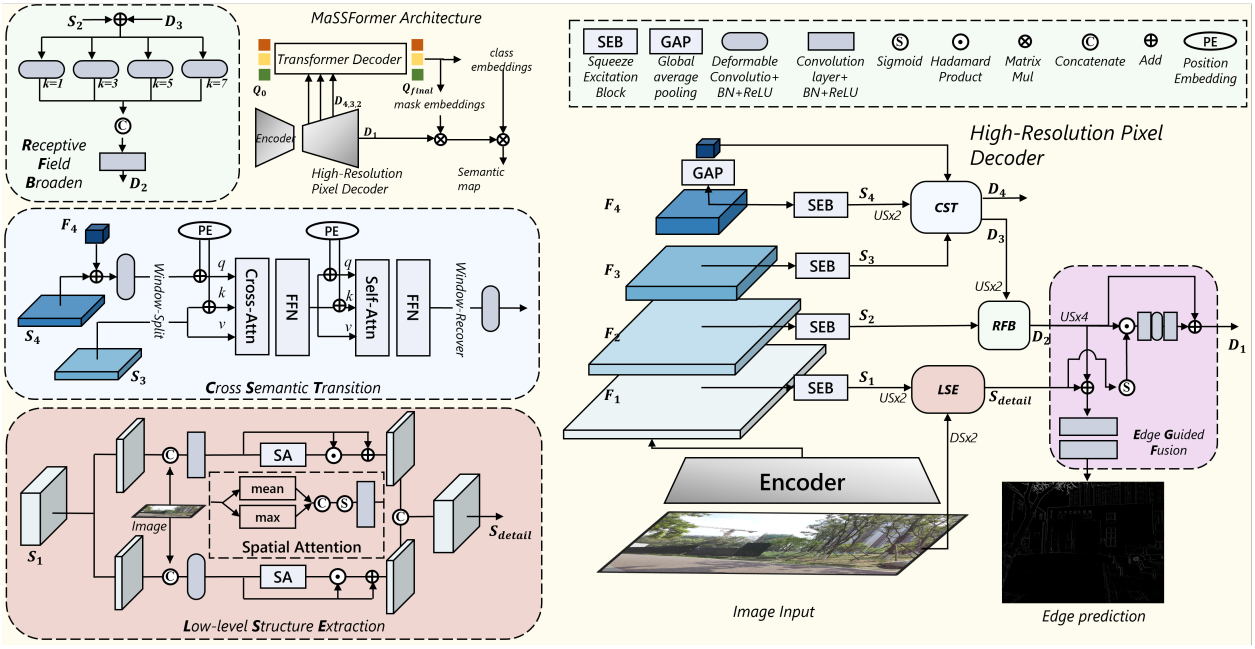

Figure 3. Architecture of MaSSFormer. The model architecture is shown on the top left corner, including the image encoder, pixel decoder and transformer decoder. The detailed structure of the high-resolution pixel decoder is shown on the right of this figure.

features, CST can gather global semantics at a lower computational cost and expand to larger spatial resolutions.

We then up-sample the CST output feature D_3 and add it to S_2 before feeding them to the RFB module for higher resolution semantic features D_2:

$$D_2 = \mathrm{RFB}(\mathrm{US}_2(D_3) + S_2). \tag{3}$$

As shown in Fig. 3, the RFB module includes 4 parallel deformable convolution modules with different kernel sizes $[1, 3, 5, 7]$ and a 1×1 point-wise convolution. What's more, RFB can capture multi-scale structure by setting convolution modules with different kernel size.

Local Structure Branch. Traditional semantic segmentation methods [6, 7, 9] typically do not consider high-resolution features for computational efficiency, which is unsuitable for high-resolution semantic segmentation. In this paper, we design the LSE module to extract low-level structure information with low computational cost. To generate high-quality masks, we upsample S_1 with resolution $\frac{H}{4} \times \frac{H}{4}$ to $\frac{H}{2} \times \frac{H}{2}$ and incorporate the down-sampled image to directly capture low-level features. However, the resolution of $\frac{H}{2} \times \frac{H}{2}$ results in a quadratic computational burden. As shown in Fig. 3, we split the input into two sets of features with fewer channels, each concatenated with the image to learn distinctive features. We then use spatial attention [45] to activate and extract low-level structural information. Finally, the features are concatenated and passed through a 1×1 convolution to output high-resolution features S_{detail}.

Edge Guided Fusion. To reduce the cost in fusing low-level structure features at resolution $\frac{H}{2} \times \frac{W}{2}$, we design an EGF module to efficiently aggregate detailed information. First, we apply consecutive 1×1 convolutions to predict edges $P^{edge} \in \mathbb{R}^{1 \times H \times W}$ as follows:

$$P^{edge} = \mathrm{Conv}(\mathrm{CBR}(S_{detail} + D_2)), \tag{4}$$

where Conv means convolution and CBR means convolution with batch-norm and ReLU. During training, the edge detection task forces the network to learn the low-level structures of the image, allowing S_{detail} to receive feedback and focus on edge regions. Then we use S_{detail} to refine D_2, which can be expressed as:

$$D_2^{refine} = \mathrm{Sigmoid}(S_{detail}) \times D_2. \tag{5}$$

As shown in Fig. 3, after reducing the channels with a 1×1 convolution, we apply a deformable convolution and restore the expanded channels with another 1×1 convolution. Finally, a residual connection is used to output high-resolution features D_1. The intermediate features $\{D_2, D_3, D_4\}$ are then fed into the transformer decoder.

Loss Function. We adopt the loss of Mask2Former [7], which includes classification loss L_{cls} and mask loss L_{mask} (a combination of BCE loss and Dice loss). We assign an extra weight in the mask loss to highlight supervision on edge regions. The weights map for each target label i can be generated using ground-truth label by $W^i = G^i - f_{avg}(G^i, k)$, where $f_{avg}(., k)$ is an averaging filter with kernel size k and G^i is the i-th binary

ground truth label. We assign the weights for the BCE loss as $L^w_{BCE} = \frac{1}{mn}\sum_i^n \sum_j^m l_{bce}(P^i_j, G^i_j)(1 + \lambda W^i_j)$, where i, j represent the j-th pixel on the i-th map and λ is a hyperparameter that adjusts the weight ratio. Additionally, we incorporate an edge detection task in the BCE loss to supervise the edge prediction results, denoted by $L_{edge} = \frac{1}{n}\sum_i^n l_{bce}(P^{edge}_i, G^{edge}_i)$, where G^{edge} represents the edges between different semantic regions in the ground truth. The overall loss function is formulated as:

$$L_{total} = L^w_{BCE} + L_{Dice} + L_{cls} + L_{edge}. \tag{6}$$

4.3. Segmentation on New Classes

Though our MaSS13K detaset consists of six commonly used categories and an 'others' category, the '`others`' category actually contains various well segmented but unnamed objects. To fully exploit the precise segmentations of those objects in our dataset, we present a novel pipeline to enable MaSSFormer to segment higher-quality masks for new classes beyong the six predefined categories.

We start with the image training set $\{I_i\}^n_{i=1}$ and the corresponding annotation set for 7 classes $\{G^j_i\}^6_{j=0}$, where G^j_i represents the mask of the i-th image in the j-th class, and $\{G^0\}$ denotes the annotation set for '`others`'. Using existing semantic segmentation models [25, 40], we can automatically annotate a new class in $\{I_i\}^n_{i=1}$ to generate a set of pseudo-labels, denoted by $\{\hat{G}^7_i\}$. By mixing the precise labels in the MaSS13K dataset and the pseudo-labels, we obtain a new set of labels to train the model. We then design a label decoupling strategy, which employs the weighting scheme described in Sec. 4.2 to treat the labels differently. Specifically, for the precise labels annotated in MaSS13K, we emphasize the loss on edges to update the network. For pseudo-labels, we apply inverse weights to ignore potentially erroneous edge annotations, allowing the network to focus on learning the main regions of the new classes. The training loss can be expressed as follows:

$$\begin{aligned} L^{new}_{BCE} = &\sum\nolimits_j^{1-6}(1 + \lambda_1 W^j)L_{BCE}(P^j, G^j) + \\ &\sum\nolimits_j^{0,7}(1 - \lambda_2 W^j)L_{BCE}(P^j, G^j), \end{aligned} \tag{7}$$

where W^j, P^j, G^j represent weights map, predicted label and ground-truth label for the j-th class. In this way, we can train MaSSFormer to segment new classes one by one.

5. MaSS13K Benchmark

Experiment Settings. We present a comprehensive benchmarking on our MaSS13K dataset with existing semantic segmentation models and our MaSSFormer. We split the 13,348 images in MaSS13K into three subsets: **MaSS-train (11,348)**, **MaSS-val (500)**, and **MaSS-test (1,500)** for model training, validation, and testing, respectively. We select 14 representative and advanced semantic segmentation methods in the evaluation, which can be categorized into lightweight methods (including STDC2 [18], BeSeNetV2 [50], PIDNet [49], FeedFormer [41], SegNext [20], SeaFormer [43] and CGRSeg [36]), FCN-based methods (including DeepLabv3+ [6], UperNet [46] and OCRNet [51]), and Transformer-based methods (including MaskFormer [9], Mask2Former [7], PEM [2] and MPFormer [53]). We provide two variants of our MaSSFormer: the baseline version MaSSFormer with ResNet-50 backbone and the lightweight version MaSSFormer-Lite with ResNet-18 backbone. We implement MaSSFormer by mmsegmentation [11]. We use 1024×1024 crops during training and use the original resolution during testing for all methods. For the competing methods, we adopt their default settings in training on our MaSS13K. More training details of our MaSSFormer can be found in **supplementary materials**.

Evaluation Metrics. Besides mIoU, which measures the overall segmentation accuracy, we use BIoU [8] and boundary F1-score [14] to evaluate the boundary accuracy and quality for high-resolution semantic segmentation. More details can be found in the **supplementary materials**.

5.1. Quantitative Evaluation

Tab. 2 shows the quantitative comparison between the proposed MaSSFormer and previous state-of-the-art semantic segmentation methods. One can see that MaSSFormer achieves the best results in terms of all the mIoU, BIoU and BF1 metrics while maintaining relatively low computational cost and parameter size, demonstrating its effectiveness and efficiency for high-resolution image segmentation. It is generally observed that correct segmentation of the main body is essential for accurate boundary segmentation, hence BIoU and BF1 tend to improve along with mIoU. However, some methods, such as FeedFormer and SegNext, show significant differences (+2.13% on MaSS13K-val) in BIoU despite having similar (-0.54%) mIoU scores. In addition, some segmentation errors in details edges and boundaries will have a small impact on the overall metrics like mIoU, but will significantly affect BIoU and BF1 scores, which can reflect a method's capability of segmenting object details. Our MaSSFormer achieves +0.69% higher in mIoU and +1.57% higher in BIoU than the second-best method, respectively, with the same ResNet50 backbone. Meanwhile, MaSSFormer with ResNet18 backbone also demonstrates competitive performance for high-resolution image segmentation, even higher than many networks with ResNet50 backbone, but with much fewer parameters and Flops. The results of MaSSFormer on each category can be found in the **supplementary materials**.

Table 2. Quantitative evaluation on MaSS13K validation and test sets. The best and the second-best results are highlighted in **bold** and in underlined respectively. FLOPs are all calculated for an input resolution of 4096×4096.

Method	Ori.	Backbone	MaSS-val(500)			MaSS-test(1,500)			Model Stat.	
			mIoU↑	BIoU↑	BF1↑	mIoU↑	BIoU↑	BF1↑	Param	FLOPs
STDC2 [18]	CVPR21	-	83.08	27.99	.3334	83.76	27.72	.3332	12.30M	564G
BiSeNetv2 [50]	IJCV21	-	71.55	25.05	.3182	72.92	24.48	.3171	3.35M	591G
SegNeXt [20]	NIPS22	MSCAN-B	87.71	39.93	.4615	88.11	39.45	.4596	27.57M	1536G
PIDNet-L [49]	CVPR23	-	82.28	31.30	.3475	81.77	30.70	.3479	37.08M	1653G
FeedFormer [41]	AAAI23	lvt	87.17	42.06	.4838	86.56	41.07	.4789	4.65M	300G
SeaFormer-L [43]	ICLR23	-	86.78	38.61	.4498	87.36	38.28	.4489	13.95M	303G
CGRSeg-L [36]	ECCV24	EFv2-L	81.29	34.48	.4090	81.45	33.95	.4071	35.64M	1536G
DeepLabv3+ [6]	ECCV18	R50	86.66	40.08	.4718	85.14	38.65	.4678	41.22M	8008G
UperNet [46]	ECCV18	R50	82.03	35.85	.4181	81.98	35.61	.4170	64.04M	11373G
OCRNet [51]	ECCV20	R50	86.99	32.68	.3808	83.02	31.33	.3731	36.52M	7352G
MaskFormer [9]	NIPS21	R50	83.27	38.61	.4399	83.22	37.90	.4393	41.31M	2396G
Mask2Former [9]	CVPR22	R50	88.28	47.40	.5458	88.00	46.13	.5330	44.01M	3123G
MPFormer [53]	CVPR23	R50	87.76	47.81	.5513	87.18	47.17	.5486	43.9M	4155G
PEM [2]	CVPR24	R50	83.41	40.51	.4675	83.38	39.99	.4644	35.5M	1859G
MaSSFormer-Lite	-	R18	87.11	45.35	.5137	86.13	43.28	.5086	15.07M	771G
MaSSFormer	-	R50	**88.97**	**48.97**	**.5639**	**88.21**	**48.39**	**.5593**	37.42M	2036G

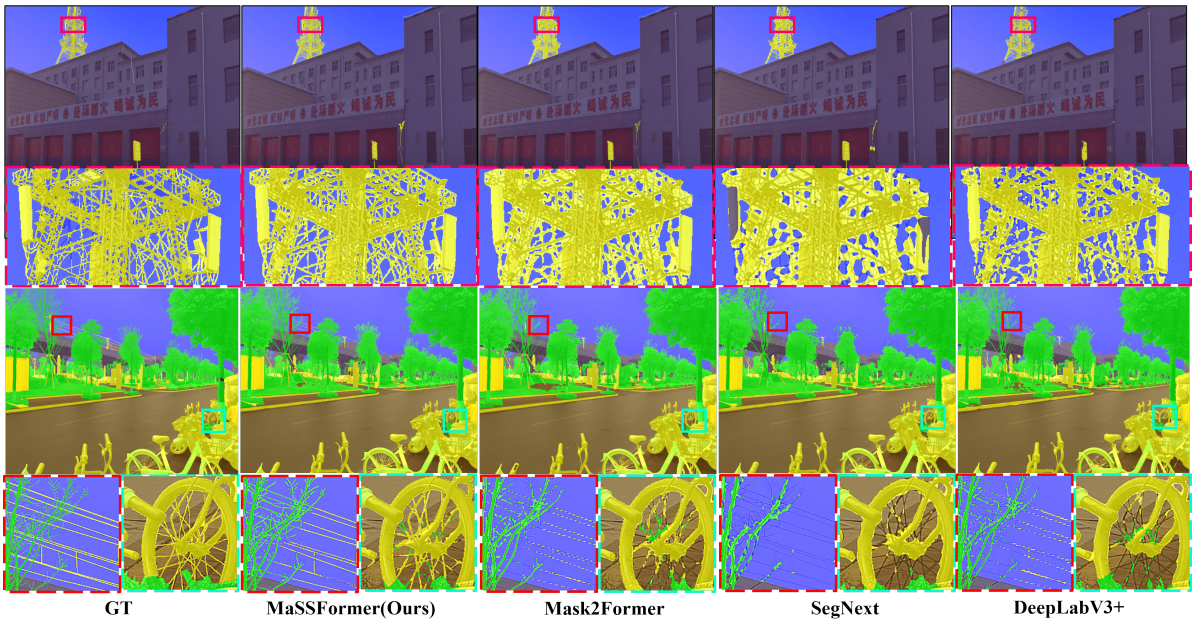

Figure 4. Qualitative comparisons of MaSSFormer with other three baseline methods. Please zoom-in for better view.

5.2. Qualitative Evaluation

Fig. 4 presents qualitative comparisons between our MaSSFormer and three representative methods: Mask2Former, SegNext, and DeepLabV3+, which have the best overall mIoU among the competing methods. We can see that the current methods struggle to distinguish fine details in high-resolution images, often resulting in blurry, discontinuous, or even missing segments. While this may have only a small impact on the mIoU and other metrics, it significantly degrades the visual quality and user experience of the segments. In contrast, our MaSSFormer effectively captures the details, producing higher-quality segmentation results. Besides, MaSSFormer demonstrates high precision in various categories, including meticulous tower structures (top two rows) and thin branches and lines (bottom two rows).

Real-world Application. High-precision semantic image segmentation has many practical real-world applications. For example, mobile photography relies on accurate segmentation of portrait areas to achieve realistic bokeh effects. As shown in Fig. 5, the bokeh effect generated using the mask predicted by the model trained on MaSS13K is more natural and realistic than that by using the model trained on ADE20K, especially in the areas of the hands and legs. This demonstrates the importance of high-precision segmenta-

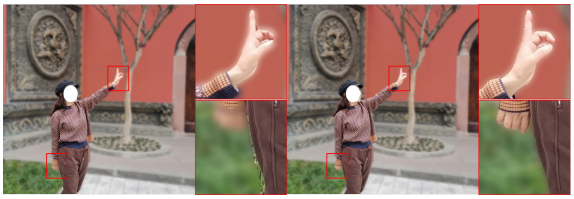

Figure 5. Comparison of bokeh effects with different masks. Please zoom-in for a better view.

Table 3. Quantitative evaluation on novel class **Car**.

Settings	Car	
	mIoU	BIoU
Pseudo Label	94.18	20.44
w/o Accurate Label	92.43	22.52
w/o Label Re-weight	83.23	31.81
w Label Re-weight	95.21	35.68

tion, which depends on high-quality and high-resolution datasets such as MaSS13K.

5.3. Segmentation on New Class

Tab. 3 and Fig. 6 show the segmentation results of MaSSFormer on a new class, 'Car'. The pseudo-labels generated by existing semi-automatic tools achieve high mIoU but struggle with precise edge segmentation, especially when the target and surrounding areas are visually similar. Besides, direct training with these pseudo-labels (upper-right of Fig. 6) does not improve segmentation quality. The bottom two rows of Tab. 3 show the results by mixed training with precise labels. We see that the model learns to enhance edge segmentation from precise annotations, improving the quality of car edges (see the bottom-left of Fig. 6), thus increasing BIoU. However, while the car roof is accurately segmented, the surrounding regions are mistakenly classified as the 'others' class because the incorrect pseudo-labels in the edge region can misguide the model training, leading to a 10.95% decrease in mIoU (see the 3rd row of Tab. 3). With our re-weighting strategy, we reduce the weight of mislabeled edges, forcing the network to learn edge-aware capabilities from accurately labeled categories and new category features with higher reliability, thus achieving high-quality segmentation of new classes. More experiments can be found in the **supplementary materials.**

5.4. Ablation Study

We conduct a series of ablation studies to validate the effectiveness of each component of MassFormer.

Global Semantic Aggregation. The top 3 rows in Tab. 4 demonstrate the effectiveness of our global semantic branch. By introducing the CST module, the mIoU metric is improved by 2.57%, indicating that the acquisition of global semantics can enhance overall precision. Meanwhile, the increase in computational cost and parameters is slight. By incorporating the RFB module, the receptive field increases,

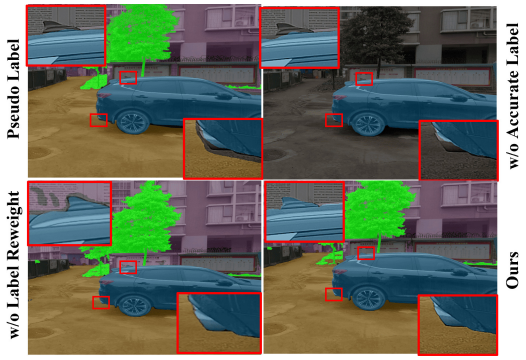

Figure 6. Visual results of MaSSFormer on novel class Car.

Table 4. Ablation on MaSS-val dataset.

Exp.	CST	RFB	LSE	EGF	mIoU	BIoU	Para	FLOPs
#1					85.54	42.69	34.72	1298
#2	✓				88.11	44.42	35.96	1348
#3	✓	✓			88.29	45.23	37.32	1692
#4	✓	✓	✓		88.02	47.36	37.40	1928
#5	✓	✓	✓	✓	88.97	48.97	37.42	2036

leading to a further improvement in mIoU. Compared to mIoU, there is a more significant increase in BIoU, indicating that our RFB module effectively fuses high-resolution features while maintaining semantic accuracy.

Local Detail Extraction and Fusion. Comparing Exps#3 and 4 in Tab. 4, we see that the LSE module significantly increases the BIoU metric, implying that more low-level details can enhance the segmentation accuracy of edge boundaries. However, there is a slight decrease in mIoU, suggesting that direct fusion of low-level features may mislead the higher-level semantics. Exp#5 demonstrates that our designed edge-guided fusion strategy effectively improves low-level structure and high-level semantic feature fusion, resulting in an increase in both mIoU and BIoU.

6. Conclusion

In this paper, we proposed MaSS13K and MaSSFormer for high-resolution semantic segmentation. The MaSS13K contained 13,348 real-word images at 4K resolution, with high-quality matting-level annotations in 7 categories. We then presented MaSSFormer, which efficiently aggregated global semantics and local structure details in high-resolution scenes, to address the challenges of high-resolution semantic segmentation. We compared 14 representative methods with MaSSFormer on MaSS13K, establishing a comprehensive benchmark for high-resolution semantic segmentation. Furthermore, we proposed a scheme to transfer and generalize the fine segmentation capabilities of MaSSFormer to novel classes beyond the original categories, further revealing the potential value of MaSS13K dataset. We hope that MaSS13k can advance the research on high-resolution and high-quality semantic segmentation.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[2] Niccolò Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15804–15813, 2024. 2, 3, 4, 6, 7

[3] Liang-Chieh Chen. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3

[4] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 2, 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 5, 6, 7

[7] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2, 3, 4, 5, 6

[8] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021. 2, 3, 5, 6, 7

[10] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8890–8899, 2020. 3, 4

[11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[12] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3, 4

[13] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[14] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation?. In *Bmvc*, pages 10–5244. Bristol, 2013. 6

[15] Omar Elharrouss, Somaya Al-Maadeed, Nandhini Subramanian, Najmath Ottakath, Noor Almaadeed, and Yassine Himeur. Panoptic segmentation: A review. *arXiv preprint arXiv:2111.10250*, 2021. 3

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 3, 4

[17] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 3

[18] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021. 2, 6, 7

[19] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017. 3

[20] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 2, 6, 7

[21] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. 3

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[23] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020. 2

[24] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 2

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[26] Tae-young Ko and Seung-ho Lee. Novel method of semantic segmentation applicable to augmented reality. *Sensors*, 20 (6):1737, 2020. 2

[27] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007. 2

[28] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 2, 3

[29] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020. 3

[30] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3, 4

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[33] Zerun Liu, Fan Zhang, Jingxuan He, Jin Wang, Zhangye Wang, and Lechao Cheng. Text-guided mask-free local image retouching. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2783–2788. IEEE, 2023. 2

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3

[35] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3

[36] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. *arXiv preprint arXiv:2405.06228*, 2024. 2, 6, 7

[37] Robert Osserman. The isoperimetric inequality. *Bulletin of the American Mathematical Society*, 84(6):1182–1238, 1978. 2, 4

[38] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 2, 3, 4

[39] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 2, 3, 4

[40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 6

[41] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2263–2271, 2023. 2, 6, 7

[42] 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*, 2017. 2

[43] Qiang Wan, Zilong Huang, Jiachen Lu, YU Gang, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The eleventh international conference on learning representations*, 2023. 2, 6, 7

[44] Chi Wang, Yunke Zhang, Miaomiao Cui, Peiran Ren, Yin Yang, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, and Weiwei Xu. Active boundary loss for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2397–2405, 2022. 3

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[46] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 2, 6, 7

[47] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11717–11726, 2022. 3, 4

[48] Chenxi Xie, Changqun Xia, Tianshu Yu, and Jia Li. Frequency representation integration for camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 1789–1797, New York, NY, USA, 2023. Association for Computing Machinery. 3

[49] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19529–19539, 2023. 2, 3, 6, 7

[50] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068, 2021. 3, 6, 7

[51] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 2, 3, 6, 7

[52] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7234–7243, 2019. 3, 4

[53] Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M Ni, and Lei Zhang. Mp-former: Mask-piloted transformer for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18074–18083, 2023. 2, 6, 7

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 3, 4

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 3, 4