IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Mobile-Seed: Joint Semantic Segmentation and Boundary Detection for Mobile Robots

Youqi Liao, Shuhao Kang, Jianping Li, Yang Liu, Yun Liu, Zhen Dong, Bisheng Yang, Xieyuanli Chen

***Abstract*—Precise and rapid delineation of sharp boundaries and robust semantics is essential for numerous downstream robotic tasks, such as robot grasping and manipulation, real-time semantic mapping, and online sensor calibration performed on edge computing units. Although boundary detection and semantic segmentation are complementary tasks, most studies focus on lightweight models for semantic segmentation but overlook the critical role of boundary detection. In this work, we introduce Mobile-Seed, a lightweight, dual-task framework tailored for simultaneous semantic segmentation and boundary detection. Our framework features a two-stream encoder, an active fusion decoder (AFD) and a dual-task regularization approach. The encoder is divided into two pathways: one captures category-aware semantic information, while the other discerns boundaries from multi-scale features. The AFD module dynamically adapts the fusion of semantic and boundary information by learning channel-wise relationships, allowing for precise weight assignment of each channel. Furthermore, we introduce a regularization loss to mitigate the conflicts in dual-task learning and deep diversity supervision. Compared to existing methods, the proposed Mobile-Seed offers a lightweight framework to simultaneously improve semantic segmentation performance and accurately locate object boundaries. Experiments on the Cityscapes dataset have shown that Mobile-Seed achieves notable improvement over the state-of-the-art (SOTA) baseline by 2.2 percentage points (pp) in mIoU and 4.2 pp in mF-score, while maintaining an online inference speed of 23.9 frames-per-second (FPS) with 1024×2048 resolution input on an RTX 2080 Ti GPU. Additional experiments on CamVid and PASCAL Context datasets confirm our method's generalizability. Code and additional results are publicly available at: https://whu-usi3dv.github.io/Mobile-Seed/.**

***Index Terms*—Deep learning for visual perception, visual learning, deep learning methods.**

I. Introduction

Semantic segmentation and boundary detection are fundamental tasks for simultaneous localization and mapping (SLAM) [3], autonomous driving [4], behavior prediction [5] and sensors calibration [6]. Semantic segmentation predicts the categorical labels for each pixel, and the boundary detection task identifies pixels lying on the boundary area where at least one neighborhood pixel belongs to a different class.

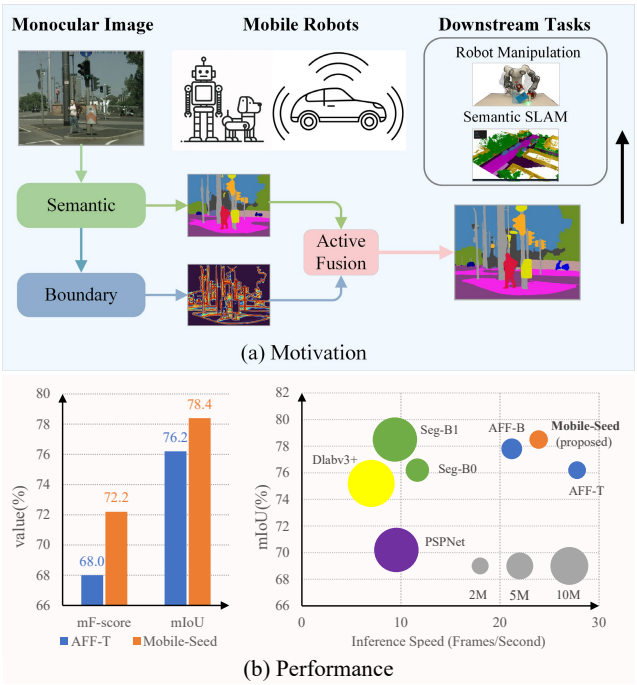


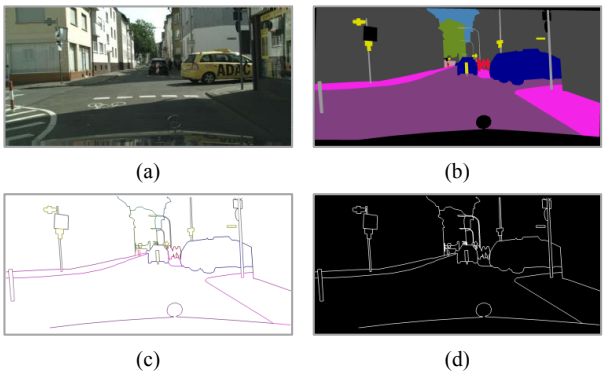
Fig. 1: (a) Motivation map: Mobile-Seed performs pixel-wise segmentation and object boundary detection simultaneously, and then fuses semantic and boundary features for accurate prediction. The boundary detection and semantic segmentation predictions could be transferred for downstream tasks, e.g., robot manipulation, semantic mapping and sensor calibration. (b) Our Mobile-Seed achieves higher performance on both semantic segmentation and boundary detection tasks while keeping real-time efficiency. The resolution of input is 1024×2048 when testing inference speed. "AFF" and "Seg" mean the AFFormer [1] and SegFormer [2], respectively.

As the boundary always surrounds the object's body [7], robust prediction of the body label guides the object boundary detection, while improving the boundary location is crucial for semantic segmentation accuracy. In other words, semantic segmentation and boundary detection are complementary tasks. Moreover, simultaneously extracting segmentation and boundary information in compact robotics is important for semantic SLAM [8], [9], in which the boundary is a strong constraint for solving the relative pose and location, and segmentation is crucial for dynamic object removal. However, on the one hand, most lightweight approaches [1], [2], [10] attempt to solve the semantic segmentation task but overlook the boundary accuracy. On the other hand, existing dual-task learning approaches [7], [11], [12] design novel architectures for performance improvement but neglect the computational burden. Overall, simultaneously capturing the segmentation and boundary has not received enough attention, but this

Manuscript received: Nov. 21, 2023; Revised: Jan. 15, 2024; Accepted: Feb. 20, 2024. This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments. Digital Object Identifier (DOI): see top of this page.

This study was supported by the National Natural Science Foundation Project (No. 42201477, No. 42130105) (Corresponding author: Jianping Li)

Y. Liao, Z. Dong and B. Yang are with Wuhan University, China. S. Kang is with the Technical University of Munich, Germany. J. Li is with Wuhan University, China and Nanyang Technological University, Singapore. Yang Liu is with the King's College London, UK. Yun Liu is with the Institute of Infocomm Research (I2R), A*STAR, and X. Chen is with the National University of Defense Technology, China.



(a) (b) (c) (d)

Fig. 2: Example diagram of color image (a), semantic mask (b), semantic boundary mask (c) and binary boundary mask (d). Semantic boundary masks are generated as [16], [17], and binary boundary masks are generated as [18].

is precisely in urgent need of the robotics society. In this paper, we investigate how to design a lightweight framework for jointly learning the semantic and boundary mask in a complementary manner, as shown in Fig. 1.

To deploy semantic segmentation for real-world online robotic and autonomous driving applications, powerful yet lightweight vision transformers (ViT) [13] have been developed. For example, the hierarchical attention [2], stride attention [10], and window attention [14] are proposed to capture the long-range context with low computation cost and outperform convolution neural network (CNN) based methods by a large margin. However, these advances are still insufficient to accurately locate object boundaries. The main reasons are: i) as the Transformer lacks inductive bias [13], it is not good at capturing fine-grained details in a local window; ii) most methods adopt very simple decoder designs, which lack the ability to capture and recover details. Some recent approaches [1], [15] even remove the decoder for efficiency, called "head free", which exacerbates boundary blurring. For the boundary detection task, most existing approaches [16], [17], [18] overlooked the computational efficiency. In the field of dual-task learning for semantic segmentation and boundary detection, several approaches [11], [12], [19] pointed out that jointly learning the boundary detection and semantic segmentation tasks with reasonable designs benefits both tasks, but none of them discussed how to implement with a lightweight design for mobile robots. It should be retained that the boundary detection task here is significantly different from the edge detection task [20]. Fig. 2 shows the semantic mask, the semantic boundary mask, and the binary boundary mask of a color image. Boundary detection aims to find semantically discontinuous areas instead of dramatic intensity, illumination, or texture changes in edge detection task.

To address the limitations mentioned above, we present a lightweight framework for simultaneous semantic segmentation and boundary detection. The workflow is shown in Fig. 3. Our objective is to utilize the semantic stream to offer fundamental knowledge for the boundary stream while supplementing the semantic segmentation task with fine-grained details captured by the boundary branch. Additionally, we introduce the active fusion decoder (AFD) to learn the fusion weights from inputs and fuse the semantic and boundary features in a dynamic way. Furthermore, we incorporate the dual-task regularization losses to alleviate conflicts arising from deep diverse supervision (DDS) [18]. Experiments on multiple public datasets demonstrate our Mobile-Seed outperforms existing methods by a large margin, especially in predicting crisper boundaries and segmenting small and thin objects. Overall, the main contributions of this paper include:

1) We propose a lightweight joint semantic segmentation and boundary detection framework for mobile robots. This framework can concurrently learn both the boundary mask and semantic mask.
2) We present the AFD for learning the channel-wise relationship between semantic features and boundary features. Compared to the fixed weight methods (fusion weights independent of the input), our AFD is more flexible in assigning proper weights for semantic features and boundary features.
3) We introduce the dual-task regularization loss to effectively mitigate conflicts arising from DDS, allowing the tasks of semantic segmentation and boundary detection to contribute to each other.

II. Related Work

A. Lightweight Semantic Segmentaiton

Since the pioneering approaches fully convolution network (FCN) [21] ushered in a new era, a significant amount of works [22], [23] have been dedicated to addressing semantic segmentation tasks. To reduce the computational burden caused by dense convolution operations on feature maps, the MobileNet [24] proposed the depthwise separable convolution and ShuffleNet [25] proposed the channel shuffle to maintain accuracy. Fast-SCNN [26] proposed the "learning to downsample" module to produce shared low-level fundamental features. During the transformer era, SegFormer [2] was the first transformer-based lightweight design for mobile devices. Activated by Swin-Transformer [27], MobileViT [14] proposed the hybrid CNN and Transformer blocks for local and global processing. TopFormer [10] designs the token pyramid module for scale-aware features. Experiments show that Topformer achieves a better trade-off between accuracy and efficiency than previous approaches. AFFormer [1] proposed the channel-wise attention module and SeaFormer [15] proposed the axial-attention module. Coincidentally, both of them utilized the "head free" decoder design: a simple classification head with several convolution layers, which means predicting at a low resolution without progressive upsampling and refinement. PP-mobileSeg [28] inherited the stride-Former frame [15] and proposed the aggregated attention module (AAM) and valid interpolation module (VIM) to enhance the semantic features. Unlike the above approaches which design single-branch models for semantic segmentation, we introduce a dual-branch framework to simultaneously learn semantic segmentation and boundary detection.

B. Boundary Detection

CASENet [16] is the first multi-label learning framework to identify semantic boundaries. Based on the ResNet-101 [29],

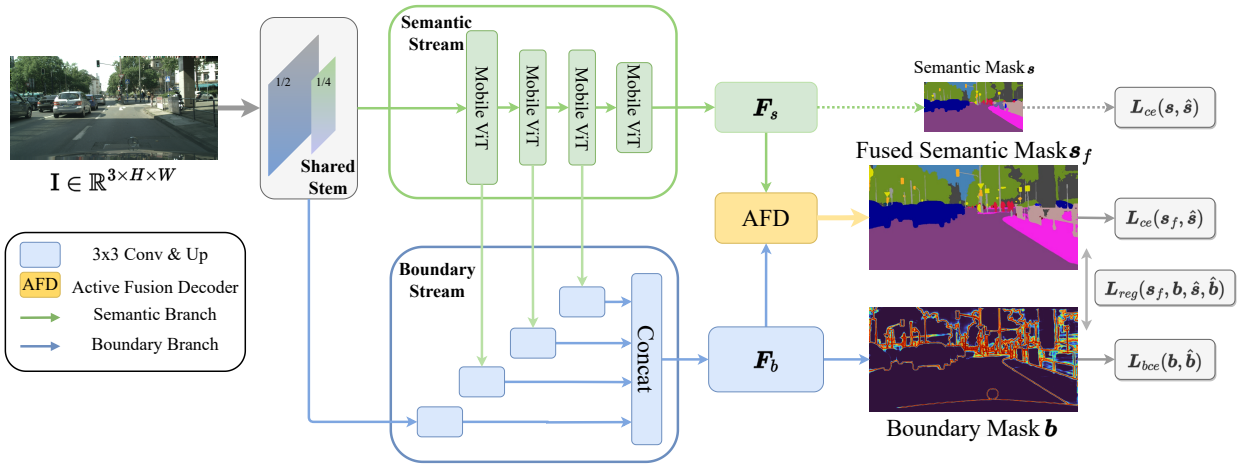


Fig. 3: Workflow of Mobile-Seed, where the semantic stream $\mathcal{S}$ and boundary stream $\mathcal{B}$ extract semantic and boundary features respectively. AFD estimates the relative weights for each channel of semantic features $\boldsymbol{F}_s$ and boundary features $\boldsymbol{F}_b$. An auxiliary classification head is applied to the semantic stream for direct supervision during training. Semantic prediction $\boldsymbol{s}$, fused semantic prediction $\boldsymbol{s}_f$, and boundary prediction $\boldsymbol{b}$ are supervised separately and accordingly. Regularization loss $\mathcal{L}_{reg}$ mitigates the divergences caused by dual-task learning.

CASENet utilizes the bottom layers for details and the top layers for category-aware features. STEAL [30] detects semantic boundaries and corrects noise labels iteratively for crisper prediction. DFF [17] proposed to learn dynamic weights for different input images and locations. RPCNet [7] proposed to jointly learn the semantic and semantic boundary with iterative pyramid context modules. DDS [18] proposed the information converter consisting of several ResNet blocks [29] to mitigate the conflicts caused by deep diverse supervision. However, integrating the information converter into the network will significantly increase the computational burden, especially for high-resolution images.

GSCNN [11], DecoupleNet [12] and BASeg [19] are the most similar approaches to our work. They take binary boundary detection as a supplement for semantic segmentation, which is performed in an auxiliary manner like the auxiliary loss function. However, our approach has significant differences compared to them: (i) Mobile-Seed is a joint boundary detection and semantic segmentation framework instead of a boundary-auxiliary semantic segmentation; (ii) we focus on designing a lightweight model with the least computational burden in contrast to previous cumbersome models.

III. Mobile-Seed Overview

In this section, we present the lightweight Mobile-Seed for joint semantic segmentation and boundary detection learning. As illustrated in Fig. 3, the Mobile-Seed contains a two-stream encoder for semantic segmentation and boundary detection, and then an active fusion decoder (AFD) for features fusion. Each branch's output is supervised with the corresponding ground truth. Moreover, regularization loss is introduced to direct dual-task learning in a complementary way.

A. Architecture Overview

Since the goal is to learn the semantic and boundary information simultaneously, we propose a two-stream encoder to capture the corresponding features from the input image. Firstly, a simple shared stem module consisting of two MobileNetV2 blocks [31] is utilized to embed the original image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ into high-dimension feature space, where H and W mean the height and width of image $\mathbf{I}$ respectively. The semantic stream $\mathcal{S}$ takes the second embed feature map as input and generates semantic-rich features. We emphasize that the semantic stream could be any lightweight semantic segmentation backbone, e.g., [1], [2], [10], [15], [24], [25], [32]. In this paper, we select one of the most recent SOTA methods, AFFormer-T [1] ('T' means the tiny model of AFFormer) as our semantic stream backbone. A simple classification head is used to generate the auxiliary semantic map $\boldsymbol{s}$ during training.

The boundary stream $\mathcal{B}$ takes the first embedded feature map and intermediate feature maps of the semantic stream as input, and feeds into a 3×3 convolution layer, group normalization layer and ReLU layer to differentiate the semantic features to boundary features. Let m denote the stage number and $i \in \{1, 2, ..., m\}$ denote the running index, the i-th stage's representation of the semantic stream is denoted as $\boldsymbol{F}_s^i$. For the i-th location, the information conversion process in the boundary stream is denoted as:

$$\boldsymbol{F}_b^i = \sigma(C_{3\times3}(\boldsymbol{F}_s^i)), \tag{1}$$

where $\boldsymbol{F}_b^i$ means boundary feature of the i-th stage, $C_{3\times3}$ means normalized 3×3 convolution layer, and σ means activation operation. Then, the multi-scale boundary features are upsampled with bilinear interpolation and concatenated together. Finally, a simple classification head is applied for predicting boundary map $\boldsymbol{b} \in \mathbb{R}^{H\times W}$, as shown in Fig. 4.

B. Active fusion Decoder

After obtaining high-dimensional semantic and boundary features, the ensuing problem is how to efficiently fuse features from different domains. As the semantic stream is supervised to learn category-aware semantics and the boundary stream is supervised to learn category-agnostic boundaries, there is a significant domain divergence in two types of features. Most previous approaches use addition [10], [21] or concatenation [12] to fuse features from multiple scales or streams, and some others introduce atrous spatial pyramid pooling

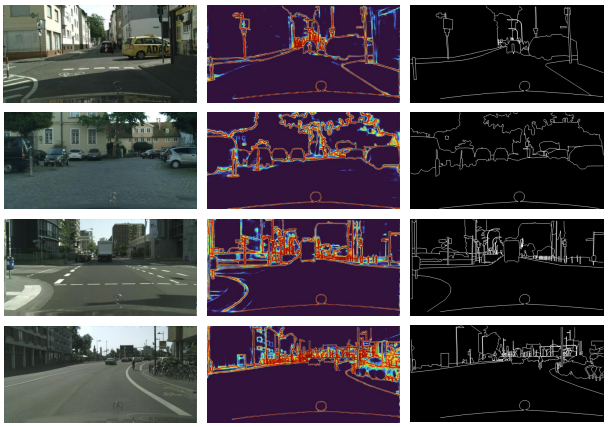


Image Boundary Prediction Ground Truth

Fig. 4: Examples of boundary maps from the boundary stream. The first column shows the input, the second shows the boundary predictions, and the last column shows the ground-truth boundaries.

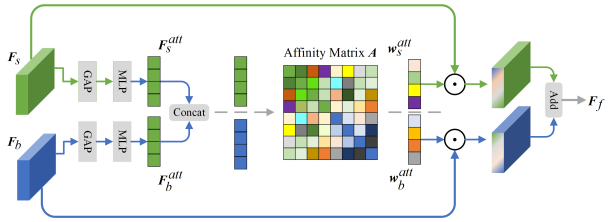


Fig. 5: Illustration of the proposed AFD.

(ASPP) [11] or pyramid context module (PCM) [7] for well-mixed in spatial dimension. The above methods could be classified as fixed weights methods, where the fusion weights in the channel dimension are image-independent. However, the importance of each channel in semantic features and boundary features may vary for different images. Therefore, the fusion weights should be conditioned on the input. There are dynamic fusion methods [17], [19] that can adapt the weights for semantic edge detection and semantic segmentation tasks. However, calculating fusion weights in both spatial and channel dimensions is still too cumbersome for the lightweight framework.

To tackle this issue, we propose the active fusion decoder (AFD) module to learn the fusion weights for semantic stream and boundary stream, as shown in Fig. 5. With the semantic feature map $\boldsymbol{F}_s \in \mathbb{R}^{C\times H\times W}$ from semantic stream and boundary feature map $\boldsymbol{F}_b \in \mathbb{R}^{C\times H\times W}$ from boundary stream, we first calculate the semantic channel-wise attention vector $\boldsymbol{F}_s^{att} \in \mathbb{R}^{C\times 1\times 1}$ and boundary channel-wise attention vector $\boldsymbol{F}_b^{att} \in \mathbb{R}^{C\times 1\times 1}$ with global average pooling (GAP):

$$\begin{aligned} \boldsymbol{F}_s^{att} &= f_s(GAP(\boldsymbol{F}_s)), \\ \boldsymbol{F}_b^{att} &= f_b(GAP(\boldsymbol{F}_b)), \end{aligned} \tag{2}$$

where $\boldsymbol{F}_s^{att}, \boldsymbol{F}_b^{att} \in \mathbb{R}^{C\times 1\times 1}$, and $f(\cdot)$ means multi-layer perception (MLP) modules. We stack attention vectors of semantic and boundary features in channel dimension:

$$\boldsymbol{F}_f^{att} = (\boldsymbol{F}_s^{att} \| \boldsymbol{F}_b^{att}), \tag{3}$$

where $\|$ means channel-wise concatenation operation. To metric the affinity among each channel of semantic and boundary features, we split the fusing attention vector $\boldsymbol{F}_f^{att} \in \mathbb{R}^{2C}$ into H groups and generate the channel-wise affinity matrix $\boldsymbol{A} \in \mathbb{R}^{2C/H\times 2C/H}$. We first calculate the query vector $\boldsymbol{q}$, key vector $\boldsymbol{k}$ and value vector $\boldsymbol{v}$ from $\boldsymbol{F}_f^{att}$ by linear projection, and then each component of affinity matrix $\boldsymbol{A}_{i,j}$ is computed as:

$$\boldsymbol{A}_{i,j} = e^{\boldsymbol{q}_i \boldsymbol{k}_j{}^\top} / \sum_{i=1}^{H} e^{\boldsymbol{q}_i}, \tag{4}$$

where $\boldsymbol{q}_i$ is the i-th head of query, and $\boldsymbol{k}_j$ is the j-th head of key. The i-th head of active fusion weight $\boldsymbol{w}$ is calculated with:

$$\boldsymbol{w}_i = \boldsymbol{A}\boldsymbol{v}_i, \tag{5}$$

where $\boldsymbol{v}_i$ means the i-th head of value vector $\boldsymbol{v}$. Then we divide the weight vector $\boldsymbol{w}$ into $\boldsymbol{w}_s$ for semantic stream and $\boldsymbol{w}_b$ for boundary stream by inverse operation of Eq. 3. The fused features $\boldsymbol{F}_f$ is calculated through residual connection:

$$\boldsymbol{F}_f = (1+\boldsymbol{w}_s)\boldsymbol{F}_s + (1+\boldsymbol{w}_b)\boldsymbol{F}_b. \tag{6}$$

A 1×1 convolution layer is followed as classification head to compact the $\boldsymbol{F}_f$ into the final semantic prediction map $\boldsymbol{s}_f \in \mathbb{R}^{N\times H\times W}$, N means the category number. Overall, the AFD estimates the channel-wise relationship within and among semantic features and boundary features for learning optimal fusion weights.

C. Loss Functions

In the Mobile-Seed framework, we jointly train the semantic and boundary stream in an end-to-end way and use the AFD to fuse the dual-task features for final prediction. Therefore, available supervisions include semantic label $\hat{\boldsymbol{s}} \in \mathbb{R}^{N\times H\times W}$, semantic boundary label $\hat{\boldsymbol{b}_s} \in \mathbb{R}^{N\times H\times W}$ and binary boundary label $\hat{\boldsymbol{b}} \in \mathbb{R}^{H\times W}$ (as shown in Fig. 2 (b), (c) and (d)). The cross-entropy (CE) loss and binary cross-entropy (BCE) loss function are used to supervise the semantic and boundary predictions:

$$\mathcal{L}_{cls} = \mathcal{L}_{ce}(\boldsymbol{s}, \hat{\boldsymbol{s}}) + \mathcal{L}_{ce}(\boldsymbol{s}_f, \hat{\boldsymbol{s}}) + \mathcal{L}_{bce}(\boldsymbol{b}, \hat{\boldsymbol{b}}). \tag{7}$$

Dual-task regularization. With the top supervision of semantic label $\hat{\boldsymbol{s}}$, the top layers are acquired to learn abstracted semantic representation, enabling it to cover diverse object shapes, lighting conditions and textures. In contrast, bottom supervision of the boundary label $\hat{\boldsymbol{b}}$ leads the bottom layers to distinct boundaries or non-boundaries, rather than category-aware semantics. Since the bottom layers provide basic representations for both semantic segmentation and boundary detection, the bottom layers receive two distinctively different supervisions under back-propagation. Liu et al. [18] pointed out that applying deep diverse supervision (DDS) directly may lead to conflicts and performance degradation, while a single convolution layer in the boundary stream is too weak to alleviate the supervision conflicts. Our ablation studies in Sec. IV-C also support this finding. To address these conflicts, authors of the DDS introduced buffer blocks to isolate the backbone and side layers. Unlike that, we design bi-directional consistency loss $\mathcal{L}_{reg}$ consisting of the semantic-to-boundary consistency loss $\mathcal{L}_{s2b}$ and the boundary-to-semantic consistency loss $\mathcal{L}_{b2s}$ to soften the conflict and free computation burden during

inference. The semantic-to-boundary consistency loss $\mathcal{L}_{s2b}$ is designed to align pseudo semantic boundary $\boldsymbol{b}_{ps}$ generated from semantic prediction $\boldsymbol{s}_f$ with the semantic boundary label $\hat{\boldsymbol{b}}_s$. We introduce a filtering template $\boldsymbol{T} \in \mathbb{R}^{N\times(2r+1)\times(2r+1)}$ to look into neighbors of each pixel and seek for the maximum difference in each category, where r is the search radius. For ease of description, we set $r = 2$ here and each channel of the filtering template $\boldsymbol{T}$ is:

$$\boldsymbol{T}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, i \in 1, 2, ..., N. \tag{8}$$

We slide the template on the fused semantic prediction map $\boldsymbol{s}_f \in \mathbb{R}^{N\times H\times W}$ and select the max absolute value in the filtering window as the pseudo semantic boundary prediction:

$$\boldsymbol{b}_{ps} = \max_{\boldsymbol{T}}(\|\boldsymbol{T} \circledast \boldsymbol{s}_f\|). \tag{9}$$

Intuitively, the template $\boldsymbol{T}$ mimics the generation process of the semantic boundary label $\hat{\boldsymbol{b}}_s$ by checking whether neighboring pixels have different labels or not. Mean absolute loss is used to supervise the pseudo semantic boundaries:

$$\mathcal{L}_{s2b} = \|\boldsymbol{b}_{ps} - \hat{\boldsymbol{b}_s}\|. \tag{10}$$

On the other hand, boundary prediction provides important prior knowledge to ensure semantic consistency between body and boundary. We combine weighted cross-entropy loss with boundary prior to formulating the boundary-to-semantic consistency loss $\mathcal{L}_{b2s}$:

$$\mathcal{L}_{b2s} = \sum_{k,p} \mathbb{1}_{\boldsymbol{b},\hat{\boldsymbol{b}}}\big(\hat{\boldsymbol{s}}(k,p)\log(\boldsymbol{s}_f(k,p))\big), \tag{11}$$

where k and p walk over the categories and pixels. $\mathbb{1}_{\boldsymbol{b},\hat{\boldsymbol{b}}} = \{1 : \boldsymbol{b} > \epsilon \cup \hat{\boldsymbol{b}} = 1\}$ marks ground-truth (GT) pixels and high confidence pixels on the boundary prediction map $\boldsymbol{b}$. ϵ is the confidence threshold and we use 0.8 in the experiments. With the bi-direction consistency losses $\mathcal{L}_{s2b}$ and $\mathcal{L}_{b2s}$, the regularization function can be formulated as :

$$\mathcal{L}_{reg} = \mathcal{L}_{s2b} + \mathcal{L}_{b2s}. \tag{12}$$

The total loss function is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{reg}, \tag{13}$$

where λ_1, λ_2 are hyper-parameters to control weights of classification loss and dual-task regularization.

IV. Experimental Evaluation

In this section, we conduct experiments on three publicly available datasets: Cityscapes [33], CamVid [34] and PASCAL Context [35] to show the capability of our method in various environments. Sec. IV-A introduces the implementation details and evaluation metrics. In Sec. IV-B, we compare our method with SOTA methods on the Cityscapes dataset and extensively validate on CamVid and PASCAL Context to demonstrate the generalization and robustness. In Sec. IV-C, ablation studies on the AFD and regularization loss demonstrate the effectiveness of our design. Overall, the results prove that our approach could (i) jointly learn the semantic segmentation and boundary detection tasks; (ii) improve the semantic segmentation performance while maintaining online operation; (iii) accurately detect object boundaries in complex scenes.

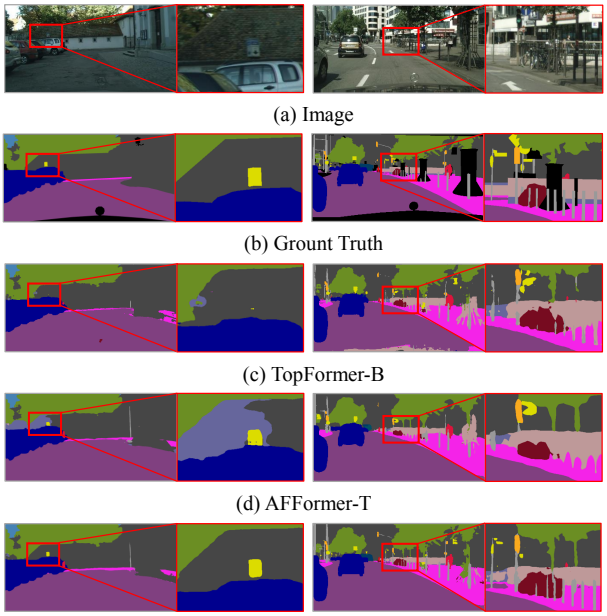
A. Experimental Setup

Implementation details. We build our model based on the MMsegmentation toolbox. All experiments were performed on an NVIDIA RTX 4090 GPU. We select the AFFormer-T [1] as our semantic stream and pre-train on ImageNet-1k [36], while the boundary stream learns from scratch. Most training details follow previous approaches [1], [10]. The hyperparameters for controling loss weight are set to $\lambda_1 = \lambda_2 = 1$. We use the AdamW optimizer [37] for all datasets to update model parameters. The data augmentation methods include random resize, random scaling, random horizontal flipping and color jittering. The training iterations, batch size and input image size for Cityscapes, CamVid and PASCAL Context datasets are set to [160K, 8, 1024×1024], [20K, 16, 520×520], [80K, 16, 480×480] respectively. For more training details, please refer to our open-source code.

Evaluation metrics. We report three quantitative measures to evaluate the performance of our method. (i) We evaluate the semantic segmentation results with the widely used Intersection over Union (IoU) metric. (ii) To evaluate our purpose that the Mobile-Seed extracts high-quality semantic boundary, we use the F-score as previous approaches [11], [12] on the Cityscapes *val* dataset. This boundary metric computes the F1 score between dilated semantic boundary prediction and ground truth with a threshold to control the bias degree. We set thresholds 0.00088, 0.001875, 0.00375, and 0.005, which correspond to 3, 5, 9, and 12 pixels respectively. (iii) The boundary IoU (BIoU) [38] is introduced to further evaluate both the semantic boundary and binary boundary performance on various datasets. Compared with the F-score, the BIoU is more sensitive to small object errors. For efficiency analysis, we report the FLOPs, params number and FPS evaluated on an RTX 2080 Ti GPU with batch size of 1. For a fair comparison, inferences are conducted on the origin image resolution instead of multi-scale inference.

B. Quantatitave and Qualitative Results

The comparison of the semantic segmentation results with SOTA methods on the Cityscapes *val* dataset is shown in Tab. I. As can be seen, the Mobile-Seed owns fewer parameters, lower computation costs, and higher mIoU performance than AFFormer-B, validating that our two-stream design achieves a better balance of accuracy and efficiency. Tab. II shows the category-wise comparison in terms of IoU with our strong baseline method AFFormer-T. Our method significantly outperforms the baseline method in most categories (18/19), improving the mIoU score from 76.2 to 78.4 (2.2% improvement) over the strong baseline. Moreover, our method could still keep near real-time (23.9 FPS) inference speed. The qualitative results are shown in Fig. 6, with additional results available on the project page.



(a) Image

(b) Grount Truth

(c) TopFormer-B

(d) AFFormer-T

(e) Mobile-Seed

Fig. 6: Qualitative comparisons for semantic segmentation. The "unlabeled" area is rendered as black in ground truth.

To demonstrate that Mobile-Seed achieves a more precise boundary location, we evaluate the semantic boundary accuracy with the F-score metric reported in Tab. III. It shows that our method outperforms the baseline method by a large margin, especially in the strictest condition with the 3px threshold (about 4.2% improvement in the F-score metric). The results of semantic boundary validate that jointly learning semantic segmentation and boundary detection boosts the segmentation performance in boundary areas. The qualitative results of the semantic boundary in Fig. 7 show that our method predicts sharper and more continuous boundaries.

Lastly, we report the semantic boundary and binary boundary performance with the BIoU metric. Fig. 8 (a) shows the Mobile-Seed achieves higher BIoU value under several thresholds. As the baseline method is a semantic segmentation framework and has no boundary stream, we retrained it with the binary boundary $\hat{b}$ supervision (called AFF-T-B in the following). The BIoU scores of Mobile-Seed and AFF-T-B shown in Fig. 8 (b) demonstrate that our framework extracts crisper and more accurate boundaries than independently learning object boundaries.

We additionally visualize the activation maps of each stage from the boundary stream in Fig. 9, illustrating that the lower stages are interested in sharp intensity change (Fig. 9 (b), (c)) and higher stages (Fig. 9 (d), (e)) focus on semantic inconsistency. Intuitively, the bottom layers capture low-level details and the top layers obtain high-level semantics, and in the end, the boundary stream head adaptively combines multi-level features for boundary prediction.·

Extensive validations. Furthermore, we evaluate our method on the CamVid and PASCAL Context datasets. We retrain the baseline AFFormer-T and report the segmentation and boundary result in terms of IoU and BIoU respectively. Quantitative results in Tab. IV show that our method significantly improves semantic segmentation accuracy in various datasets, demonstrating the generalization ability.

TABLE I: Semantic segmentation results on Cityscapes *val* dataset. LRFormer-T*: code of LRFormer is not available.

Method	#Params	FLOPs	mIoU	FPS
FCN [21]	9.8M	317G	61.5	11.2
PSPNet [22]	13.7M	423G	70.2	9.5
DeepLabV3+ [23]	15.4M	555G	75.2	8.2
SegFormer-B0 [2]	3.8M	125G	76.2	11.7
TopFormer-B [10]	5.1M	11.2G	75.2	55.6
PIDNet-S [32]	7.6M	47.6G	78.7	15.3
LRFormer-T* [39]	13.0M	122.0G	80.7	-
AFFormer-T [1]	2.2M	23.6G	76.2	27.8
AFFormer-B [1]	3.0M	33.5G	77.8	21.2
Mobile-Seed(Ours)	2.4M	31.6G	78.4	23.9

C. Ablation Studies and Insights

Effectiveness of dual-task learning framework. We conduct an ablation study to demonstrate thaxt our dual-task learning framework is better than learning semantic segmentation task individually, as shown in Tab. V. This ablation experiment employs the single semantic stream $\mathcal{S}$ as the baseline [A] and test boundary stream $\mathcal{B}$, boundary loss function $\mathcal{L}_b$ and dual-task regularization loss $\mathcal{L}_{reg}$, respectively. [B] shows that adding "multi-scale" features from the boundary stream boosts the semantic segmentation performance. We do not refer to the "multi-scale" features as "boundary" features because the boundary supervision is removed. [C] shows that explicitly supervising the boundary stream with the $\mathcal{L}_b$ leads to performance degradation, as the mIoU drops about 0.7%. This circumstance proves our suppose that applying distinctive supervision to different modules may harm the framework. [D] shows that our dual-task regularization loss $\mathcal{L}_{reg}$ could mitigate the learning divergence and promote the semantic segmentation and boundary detection tasks learning in a complementary way.

Comparison of feature fusion methods. We conduct ablation studies to prove the effectiveness of our AFD. We take the semantic stream $\mathcal{S}$, boundary stream $\mathcal{B}$ and total loss $\mathcal{L}$ as baseline and check out the influence of feature fusion methods. We compare our AFD with fixed weight fusion methods, as addition and concatenation in previous approaches. Tab. VI shows that our AFD is more lightweight than concatenation and achieves better performance compared to both addition and concatenation. The results support our assumption that the fusion weights should be conditioned on the input, where our AFD dynamically assigns proper weights to each channel of semantic and boundary features and outperforms addition and concatenation.

V. Conclusion

In this paper, we present a novel lightweight framework Mobile-Seed for joint semantic segmentation and boundary detection. Our method consists of a two-stream encoder and an active fusion decoder (AFD), where the encoder extracts semantic and boundary features respectively, and the AFD assigns dynamic fusion weights for two kinds of features. Moreover, regularization loss is introduced to alleviate the divergence in dual-task learning. We have implemented and evaluated our approach on various datasets and provided comparisons to other existing techniques. The experimental results validate that our method outperforms all the existing

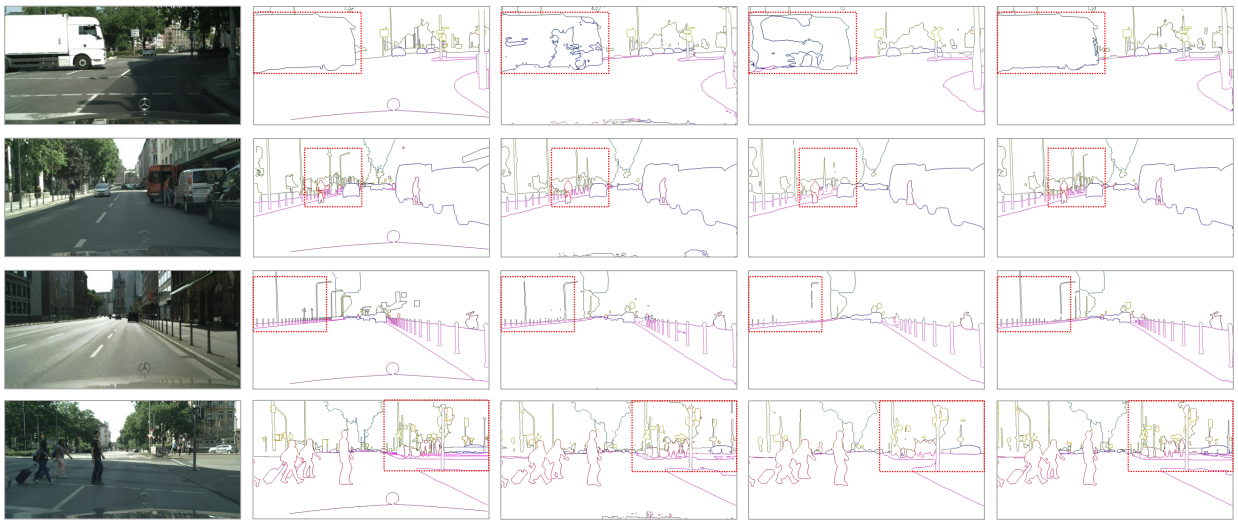


Image Ground Truth TopFormer-B AFFormer-T Mobile-Seed

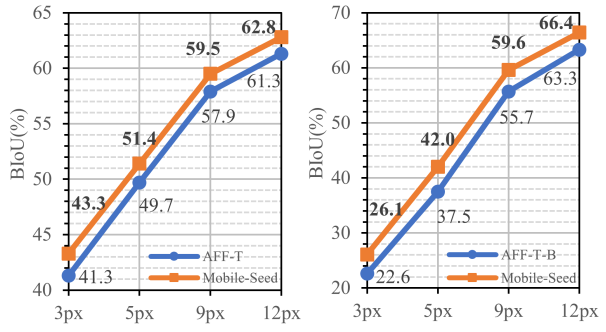
Fig. 7: Qualitative results of the semantic boundary.

TABLE II: Comparison class-aware semantic segmentation results to the baseline method. AFF-T is short of AFFormer-T.

mIoU	road	s.walk	build	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mean
AFF-T	98.2	85.3	92.5	54.7	57.6	63.9	70.1	78.5	92.7	**66.3**	94.8	81.1	60.0	94.7	70.2	80.0	69.5	61.7	75.9	76.2
Ours	**98.3**	**85.9**	**92.8**	**61.8**	**58.7**	**66.7**	**71.6**	**79.6**	**92.9**	65.9	**95.1**	**82.0**	**61.4**	**94.9**	**78.9**	**85.8**	**77.9**	**63.0**	**77.5**	**78.4**

TABLE III: Quantatitave results of semantic boundary on the Cityscapes *val* dataset. AFF-T is short of AFFormer-T.

Thrs	Method	road	s.walk	build	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mean
3px	AFF-T	81.3	61.0	66.1	47.8	47.4	66.1	65.5	66.9	68.2	53.6	78.8	57.6	65.2	74.7	77.7	85.1	91.3	**77.5**	61.2	68.0
	Ours	**84.2**	**66.8**	**72.0**	**57.0**	**53.4**	**73.6**	**68.5**	**70.1**	**73.8**	**59.5**	**82.2**	**61.1**	**66.5**	**79.3**	**81.0**	**88.2**	**95.2**	76.8	**62.6**	**72.2**
5px	AFF-T	86.9	70.1	75.2	50.5	50.1	72.9	71.8	74.9	78.6	57.6	85.5	65.7	69.9	82.7	78.8	86.3	91.6	**78.8**	68.7	73.5
	Ours	**88.6**	**74.2**	**80.0**	**59.4**	**56.2**	**78.6**	**74.0**	**76.4**	**82.8**	**63.0**	**87.9**	**68.2**	**71.2**	**85.7**	**81.8**	**89.3**	**95.6**	78.4	**69.8**	**76.9**
9px	AFF-T	90.7	76.8	82.5	53.2	53.0	77.1	76.4	79.8	86.5	61.3	89.1	71.7	74.2	87.9	79.8	87.6	91.9	**80.0**	75.8	77.6
	Ours	**91.5**	**79.4**	**86.1**	**61.8**	**59.0**	**82.0**	**77.4**	**80.3**	**89.0**	**66.3**	**90.9**	**73.5**	**75.4**	**89.9**	**82.6**	**90.3**	**95.9**	79.8	**75.9**	**80.4**
12px	AFF-T	91.9	79.0	85.1	54.3	54.3	78.8	77.6	81.4	89.1	62.6	90.3	73.8	75.8	89.7	80.4	88.1	92.0	**80.6**	78.4	79.1
	Ours	**92.4**	**81.4**	**88.3**	**62.9**	**60.3**	**83.3**	**78.4**	**81.6**	**91.1**	**67.6**	**91.8**	**75.3**	**77.0**	**91.3**	**83.0**	**90.6**	**96.0**	80.4	**78.1**	**81.6**



(a) Semantic Boundary (b) Binary Boundary

Fig. 8: (a) Semantic boundary results and (b) Binary boundary results on the Cityscapes *val* dataset. AFF-T-B means AFFormer-T for the binary boundary detection task.

TABLE IV: Comparison with baseline method on CamVid and PASCAL Context datasets. PASCAL^{59} and PASCAL^{60} mean PASCAL Context dataset with 59 and 60 categories, respectively. The threshold of BIoU is set to 3px.

Method	CamVid		PASCAL^{59}		PASCAL^{60}	
	mIoU	BIoU	mIoU	BIoU	mIoU	BIoU
AFFormer-T	71.6	41.2	45.7	20.7	41.4	14.9
Mobile-Seed	**73.4**	**45.2**	**47.2**	**22.1**	**43.0**	**16.2**

methods and support all claims made in this paper. We believe that the Mobile-Seed can be deployed on lightweight robotics platforms and serves for semantic SLAM, robot manipulation and other downstream tasks.

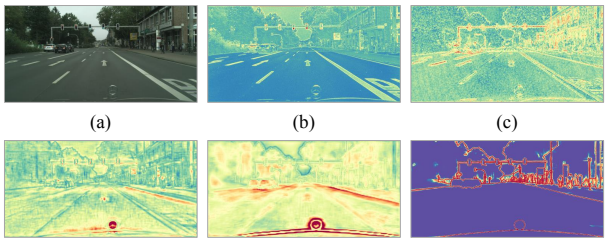


Fig. 9: Visulization of multi-scale activation maps in boundary stream. (a) input image. (b) stage I. (c) stage II. (d) stage III. (e) stage IV. (f) final prediction.

TABLE V: Ablation study on our dual-task learning framework. $\mathcal{S}$ means semantic stream, and $\mathcal{B}$ means boundary stream. $\mathcal{L}_b$ means boundary loss and $\mathcal{L}_{reg}$ means dual-task regularization loss. The threshold of BIoU metric is set to 3px.

	$\mathcal{S}$	$\mathcal{B}$	$\mathcal{L}_b$	$\mathcal{L}_{reg}$	mIoU	BIoU
$[A]$	✓				76.2	41.3
$[B]$	✓	✓			77.7	42.1
$[C]$	✓	✓	✓		76.9	41.6
$[D]$	✓	✓	✓	✓	**78.4**	**43.3**

REFERENCES

[1] D. Bo, W. Pichao, and F. Wang, "Afformer: Head-free lightweight semantic segmentation with linear transformer," in *Proc. of the Conf. on Advancements of Artificial Intelligence (AAAI)*, 2023.

[2] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with

TABLE VI: Ablation study on feature fusion methods. 'ADD' means features addition, 'CAT' means features concatenation and 'AFD' means active fusion decoder proposed in our method.

ADD	CAT	AFD	mIoU	FLOPs	FPS
✓			77.7	0.96G	**24.3**
	✓		78.0	1.91G	23.6
		✓	**78.4**	**0.96G**	23.9

transformers," *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, vol. 34, pp. 12 077–12 090, 2021.
[3] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss, "OverlapNet: Loop Closing for LiDAR-based SLAM," in *Proc. of Robotics: Science and Systems (RSS)*, 2020.
[4] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet, "Image-to-lidar self-supervised distillation for autonomous driving data," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9891–9901.
[5] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 525–11 533.
[6] Y. Liao, J. Li, S. Kang, Q. Li, G. Zhu, S. Yuan, Z. Dong, and B. Yang, "Se-calib: Semantic edges based lidar-camera boresight online calibration in urban scenes," *IEEE Trans. on Geoscience and Remote Sensing*, 2023.
[7] M. Zhen, J. Wang, L. Zhou, S. Li, T. Shen, J. Shang, T. Fang, and L. Quan, "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 666–13 675.
[8] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Proc. of Robotics: Science and Systems (RSS)*, 2014, pp. 1–9.
[9] X. Chen, A. Milioto, E. Palazzolo, P. Giguere, J. Behley, and C. Stachniss, "Suma++: Efficient lidar-based semantic slam," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 4530–4537.
[10] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, "Topformer: Token pyramid transformer for mobile semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 083–12 093.
[11] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5229–5238.
[12] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020, pp. 435–452.
[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2021.
[14] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2022.
[15] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2023.
[16] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "Casenet: Deep category-aware semantic edge detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5964–5973.
[17] Y. Hu, Y. Chen, X. Li, and J. Feng, "Dynamic feature fusion for semantic edge detection," in *Proc. of the Intl. Conf. on Artificial Intelligence (IJCAI)*, 2019.
[18] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," *Intl. Journal of Computer Vision (IJCV)*, vol. 130, no. 1, pp. 179–198, 2022.
[19] X. Xiao, Y. Zhao, F. Zhang, B. Luo, L. Yu, B. Chen, and C. Yang,
[20] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3000–3009.
"Baseg: Boundary aware semantic segmentation for autonomous driving," *Neural Networks*, vol. 157, pp. 460–470, 2023.
[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018, pp. 801–818.
[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.
[26] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," in *Proc. of British Machine Vision Conference (BMVC)*, 2019.
[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.
[28] S. Tang, T. Sun, J. Peng, G. Chen, Y. Hao, M. Lin, Z. Xiao, J. You, and Y. Liu, "Pp-mobileseg: Explore the fast and accurate semantic segmentation model on mobile devices," *arXiv preprint arXiv:2304.05152*, 2023.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[30] D. Acuna, A. Kar, and S. Fidler, "Devil is in the edges: Learning semantic boundaries from noisy annotations," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 075–11 083.
[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
[32] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 529–19 539.
[33] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.
[34] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.
[35] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 891–898.
[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2017.
[38] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary iou: Improving object-centric image segmentation evaluation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 334–15 342.
[39] Y.-H. Wu, S.-C. Zhang, Y. Liu, L. Zhang, X. Zhan, D. Zhou, J. Feng, M.-M. Cheng, and L. Zhen, "Low-resolution self-attention for semantic segmentation," *arXiv preprint arXiv:2310.05026*, 2023.