Revisiting Efficient Semantic Segmentation: Learning Offsets for Better Spatial and Class Feature Alignment

Shi-Chen Zhang[1] Yunheng Li[1] Yu-Huan Wu[2] Qibin Hou[1†] Ming-Ming Cheng[1]
[1]VCIP, CS, Nankai University [2]IHPC, A*STAR, Singapore
zhangshichen@mail.nankai.edu.cn

Project page: https://github.com/HVision-NKU/OffSeg.

Abstract

Semantic segmentation is fundamental to vision systems requiring pixel-level scene understanding, yet deploying it on resource-constrained devices demands efficient architectures. Although existing methods achieve real-time inference through lightweight designs, we reveal their inherent limitation: misalignment between class representations and image features caused by a per-pixel classification paradigm. With experimental analysis, we find that this paradigm results in a highly challenging assumption for efficient scenarios: Image pixel features should not vary for the same category in different images. To address this dilemma, we propose a coupled dual-branch offset learning paradigm that explicitly learns feature and class offsets to dynamically refine both class representations and spatial image features. Based on the proposed paradigm, we construct an efficient semantic segmentation network, OffSeg. Notably, the offset learning paradigm can be adopted to existing methods with no additional architectural changes. Extensive experiments on four datasets, including ADE20K, Cityscapes, COCO-Stuff-164K, and Pascal Context, demonstrate consistent improvements with negligible parameters.

1. Introduction

Semantic segmentation, which aims to assign category labels to every image pixel, plays a vital role in computer vision applications [13, 15, 20, 25, 26, 28, 36, 48, 49, 63]. While recent advances in standard models [29, 30, 38, 39, 56, 60, 62] have achieved remarkable segmentation accuracy, their computational and parametric complexity renders them impractical for resource-constrained scenarios. This has spurred significant interest in efficient semantic segmentation models [18, 34, 37, 41, 46, 50], which prioritize real-time inference and minimal parameters.

†Corresponding Author. Email: houqb@nankai.edu.cn

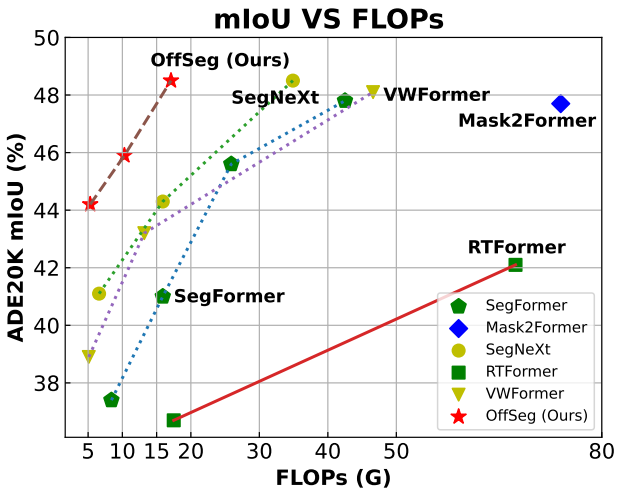

Figure 1. Comparisons with popular efficient segmentation methods on the ADE20K [61] dataset. We can see from the figure that our method achieves the best trade-off between performance and computations.

Conventional segmentation frameworks typically use high-dimensional image features, rich class representations, and a large number of parameters, which together yield superior performance compared to more compact architectures. This phenomenon aligns with the neural scaling law, where model capacity positively correlates with segmentation accuracy until it is able to reach computational resource constraints. In contrast, efficiency-oriented architectures [43, 46, 50, 51] face an inherent trade-off: aggressive model compression weakens their capacity to align category semantics with localized visual cues. This misalignment results in blurred object boundaries, missed small instances, and inconsistent predictions, all of which are further intensified by the prevailing per-pixel classification paradigm (see Fig. 2(a)). While existing works employ lightweight backbones [18, 23, 41] or spatial downsampling [22, 45, 58] to achieve efficient performance, they largely overlook the

Table 1. Comparison of different semantic segmentation paradigms. 'Fea.' and 'Rep.' donate feature adaptation and class representation adaptation, respectively.

Paradigm	Fea.	Rep.	Interaction	Alignment	Overhead
Per-Pixel Classification	✗	✗	✗	Static unidirectional	Matrix multiplication
Mask Classification	✗	✓	Cross-attention	Dynamic but asymmetric	Transformer decoder
Offset Learning	✓	✓	Dual-decoupled offsets	Elastic bidirectional	Matrix multiplication

fundamental challenge of jointly refining category and feature representations under strict parametric constraints.

To uncover the fundamental issues inherent in the per-pixel classification paradigm, we employ the ideal class representation (feature) mining method to derive optimal class-specific representations for individual images. Statistical analysis (Fig. 3) reveals that the similarity between optimal class representations of the same category across different images is remarkably low. This finding shows that using fixed class representations for all images, as in the per-pixel classification approach, is suboptimal since it fails to adapt to the unique image features and class-specific details in each image.

Based on the observations of the fundamental challenge, we propose an offset learning paradigm, a novel segmentation method that can explicitly learn and rectify the deviation between class representations and image features through learnable feature offsets (FOs) and class offsets (COs). Our key insight is that, while efficient models lack sufficient parameters to model ideal category-feature relationships, they can effectively learn to predict the offset between initially coarse representations and their optimal counterparts. Specifically, our offset learning paradigm consists of two primary branches: the Class Offset Learning branch and the Feature Offset Learning branch. These two branches are designed to learn COs and FOs, respectively, enabling flexibility of both image features and class representations.

As shown in Fig. 2(b), in addition to the per-pixel segmentation paradigm, there exists a mask-based segmentation paradigm [4, 10, 11], which employs cross-attention to facilitate interaction between learnable queries and image features. This approach enables queries to learn image-specific characteristics adaptively. However, it has two inherent limitations: (1) It only adjusts the queries while leaving the image features static, and (2) the cross-attention introduces significant computational overhead. As summarized in Tab. 1, our method distinguishes itself from these two paradigms through two key advantages: (1) the dual adaptability of both image features and class representations, and (2) negligible interaction overhead.

Based on our proposed offset learning paradigm (Fig. 2(c)), we design a straightforward segmentation network named OffSeg, which consists solely of a backbone and a pixel decoder. As a plug-and-play paradigm, we apply our framework to SegNeXt [18] (CNN-based), SegFormer [46] (Transformer-based), and Mask2Former [11] (mask classification) to demonstrate its effectiveness and flexibility. Extensive experiments across four benchmark datasets show that the results consistently validate the efficiency and effectiveness of our method. Performance improvements achieved across different architectures and datasets highlight the robustness and generalizability of our approach. In Fig. 1, we present the performance of our model across different scales. The results demonstrate that our OffSeg achieves a superior balance between performance and computational efficiency.

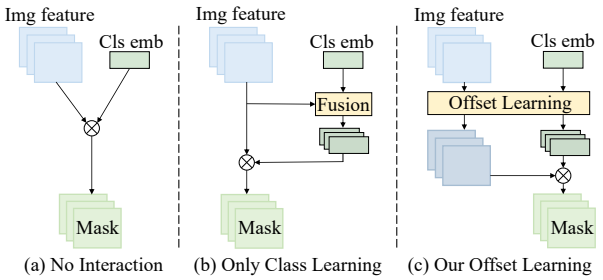

Figure 2. Visual comparison of different semantic segmentation paradigms. From left to right, the figures sequentially illustrate per-pixel classification, mask classification, and our proposed offset learning paradigm.

To sum up, our main contributions can be summarized as follows:

- We identify the core limitation of per-pixel segmentation through statistical analysis and ideal class representation (feature) mining, exposing the intrinsic misalignment between static image features and class representations.
- We propose a parameter-efficient offset learning paradigm with dual branches that jointly adapt image features and class representations with nearly negligible computational overhead.
- Extensive experiments demonstrate superior performance of our proposed OffSeg and effectiveness over previous per-pixel classification (CNN, Transformer) and mask classification paradigms.

2. Related Work

2.1. Traditional Semantic Segmentation

Semantic segmentation has witnessed significant advancements through large-scale models that prioritize accuracy over computational efficiency. Pioneering works like fully convolutional networks (FCN) [31] established the foundation by replacing fully connected layers with convolutional operations, enabling dense pixel-wise predictions. With this paradigm established, subsequent CNN-based works [1, 17, 27, 36, 42, 52, 54, 57, 59] have enhanced FCN from various perspectives. For example, U-Net [36] further enhances feature localization through symmetric encoder-decoder structures and skip connections. From the perspective of context aggregation, DeepLab series [6–9] employ atrous spatial pyramid pooling (ASPP) to capture multi-scale contextual information. PSPNet [59] proposes pyramid pooling modules to aggregate global context across different sub-regions. Benefiting from the success of attention mechanisms [14, 40], Transformer-based approaches [21, 35, 38, 55, 60] have achieved remarkable results. For instance, SERE [60] redefines semantic segmentation as a sequence-to-sequence prediction task, leveraging global self-attention to model full-image context. Unlike the per-pixel classification paradigm, MaskFormer series [10, 11] introduces a mask classification paradigm, where learnable queries interact with image features through a transformer decoder.

2.2. Efficient Semantic Segmentation

While traditional models achieve high segmentation accuracy, their computational demands hinder real-time applications, driving the development of efficient semantic segmentation works [4, 16, 18, 37, 41, 43–46, 50, 51, 58]. From the perspective of the backbone network, SegFormer [46] proposes a lightweight and hierarchically structured transformer encoder, while SegNeXt [18] proposes a more effective convolutional attention solely through multi-scale convolutions to construct an efficient backbone. LRFormer [45] introduces a highly-efficient transformer with linear attention, which is computed in a very low resolution space. For the decoder, FeedFormer [37] employs a transformer that treats image features as queries to extract structural information. VWFormer [47] augments multi-scale representations by interacting with multiple windows of different scales through cross-attention. CGRSeg [34] utilizes pyramid context-guided spatial feature reconstruction to enhance the ability of foreground objects representation from both horizontal and vertical dimensions.

For efficient segmentation models, since the mask classification paradigm requires a computationally heavy transformer decoder for feature interaction, they all adopt the per-pixel classification segmentation paradigm. Although per-pixel classification incurs little computational overhead, it inherently suffers from the misalignment between image features and class representations. The dilemma becomes more pronounced in lightweight scenarios (Fig. 3), which motivates us to develop a segmentation paradigm tailored for efficient semantic segmentation.

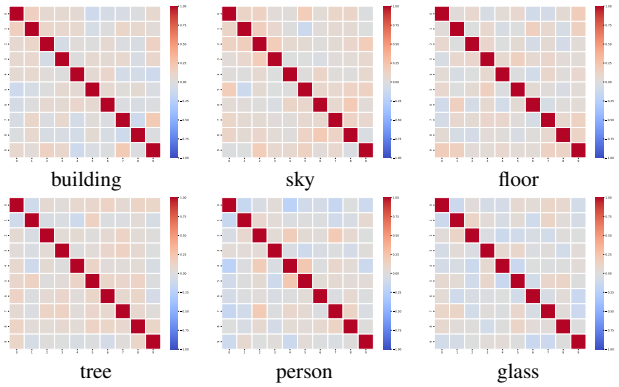

Figure 3. Heatmap visualizations of the ideal class representations similarity. We can observe that the correlations between different ideal class representations of the same category are very low.

3. Method

3.1. Revisiting Per-pixel Classification

Per-pixel classification, the cornerstone of conventional semantic segmentation, independently assigns labels to each pixel by comparing its feature vector with predefined category prototypes. Traditional per-pixel classification maps pixel embeddings $\mathbf{E} \in \mathbb{R}^{HW \times C}$ to class scores $\mathbf{P}$ via a 1×1 convolution:

$$\mathbf{P}_{i,j} = \mathbf{W}_c \cdot \mathbf{E}^{\top}{}_{i,j}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{K \times C}$ is learnable parameter for K classes, c is the class index. This paradigm treats each pixel independently, ignoring contextual correlations.

While widely adopted, this paradigm suffers from two critical issues in efficient segmentation scenarios. First, the per-pixel classification paradigm relies on fixed class representations to categorize pixels. Second, this approach assumes that the network can learn identical features for the same category across different images. However, our subsequent analysis in §3.2 demonstrates that this assumption is fundamentally unattainable. This misalignment between fixed class representations and diverse image features serves as compelling evidence for the necessity of adaptive mechanisms in modern segmentation frameworks.

3.2. Ideal Class Representation (Feature) Mining

To theoretically demonstrate the necessity of adaptive class representations and image features, we derive optimal per-image class prototypes through inverse reasoning based on

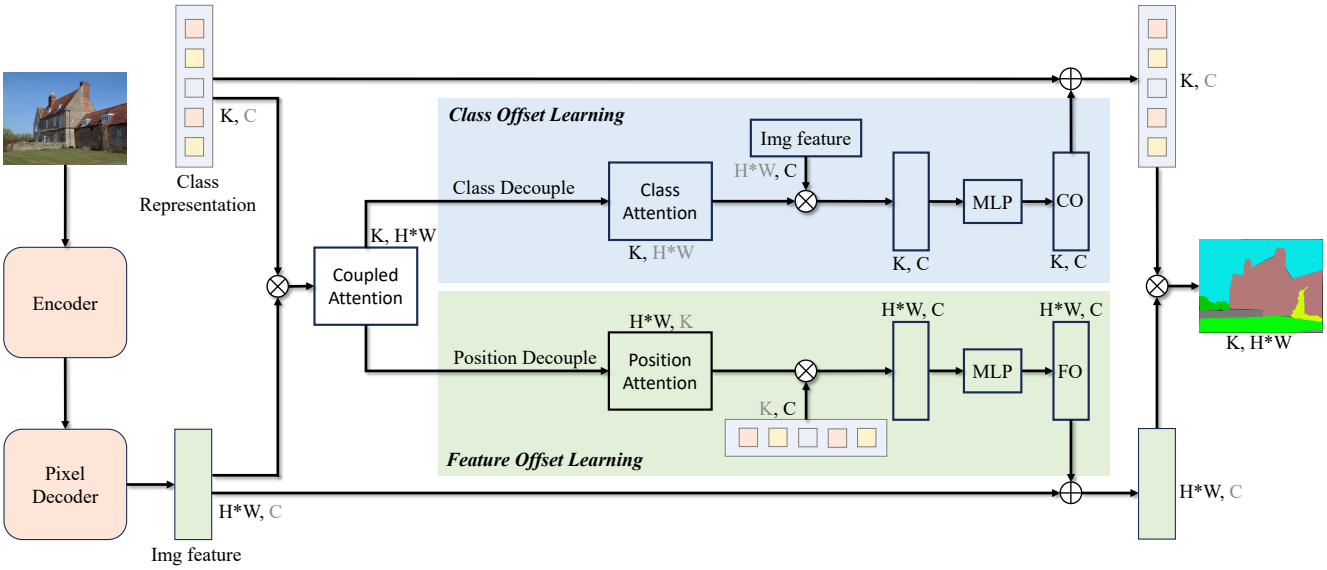

Figure 4. Framework of the proposed OffSeg. Given an input image, we first use the encoder to extract the multi-scale features and then use the pixel decoder to generate the image feature. The offset learning paradigm contains two branches: Class offset learning branch and Feature offset learning branch. With the learned class offset (CO) and feature offset (FO), we guide the initial feature to aligned space. We denote the dimension of matrix multiplication by grey.

ground-truth masks. Given an input image with ground-truth mask $\mathbf{M} \in \mathbb{R}^{K \times HW}$ and its deep feature $\mathbf{E} \in \mathbb{R}^{HW \times C}$, the ideal class prototypes $\mathbf{W}^* \in \mathbb{R}^{K \times C}$ should satisfy: $\mathbf{M} = \mathbf{W}^* \cdot \mathbf{E}^\top$, where each row of $\mathbf{W}^*$ represents the optimal prototype for a specific class. Solving this linear system yields:

$$\mathbf{W}^* = \mathbf{M} \cdot \left(\mathbf{E}^\top\right)^\dagger, \tag{2}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse.

By recomputing masks via $\mathbf{M}_{\text{pred}} = \mathbf{W}^* \cdot \mathbf{E}^\top$, we achieve near-perfect reconstruction with around 95% mIoU on the ADE20K dataset, confirming the theoretical validity of $\mathbf{W}^*$. With such a mathematical derivation, we conduct a similarity analysis using SegFormer [46] (an efficient network with 4.3M learnable parameters). Firstly, we randomly select six categories (*i.e.*, building, sky, floor, tree, person, glass) and compute $\mathbf{W}^*$ for 10 images per category for a better view. As shown in Fig. 3, we visualize their pairwise similarity via heatmaps. Strikingly, the ideal class representations exhibit less similarity than our common sense. The low correlation heatmap patterns reveal that optimal class representations vary drastically across images for the same class. This phenomenon stems from a fundamental tension in efficient models: aggressive feature compression amplifies intra-class feature variance, forcing $\mathbf{W}^*$ to diverge significantly to fit distorted features. The above analysis reveals two critical implications:

- Fixed prototypes fail: Fixed class representations cannot universally align with highly variable features of different images in efficient models.
- Fixed features fail: The relationship $\mathbf{W}^* \propto f(\mathbf{E})$ implies that distorted features also hinder prototype stability, which needs a vicious cycle requiring joint correction.

3.3. Offset Learning Paradigm

Our offset learning paradigm redefines the segmentation paradigm as a dual-decoupled alignment process:

$$\mathbf{M} = (\mathbf{W} + \Delta\mathbf{W}) \cdot (\mathbf{E} + \Delta\mathbf{E})^\top, \tag{3}$$

which departs our method from conventional per-pixel classification [7, 18, 19, 59] and mask-centric approaches [10, 11]. The core innovation lies in the class offset learning and feature offset learning branches that collaboratively refine class prototypes and spatial features through decoupled attention mechanisms as shown in Fig. 4.

To be specific, given the image feature $\mathbf{E} \in \mathbb{R}^{HW \times C}$ and class embedding $\mathbf{W} \in \mathbb{R}^{K \times C}$, we compute a coupled attention matrix $\mathbf{A}_c$:

$$\mathbf{A}_c = \mathbf{W} \cdot \mathbf{E}^\top, \tag{4}$$

where $\mathbf{A}_c \in \mathbb{R}^{K \times HW}$. This matrix encodes correlations between classes and spatial positions and is used in the subsequent class offset learning branch and feature offset learning branch.

Class offset learning dynamically adjusts class representations based on spatial context, alleviating the rigidity of fixed class embeddings. First, we apply the softmax normalization along spatial dimensions $\mathrm{Softmax}_S$ to generate

class-wise attention weights:

$$\mathbf{A}_{\text{cls}} = \text{Softmax}_S(\mathbf{A}_c) \in \mathbb{R}^{K \times HW}, \tag{5}$$

where each row $\mathbf{a}_k \in \mathbf{A}_{\text{cls}}$ indicates the spatial importance distribution for class k. Then, we aggregate spatial features weighted by class attention:

$$\mathbf{F}_{\text{cls}} = \mathbf{A}_{\text{cls}} \cdot \mathbf{E} \in \mathbb{R}^{K \times C}, \tag{6}$$

where $\mathbf{F}_{\text{cls}}$ contains class-specific prototypes that encode global spatial distributions. Finally, we generate class offsets via an MLP:

$$\Delta\mathbf{W} = \text{MLP}(\mathbf{F}_{\text{cls}}) \in \mathbb{R}^{K \times C}, \tag{7}$$

and adjust original representations as follows:

$$\mathbf{W}_{\text{adj}} = \mathbf{W} + \Delta\mathbf{W}. \tag{8}$$

This branch learns image-specific offsets to align class embeddings with corresponding image features, narrowing their representational gap.

Feature offset learning refines image features by injecting class-aware semantics. The intention is to overcome the local ambiguity in per-pixel classification. As shown in Fig. 4, the feature offset learning branch is dual to the class offset learning branch.

To be specific, like the class offset learning branch, we first apply softmax along the class dimension Softmax_K and transpose for spatial alignment:

$$\mathbf{A}_{\text{pos}} = (\text{Softmax}_K(\mathbf{A}_c))^T \in \mathbb{R}^{HW \times K}, \tag{9}$$

where each row $\mathbf{a}_i \in \mathbf{A}_{\text{pos}}$ represents the class probability distribution at position i. Then, we fuse class semantics into spatial positions via the following equation:

$$\mathbf{F}_{\text{pos}} = \mathbf{A}_{\text{pos}} \cdot \mathbf{W} \in \mathbb{R}^{HW \times C}, \tag{10}$$

where $\mathbf{F}_{\text{pos}}$ encodes position-wise semantic guidance from all classes. Finally, we adopt an MLP to generate feature offset:

$$\Delta\mathbf{E} = \text{MLP}(\mathbf{F}_{\text{pos}}) \in \mathbb{R}^{HW \times C}, \tag{11}$$

and use it to guide the original features:

$$\mathbf{E}_{\text{adj}} = \mathbf{E} + \Delta\mathbf{E}. \tag{12}$$

The final segmentation masks are generated through bidirectional elastic alignment:

$$\mathbf{M} = \mathbf{W}_{\text{adj}} \cdot \mathbf{E}_{\text{adj}}^T, \tag{13}$$

which is another form of Eqn. (3).

We have summarized the main differences between our offset learning paradigm and other semantic segmentation paradigms in Tab. 1. Different from per-pixel classification (*e.g.*, SegFormer [46], SegNeXt [18]), which relies on static alignment between fixed features and rigid class embeddings, or mask classification (*e.g.*, MaskFormer [10], Mask2Former [11]) that dynamically refines class queries by a heavy transformer decoder, our framework uniquely introduces bidirectional offset learning with even negligible learnable parameters. This method enables symmetric adaptation: class representations can be adjusted through class-specific spatial prototypes, while image features can be refined by position-aware semantic guidance. By decoupling class- and position-wise interactions into two distinct pathways, our method achieves elastic feature-class alignment, where both modalities co-evolve to capture instance-specific geometries and contextual semantics. This contrasts sharply with the unidirectional or hard-coded alignment strategies in existing paradigms.

3.4. Overall Architecture

To validate the efficiency and effectiveness of our proposed offset learning paradigm, we design a standard semantic segmentation model with the following efficient components without structural modifications. For the backbone, we employ a hybrid architecture named EfficientFormerV2 [24], which achieves a balance between parameter efficiency and performance through a fine-grained joint search strategy. For multi-scale feature aggregation, we select FreqFusion [5], which fuses two scale features with frequency-aware operators. Notably, when combined with our offset learning paradigm, the entire model introduces nearly negligible learnable parameters (0.1-0.2M).

4. Experiments

4.1. Experimental Settings

Datesets. We evaluate our method on four widely adopted semantic segmentation benchmarks: ADE20K [61], Cityscapes [13], COCO-Stuff [3], and Pascal Context [33]. ADE20K [61] is a scene parsing dataset with 150 object/stuff categories, containing 20K/2K/3K images for training/validation/testing. It features diverse indoor and outdoor scenes with complex occlusions. Cityscapes [13] focuses on urban driving scenarios, providing 5,000 high-resolution images (2048×1024) with 19 semantic classes. COCO-Stuff [3] comprises 118K training and 5K validation images with 171 classes (80 things + 91 stuff). Its long-tailed distribution challenges model generalization. PASCAL Context [33] dataset comprises 59 semantic categories as foreground objects, with 4,996 training images and 5,104 validation images.

Implementation details. Our implementation is based on the MMSegmentation [12] with PyTorch. Following previous works [10, 11, 18, 34, 46], we adopt the AdamW [32]

Table 2. Performance comparison of state-of-the-art methods on ADE20K, Cityscapes and COCO-Stuff datasets. FLOPs (G) is computed at input resolutions of 512×512 for ADE20K and COCO-Stuff, and 2048×1024 for Cityscapes.

Method	Params (M)	ADE20K		Cityscapes		COCO-Stuff	
		FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU
SegFormer-B0 [46]	3.8	8.4	37.4	125.5	76.2	8.4	35.6
RTFormer-Slim [43]	4.8	17.5	36.7	-	76.3	-	-
FeedFormer-B0 [37]	4.5	7.8	39.2	107.4	77.9	-	-
Seaformer-L [41]	14.0	6.5	42.7	-	-	-	-
VWFormer-B0 [47]	3.7	5.1	38.9	-	77.2	5.1	36.2
CGRSeg-T [34]	9.4	4.0	43.6	-	-	4.0	42.2
EDAFormer-T [53]	4.9	5.6	42.3	151.7	78.7	5.6	40.3
OffSeg-T	6.2	5.3	44.2	44.8	78.9	5.3	41.9
SegFormer-B1 [46]	13.7	15.9	42.2	243.7	78.5	15.9	40.2
SegNeXt-S [18]	13.9	15.9	44.3	124.6	81.3	15.9	42.2
RTFormer-Base [43]	16.8	67.4	42.1	-	79.3	26.6	35.3
VWFormer-B1 [47]	13.7	13.2	43.2	-	79.0	-	41.5
PEM-STDC1 [4]	17.0	16.0	39.6	-	-	-	-
OffSeg-B	13.0	10.3	45.9	86.5	80.5	10.3	44.3
SenFormer [2]	59.0	179.0	46.0	-	-	-	-
SegFormer-B2 [46]	27.5	25.9	45.6	717.1	81.0	26.0	44.6
MaskFormer [10]	42.0	55.0	46.7	-	-	-	-
Mask2Former [11]	47.0	74.0	47.7	-	-	-	-
FeedFormer-B2 [37]	29.1	42.7	48.0	522.7	81.5	-	-
PEM-STDC2 [4]	21.0	19.3	45.0	-	-	-	-
OffSeg-L	26.4	17.1	48.5	143.4	81.6	17.1	46.0

optimizer with poly learning rate decay and 1,500 iterations linear warmup for all models, without specific tuning for any other settings. The batch size is set to 16 for the ADE20K/COCO-Stuff/Pascal Context datasets and 8 for the Cityscapes dataset. During training, the image size is cropped to 512×512 for ADE20K and COCO-Stuff, 480×480 for Pascal Context, and 1024×1024 for the Cityscapes dataset. We adopt the standard data augmentation and train 160k iterations on ADE20K and Cityscapes datasets and 80k iterations on the COCO-Stuff and Pascal Context datasets. During inference, we employ single-scale testing for all datasets. All experiments are conducted on 8 NVIDIA RTX 3090 GPUs.

4.2. Main Results

We evaluate our method on three standard semantic segmentation benchmarks: ADE20K, Cityscapes and COCO-Stuff, as detailed in Tab. 2. For the ADE20K dataset, the proposed OffSeg-T attains 44.2 mIoU on ADE20K, surpassing EDAFormer-T by 1.9 mIoU while reducing computations by 24%. OffSeg-B establishes a strong accuracy-efficiency trade-off: 45.9 mIoU on ADE20K (10.3G FLOPs), outperforming SegNeXt-S (+1.6) and PEM-STDC1 (+6.3) with 35% lower FLOPs than SegNeXt-S. At the large scale, OffSeg-L achieves a 48.5 mIoU score on ADE20K with 17.1G FLOPs, outperforming Mask2Former (+0.8) with 4.3× fewer FLOPs

On the Cityscapes dataset, our OffSeg-L achieves superior results while using only a quarter of the computational cost required by FeedFormer-B2. On the COCO-Stuff dataset, our OffSeg-B surpasses RTFormer-Base by 9.0 mIoU with less than half of its computational cost.

These experimental results suggest that our dual-decoupled offset learning paradigm can effectively address the misalignment between class representations and image features, particularly in class-dense and challenging scenarios, like ADE20K and COCO-Stuff.

4.3. Generalization Ability

To validate the broad applicability of our offset learning paradigm, we integrate it into three representative models: SegNeXt [18] (CNN-based), SegFormer [46] (Transformer-based), and Mask2Former [11] (mask classification). For the per-pixel classification framework models (SegNeXt and SegFormer), we adapt our approach by simply replacing the final 1×1 convolutional layer with our offset learning paradigm. For the mask classification framework (Mask2Former), we leverage offset learning to align mask embeddings with per-pixel embeddings while remaining other parts unchanged.

SegNeXt with offset learning. To evaluate the robustness of our method, we conduct experiments on four datasets. As shown in Table 3, integrating our paradigm into SegNeXt

Table 3. Performance comparison of SegNeXt [18] and SegNeXt w/ offset learning on ADE20K, Cityscapes, Pascal Context, and COCO-Stuff datasets. FLOPs (G) is computed at input resolutions of 2048×1024 for Cityscapes and 512×512 for other datasets.

Method	Offset	Params (M)	ADE20K		Cityscapes		Pascal Context		COCO-Stuff	
			FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU
SegNeXt-T		4.3	6.6	41.1	50.5	79.8	6.6	51.2	6.6	38.7
SegNeXt-T	✓	4.4	7.2	43.0$_{(+1.9)}$	53.1	80.0$_{(+0.2)}$	6.8	53.2$_{(+2.0)}$	7.3	40.0$_{(+1.3)}$
SegNeXt-S		13.9	15.9	44.3	124.6	81.3	15.9	54.2	15.9	42.2
SegNeXt-S	✓	14.1	16.5	45.6$_{(+1.3)}$	127.2	81.7$_{(+0.4)}$	16.1	55.9$_{(+1.7)}$	16.6	43.5$_{(+1.3)}$
SegNeXt-B		27.6	34.9	48.5	275.7	82.6	34.9	57.0	34.9	45.8
SegNeXt-B	✓	28.2	34.8	49.4$_{(+0.9)}$	269.6	82.8$_{(+0.2)}$	34.1	58.0$_{(+1.0)}$	35.0	45.8$_{(+0.0)}$

Table 4. Performance comparison of SegFormer [46] and SegFormer w/ offset learning on ADE20K and COCO-Stuff datasets. FLOPs (G) is computed at input resolutions of 512×512 for all datasets.

Method	Offset	Params	ADE20K		COCO-Stuff	
			FLOPs	mIoU	FLOPs	mIoU
SegFormer-B0		3.8M	8.4	37.4	8.6	35.6
SegFormer-B0	✓	3.9M	8.8	40.1$_{(+2.7)}$	8.9	38.3$_{(+2.7)}$
SegFormer-B1		13.7M	15.9	41.0	16.1	40.2
SegFormer-B1	✓	13.9M	16.3	43.7$_{(+2.7)}$	16.4	41.9$_{(+1.7)}$
SegFormer-B2		24.8M	25.9	45.6	26.0	44.6
SegFormer-B2	✓	24.9M	26.1	47.3$_{(+1.7)}$	26.2	45.2$_{(+0.6)}$
SegFormer-B3		44.6M	42.5	47.8	42.6	45.5
SegFormer-B3	✓	44.8M	42.8	49.5$_{(+1.7)}$	42.9	46.3$_{(+0.8)}$
SegFormer-B4		61.4M	59.2	48.5	59.3	46.5
SegFormer-B4	✓	61.6M	59.5	50.1$_{(+1.6)}$	59.6	47.0$_{(+0.5)}$
SegFormer-B5		82.0M	75.2	49.1	75.3	46.7
SegFormer-B5	✓	82.2M	75.5	50.6$_{(+1.5)}$	75.5	47.2$_{(+0.5)}$

yields consistent performance improvements with negligible parameter overhead. In summary, our method achieves average improvements of 1.4, 1.2, and 0.5 mIoU across all datasets for the Tiny, Small, and Base scales, respectively, while introducing only 0.1-0.2M additional parameters. These results demonstrate the effectiveness and efficiency of our approach in enhancing segmentation performance. We also evaluate SegNeXt-T with offset learning paradigm using the ensemble strategy (multi-scale). The model achieves mIoU of 43.2 on ADE20K, 81.5 on Cityscapes, 54.5 on Pascal Context, and 40.5 on COCO-Stuff, further enhancing segmentation accuracy.

To further show the advantages of our method, we show the segmentation results based on the SegNeXt-T model in Fig. 5. The visualizations reveal that our model achieves more precise segmentation outcomes, particularly in the identification of background regions and small objects. This qualitatively validates that our method can better align image features with class representations.

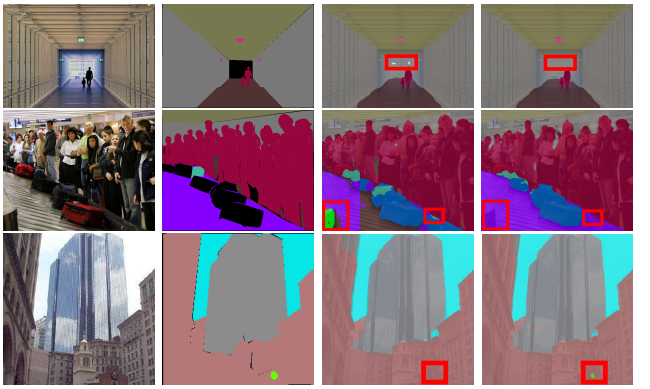

Figure 5. Visualization of the offset learning paradigm on SegNeXt. Compared to the baseline SegNeXt-T, applying offset learning paradigm enables the model to segment objects more accurately, especially small objects (*e.g.*, the clock in the third image).

Furthermore, the diminishing improvement trend (from 1.4 to 0.5 mIoU) as the model size increases suggests that larger-scale models inherently possess superior feature-representation alignment capabilities. This experimental observation directly validates our hypothesis of a positive correlation between model size and class alignment capability. The results also demonstrate that our approach resolves the identified issues with the efficiency level anticipated in our analysis.

SegFormer with offset learning. To systematically evaluate the adaptability of our offset learning paradigm across varying model capacities, we perform comprehensive experiments on six SegFormer architectures (B0-B5) and report results on ADE20K and COCO-Stuff. To ensure a fair comparison, we employ the results provided by mmsegmentation [12], wherein the model's FLOPs are smaller than those reported in the paper. From B0 to B5, the average mIoU improvements on the two datasets are 2.7, 2.2, 1.2, 1.3, 1.1, and 1.0, respectively. This gradually diminishing performance gain with increasing model scale aligns with the conclusions drawn from experiments on the SegNeXt model, further validating our hypothesis regarding the mis-

Table 5. Performance comparison of Mask2Former [11] and Mask2Former w/ offset learning on ADE20K dataset.

Method	Offset	Params (M)	mIoU
Mask2Former-Tiny		47.4	47.7
Mask2Former-Tiny	✓	47.6	$50.3_{(+2.6)}$

alignment between efficient model features and class representations. It also underscores the effectiveness of our offset learning paradigm in addressing the issue of misalignment in the traditional per-pixel classification paradigm.

Mask2Former with offset learning. We perform experiments on Mask2Former-Tiny, as even the Tiny model already has 47.4M parameters. As shown in Tab. 5, the results demonstrate that integrating our method into the mask classification paradigm improves the performance by 2.6 mIoU with only a 0.2M parameter increase. As shown in the comparison in Tab. 1, the mask classification paradigm only adjusts class representations, whereas our offset learning paradigm simultaneously aligns both image features and class representations. This demonstrates the effectiveness of our method in achieving efficient and concurrent alignment of image features (per-pixel embeddings) and class representations (mask embeddings).

4.4. Ablation Study

Ablation analysis on core components. To systematically validate the efficacy of each component in our OffSeg framework, we conduct ablation studies on the ADE20K dataset, as detailed in Tab. 6 The baseline model (first row) employs a simple convolutional pixel decoder without FreqFusion [5] and achieves 40.7 mIoU. Introducing FreqFusion alone improves the mIoU score by 1.3. Based on FreqFusion, combining class offset learning and feature offset learning improves mIoU by 0.9 and 1.5, respectively. This demonstrates that both branches can independently enhance the model performance, with adaptable features yielding a larger impact than adjustable class representations. The simultaneous incorporation of both branches achieves the best performance, demonstrating their synergistic effect in jointly aligning image features and class representations.

Ablation analysis on the effect of channel number. As shown in Tab. 7, we adopt the baseline w/ FreqFusion only as the base model to verify the effect of channel number on the per-pixel classification paradigm. When the channel number scales up from 64 to 1024, the results show that increasing feature dimensions in per-pixel classification models yields diminishing returns, and when it reaches 2048, the performance experiences a certain decline. This reveals that increasing the image feature channels and class representation channels for the per-pixel classification paradigm can improve the expressive ability of the model, as higher-dimensional vectors can represent higher-dimensional spaces.

However, the simple approach of enhancing model performance by increasing the number of channels incurs significant computational overhead and has an upper limit. For instance, at a channel of 1024, the model achieves 43.5 mIoU with 11.1G FLOPs, but further increasing the channel number does not enhance performance. In contrast, our OffSeg-T with offset learning paradigm achieves a performance with 44.2 mIoU while requiring less than half of the computational resources. This comparative analysis further substantiates the efficiency and effectiveness of our proposed method in achieving good segmentation accuracy.

Table 6. Ablation experiments on different components of the proposed OffSeg. FF., CO., and FO. represent FreqFusion, class offset learning, and feature offset learning, respectively.

FF.	CO.	FO.	Params (M)	FLOPs (G)	mIoU
			5.9	5.1	40.7
✓			6.1	6.0	42.0
✓	✓		6.2	5.3	42.9
✓		✓	6.2	5.3	43.5
✓	✓	✓	6.2	5.3	44.2

Table 7. Ablation experiments on the effect of channel number of image features and class representations on baseline. The models all belong to the per-pixel classification paradigm, not our offset learning paradigm.

Channel	64	128	256	512	768	1024	2048
Params (M)	6.0	6.0	6.1	6.2	6.3	6.4	6.8
FLOPs (G)	4.7	5.1	6.0	7.7	9.4	11.1	17.9
mIoU (%)	41.2	41.7	42.0	42.8	42.9	43.5	43.0

5. Conclusions

In this paper, we analyze the limitations of the per-pixel classification paradigm, specifically the misalignment between image features and class representations. To address this issue, we propose the offset learning paradigm, which introduces separate feature offset learning and class offset learning branches to explicitly learn the necessary offsets for aligning image features with their corresponding class representations. Building upon this paradigm, we design a series of efficient segmentation networks, named OffSeg, containing three different scales. As a general segmentation paradigm, we also integrate our offset learning paradigm into three representative segmentation methods, including SegFormer, SegNeXt, and Mask2Former with negligible parameters. Extensive experiments on four widely used datasets demonstrate the effectiveness and efficiency of our proposed offset learning paradigm.

Acknowledgment

This work was partially funded by NSFC (No. 62495061, 62276145), National Key Research and Development Project of China (No. 2024YFE0100700), the Science and Technology Support Program of Tianjin, China (No. 23JCZDJC01050), and A*STAR Career Development Fund (No. C233312006). This work was also supported in part by the Supercomputing Center of Nankai University.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 3

[2] Walid Bousselham, Guillaume Thibault, Lucas Pagano, Archana Machireddy, Joe Gray, Young Hwan Chang, and Xubo Song. Efficient self-ensemble for semantic segmentation. In *BMVC*, 2022. 6

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *IEEE CVPR*, pages 1209–1218, 2018. 5

[4] Niccolo Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer for image segmentation. In *IEEE CVPR*, pages 15804–15813, 2024. 2, 3, 6

[5] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE TPAMI*, 2024. 5, 8

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3

[7] 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 TPAMI*, 40(4):834–848, 2017. 4

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

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 3

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021. 2, 3, 4, 5, 6

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE CVPR*, pages 1290–1299, 2022. 2, 3, 4, 5, 6, 8

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5, 7

[13] 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 *IEEE CVPR*, pages 3213–3223, 2016. 1, 5

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[15] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023. 1

[16] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *IEEE CVPR*, pages 9716–9725, 2021. 3

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE CVPR*, pages 3146–3154, 2019. 3

[18] 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. *NeurIPS*, 35:1140–1156, 2022. 1, 2, 3, 4, 5, 6, 7

[19] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *IEEE CVPR*, pages 4003–4012, 2020. 4

[20] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications, 2024. 1

[21] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *IEEE CVPR*, pages 7287–7296, 2022. 3

[22] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *IEEE CVPR*, pages 9522–9531, 2019. 1

[23] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *NeurIPS*, 35:12934–12949, 2022. 1

[24] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *IEEE ICCV*, pages 16889–16900, 2023. 5

[25] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 2024. 1

[26] Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28243–28258. PMLR, 2024. 1

[27] Dong Liang, Yue Sun, Yun Du, Songcan Chen, and Sheng-Jun Huang. Relative difficulty distillation for semantic segmentation. *Science China Information Sciences*, 67(9): 192105, 2024. 3

[28] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *Computational Visual Media*, 9(4):753–765, 2023. 1

[29] 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 *IEEE ICCV*, pages 10012–10022, 2021. 1

[30] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *IEEE CVPR*, pages 12009–12019, 2022. 1

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 3

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE CVPR*, pages 891–898, 2014. 5

[34] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *ECCV*, pages 239–255. Springer, 2024. 1, 3, 5, 6

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE ICCV*, pages 12179–12188, 2021. 3

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 3

[37] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *AAAI*, pages 2263–2271, 2023. 1, 3, 6

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE ICCV*, pages 7262–7272, 2021. 1, 3

[39] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. Contextrast: Contextual contrastive learning for semantic segmentation. In *IEEE CVPR*, pages 3732–3742, 2024. 1

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3

[41] 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. 1, 3, 6

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10): 3349–3364, 2020. 3

[43] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Rtformer: Efficient design for real-time semantic segmentation with transformer. *NeurIPS*, 35:7423–7436, 2022. 1, 3, 6

[44] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2T: Pyramid pooling transformer for scene understanding. *IEEE TPAMI*, 45(11):12760–12771, 2023.

[45] Yu-Huan Wu, Shi-Chen Zhang, Yun Liu, Le Zhang, Xin Zhan, Daquan Zhou, Jiashi Feng, Ming-Ming Cheng, and Liangli Zhen. Low-resolution self-attention for semantic segmentation. *IEEE TPAMI*, 2025. 1, 3

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 1, 2, 3, 4, 5, 6, 7

[47] Haotian Yan, Ming Wu, and Chuang Zhang. Multi-scale representations by varying window attention for semantic segmentation. In *ICLR*, 2024. 3, 6

[48] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *ICLR*, 2024. 1

[49] Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19345–19355, 2025. 1

[50] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 1, 3

[51] 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. *IJCV*, 129:3051–3068, 2021. 1, 3

[52] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[53] Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, and Suk-Ju Kang. Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In *ECCV*, pages 92–110. Springer, 2024. 6

[54] 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. 3

[55] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *NeurIPS*, 34: 7281–7293, 2021. 3

[56] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *NeurIPS*, 35:4971–4982, 2022. 1

[57] Dong Zhang, Liyan Zhang, and Jinhui Tang. Augmented fcn: rethinking context modeling for semantic segmentation. *Science China Information Sciences*, 66(4):142105, 2023. 3

[58] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *IEEE CVPR*, pages 12083–12093, 2022. 1, 3

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE CVPR*, pages 2881–2890, 2017. 3, 4

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE CVPR*, pages 6881–6890, 2021. 1, 3

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE CVPR*, pages 633–641, 2017. 1, 5

[62] Tianfei Zhou and Wenguan Wang. Cross-image pixel contrasting for semantic segmentation. *IEEE TPAMI*, 2024. 1

[63] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 1