MetaSeg: MetaFormer-based Global Contexts-aware Network for Efficient Semantic Segmentation

Beoungwoo Kang*, Seunghun Moon*, Yubin Cho*, Hyunwoo Yu*, and Suk-Ju Kang
Sogang University, Republic of Korea
{beoungwoo, moonsh97, dbqls1219, hyunwoo137, sjkang}@sogang.ac.kr

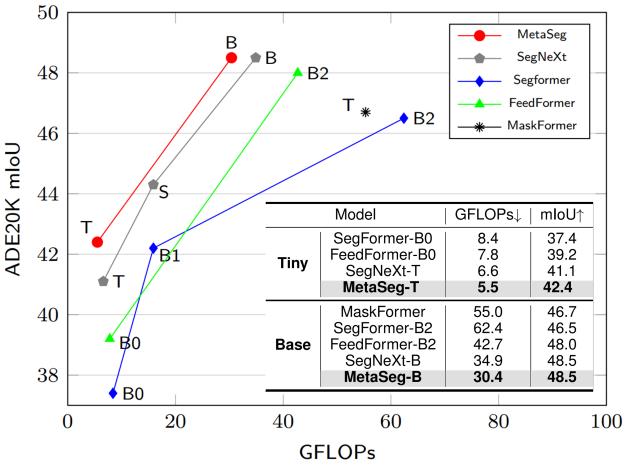
Abstract

Beyond the Transformer, it is important to explore how to exploit the capacity of the MetaFormer, an architecture that is fundamental to the performance improvements of the Transformer. Previous studies have exploited it only for the backbone network. Unlike previous studies, we explore the capacity of the Metaformer architecture more extensively in the semantic segmentation task. We propose a powerful semantic segmentation network, MetaSeg, which leverages the Metaformer architecture from the backbone to the decoder. Our MetaSeg shows that the MetaFormer architecture plays a significant role in capturing the useful contexts for the decoder as well as for the backbone. In addition, recent segmentation methods have shown that using a CNN-based backbone for extracting the spatial information and a decoder for extracting the global information is more effective than using a transformer-based backbone with a CNN-based decoder. This motivates us to adopt the CNN-based backbone using the MetaFormer block and design our MetaFormer-based decoder, which consists of a novel self-attention module to capture the global contexts. To consider both the global contexts extraction and the computational efficiency of the self-attention for semantic segmentation, we propose a Channel Reduction Attention (CRA) module that reduces the channel dimension of the query and key into the one dimension. In this way, our proposed MetaSeg outperforms the previous state-of-the-art methods with more efficient computational costs on popular semantic segmentation and a medical image segmentation benchmark, including ADE20K, Cityscapes, COCO-stuff, and Synapse.

1. Introduction

Semantic segmentation is a challenging task in the computer vision that classifies categories for each pixel in an image. This task has a wide range of applications, including autonomous driving and medical image segmentation.

*These authors contributed equally.



	Model	GFLOPs↓	mIoU↑
Tiny	SegFormer-B0	8.4	37.4
	FeedFormer-B0	7.8	39.2
	SegNeXt-T	6.6	41.1
	MetaSeg-T	**5.5**	**42.4**
Base	MaskFormer	55.0	46.7
	SegFormer-B2	62.4	46.5
	FeedFormer-B2	42.7	48.0
	SegNeXt-B	34.9	48.5
	MetaSeg-B	**30.4**	**48.5**

Figure 1. Performance-Computation curves on ADE20K validation set. Compared the performance and computation of our MetaSeg with recent models [6, 11, 20, 30]. We find that our MetaSeg has the best trade-off between the performance and computational costs.

With the great success of the vision transformer (ViT) [8] in the image classification, the transformer-based methods have been introduced in the field of semantic segmentation. Most previous studies [26–28, 30] mainly utilize the self-attention layer in the transformer block to achieve the superior performance. However, recent research [31] found that the abstracted architecture of the transformer block (*i.e.*, MetaFormer block), which consists of a token-mixer, channel MLPs and residual connections, plays a more significant role in achieving the competitive performance than the specific token mixer (*e.g.* attention, spatial MLP). Therefore, the MetaFormer architecture has the potential to be variably applied with different token mixers depending on the specific purpose.

From the MetaFormer architecture, some recent studies [13, 25] have derived their own methods. For example, EfficientFormer [13] employs the MetaFormer architecture using the self-attention as the token mixer to effectively cap-

ture the global semantic-aware features. InternImage [25] also utilizes the MetaFormer with the deformable convolution as the token mixer to capture the contextual information. These methods have exploited the capacity of the MetaFormer architecture only for the encoder. However, unlike previous studies, we take advantage of the capacity of the MetaFormer block more extensively for the semantic segmentation task. Therefore, we propose a novel and powerful segmentation network, *MetaSeg*, which utilizes the MetaFormer block up to the decoder to obtain the enhanced visual representation.

In addition, previous segmentation methods [14, 26, 30] used the transformer-based backbone with the CNN-based decoder. However, recent studies [11, 32] have shown that using the CNN-based backbone for extracting the local information and the decoder for extracting the global information is more effective in improving the performance by compensating for the globality in the local contexts. Based on this observation, we adopt the CNN-based backbone (*i.e.*, MSCAN [11]) that contains the MetaFormer block used the convolution as a token mixer, and design a novel transformer-based decoder. Since it is important to consider the globality in the decoder to complement the CNN-based encoder features, the proposed decoder leverages the MetaFormer block that uses the self-attention as a token mixer to capture the global contexts. However, the self-attention has a limitation of the considerable computational costs due to the high-resolution features in the semantic segmentation task.

To address this issue, we propose a novel and efficient self-attention module, *Channel Reduction Attention* (CRA), which embeds the channel dimension of the query and key into the one dimension for each head in the self-attention operation. Conventional self-attention methods [8, 14, 26, 28, 30], which embed the channel dimension of the query and key without the channel reduction, show great performance but have high computational costs. Compared to these methods, our method leads to competitive performance with the computational reduction. This indicates that our CRA can sufficiently consider the globality even when each query and key token is a scalar type, not a vector. Therefore, our CRA module is more efficient and effective than the previous self-attention modules.

To demonstrate the effectiveness and efficiency of our method, we conduct experiments on the challenging semantic segmentation datasets: ADE20K [34], Cityscapes [7], and COCO-stuff [2]. To verify the ability for the application, we also conduct experiments on the medical image segmentation dataset: Synapse [1]. As shown in Fig. 1, our MetaSeg-T and MetaSeg-B surpass the previous state-of-the-art methods on three public semantic segmentation benchmarks, including ADE20K, Cityscapes, and COCO-Stuff. Especially, our MetaSeg-T outperforms SegNeXt-T [11] by 1.3%, 0.3% and 1.0% mIoU improvements with 16.7%, 5.2% and 16.7% lower computational costs on ADE20K, Cityscapes, and COCO-Stuff, respectively.

In summary, the main contributions of our method are summarized as follows.

- The proposed MetaSeg is a powerful semantic segmentation network that effectively captures the local to global contexts, showing that the capacity of the MetaFormer architecture can be extended to the decoder as well as the encoder.
- We propose Channel Reduction Attention (CRA), a novel and efficient self-attention module for semantic segmentation, which can consider the globality efficiently by reducing the channel dimension of the query and key into the one dimension for the computational reduction in the self-attention operation.
- Our proposed MetaSeg outperforms the previous state-of-the-art methods in terms of efficiency, accuracy and robustness on three challenging semantic segmentation datasets and a medical image segmentation dataset to show ours applicability across different domains.

2. Related Works

2.1. MetaFormer-based architecture

MetaFormer is an general architecture of the transformer [23] where the token mixer is not specified. Recent methods [21, 22, 31] have explored various types of token mixers within the MetaFormer architecture to encourage the performance. Mlp-Mixer [21] and ResMLP [22] utilized MLP-like token mixers. PoolFormer [31] simply exploited pooling as token mixers to verify the power of the MetaFormer architecture. PVT [26], Swin [14], CvT [28], and EfficientFormer [13] adopted the self-attention as token mixers to aggregate the global information. These studies have focused on exploiting a variant token mixer based on the MetaFormer in the encoder. Therefore, we propose novel MetaFormer block which is leverage our Channel Reduction Attention (CRA) module as a token mixer. In addition, unlike the previous methods that apply the MetaFormer architecture to the encoder, we propose novel approach that the capacity of the MetaFormer architecture is extended to the decoder to consider the globality that is helpful for improving the segmentation performance.

2.2. Semantic segmentation

As ViT [8] have achieved the great success on the image classification task, self-attention based transformer backbones have also been explored in the semantic segmentation task. SETR [33] was the first to use ViT as a backbone on the segmentation task. PVT [26], Swin [14], CvT [28],

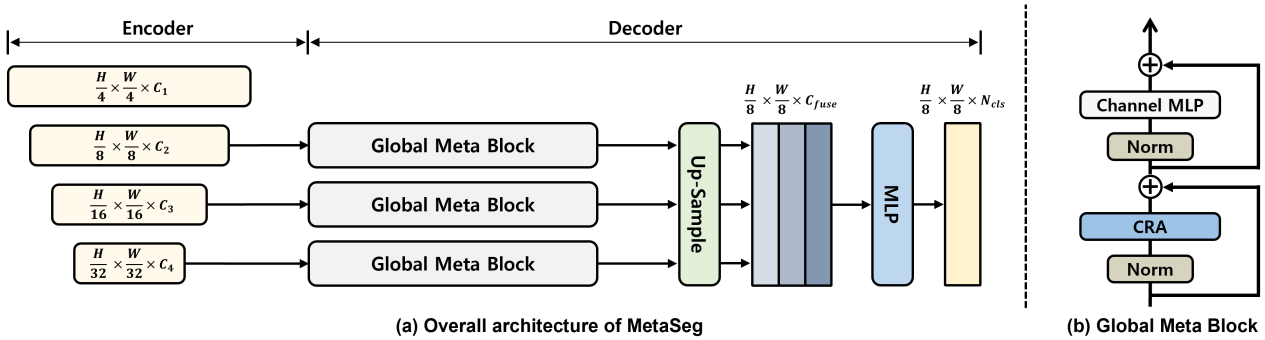


Figure 2. (a) Overall architecture of MetaSeg, consisting of two main part: hierarchical CNN-based Encoder and Global Meta Blcok (GMB) based decoder. (b) Details of the GMB, which is composed with the proposed Channel Reduction Attention (CRA) module and the channel MLP. Our MetaSeg extracts the multi-scale feature that contains local information in the encoder and complements the global information in the GMB of the decoder.

and LeViT [10] studied the hierarchical transformer-based backbone to exploit the multi-scale features. Beyond introducing transformer backbones for the segmentation, Segformer [30] designed a light-weight transformer backbone and a MLP-based decoder to consider the computational efficiency. More recent methods [11, 32] adopted the CNN-based backbone with the transformer-based decoder to aggregate the local to global information. TopFormer [32] encoded the tokens by the MobileNetV2 [18], and then fed the tokens into the transformer blocks. In SegNeXt [11], the convolution-based encoder extracts the spatial information and the transformer-based decoder extracts the global context. These methods [11, 32] have demonstrated that using the CNN-based backbone with the transformer-based decoder is effective for the semantic segmentation. According to these studies, we adopt the combination of the CNN-based backbone and transformer-based decoder.

Additionally, transformer-based segmentation methods [14, 26, 30] have considered the computational efficiency of the attention mechanism due to high-resolution features. Swin [14] proposed a shifted window self-attention by partitioning the feature maps into the windows. Some recent methods [26, 30] adopted a spatial reduction attention that reduces the resolution of the key-value. In this paper, we introduce a novel self-attention module, Channel Reduction Attention (CRA), which reduces the channel dimension of the query and key into the one dimension for efficient computational costs of the self-attention.

3. Method

This section describes our MetaSeg architecture, an efficient and powerful segmentation network. Basically, we adopt the CNN-based encoder and MetaFormer-based decoder to aggregate the local and global information. We first explain the overall architecture, and then explain the encoder and decoder. Finally, we describe the Global Meta Block (GMB) with the proposed Channel Reduction Attention (CRA) that is an efficient self-attention module.

3.1. Overall Architecture

As shown in Fig. 2 (a), our MetaSeg is based on the MetaFormer block with a hierarchical backbone network of the four stages. We utilize the CNN-based encoder that adopts a series of convolutional layers as a token mixer. The encoder aggregates the local information from the input via the token mixer. For the decoder, we design the novel CRA module as a token mixer to capture the global contexts with low computational costs.

3.1.1 Hierarchical convolutional encoder

We adopt the CNN-based pyramid encoder to acquire multi-scale features. Following previous encoder-decoder structured segmentation networks, given an image $I \in \mathbb{R}^{H \times W \times 3}$ as an input, each stage of the encoder extracts the down-sampled features $F_i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ where $i \in \{1, 2, 3, 4\}$ and C_i denote the index of the encoder stage and the channel dimension. These features provide the coarse to fine-grained features that leads to the performance improvements of the semantic segmentation. Specifically, we adopt MSCAN [11] as a encoder, which consists of MetaFormer blocks using a convolution-based token mixer.

3.1.2 Lightweight decoder

The decoder of our MetaSeg exploits the MetaFormer architecture to improve the capture the global contexts that are not considered enough in the encoder. We discovered that the MetaFormer block, with the self-attention module as a token mixer, exhibits exceptional capability in gathering global contexts from the multi-scale features of the encoder. The decoder consists of following components: the

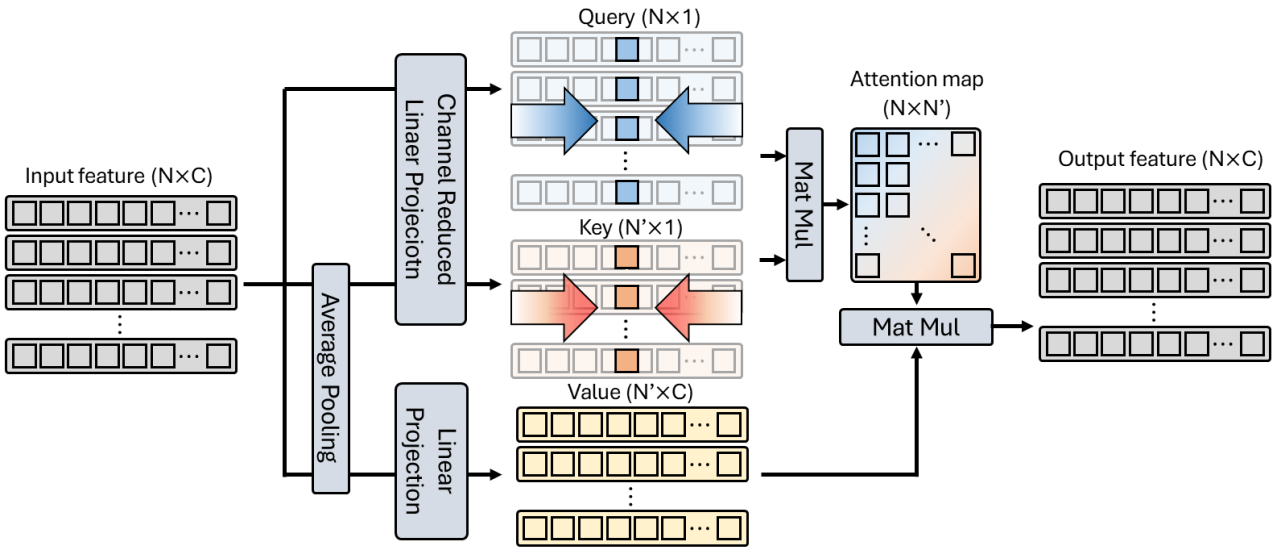


Figure 3. Illustration of the proposed Channel Reduction Attention (CRA). In our CRA, the channel dimension of the query and key is reduced to the one dimension for the computational efficiency and our CRA can capture the globality of the features effectively.

Global Meta Block (GMB), the up-sampling layer, the MLP layer and the prediction layer. The up-sampling layer expands the feature resolution to $\frac{H}{8} \times \frac{W}{8}$, unifying the size of outputs extracted from the GMB of each stage. We exclude the features of the first encoder stage since they contain too much low-level information and bring high computational costs. The MLP layer then concatenates the up-sampled features. Finally, the prediction layer predicts the segmentation mask. The overall procedure in decoder is as follows:

$$\begin{aligned} &\hat{F}_i = \text{GMB}(F_i),\ i \in \{2,3,4\} \\ &F_{up\_i} = \text{UpSample}(\frac{H}{8} \times \frac{W}{8})(\hat{F}_i),\ C_{fuse} = \sum_{i=2}^{4} C_i \\ &F = \text{Linear}(C_{fuse}, C_{MLP})(\text{Concat}(F_{up\_i})), \\ &Z = \text{Linear}(C_{MLP}, N_{cls})(F), \end{aligned} \tag{1}$$

where $\text{Linear}(a,b)(\cdot)$ denotes a linear layer with a size of a as input dimensions and a size of b as output dimensions. C_{MLP} denotes the channel dimension of the MLP. N_{cls} is defined as the number of classes.

3.2. Global Meta Block (GMB)

The proposed GMB leverages the MetaFormer block in the decoder to further enhance the global contexts of the feature representations extracted by the encoder, which mainly focuses on the local context. As illustrated in Fig. 2 (b), the GMB adopts the MetaFormer block of two residual sub-blocks and employs a novel channel reduction self-attention (CRA) module as a token mixer. Our CRA module effectively captures global contexts of the features with efficient computational costs. The GMB is performed at each stage except the first stage (*i.e.*, $i \in \{2,3,4\}$). The overall operation is defined as follows:

$$\begin{aligned} M_i &= \text{CRA}(\text{LN}(F_i)) + F_i\ , \\ \hat{F}_i &= \text{MLP}(\text{LN}(M_i)) + M_i\ , \end{aligned} \tag{2}$$

where LN and MLP denote the layer normalization and the channel MLP layer, respectively.

3.2.1 Channel Reduction Attention

We propose the Channel Reduction Attention (CRA) module as a novel token mixer utilized in the GMB to consider both the globality extraction and the computational efficiency of the self-attention for the semantic segmentation. Our CRA is based on the multi-head self-attention. The key and value are average pooled before the attention operation. As shown in Fig. 3, the channel dimensions of the query and key are embedded into the one dimension to further reduce the computational costs. We found that the channel squeezed query $Q \in \mathbb{R}^{\text{Head} \times H_i W_i \times 1}$ and key $K \in \mathbb{R}^{\text{Head} \times (H_i W_i / r_i^2) \times 1}$ can sufficiently extract global similarities. The CRA operation is formulated as follows:

$$\begin{aligned} &\text{CRA}(F_i) = \text{Concat}(\text{Head}_0, ..., \text{Head}_\text{j}) W_i^O\ , \\ &Q_i = F_i W_j^Q,\ K_i = \text{AvgPool}(F_i) W_j^K\ , \\ &V_i = \text{AvgPool}(F_i) W_j^V,\ \text{Head}_\text{j} = \text{Att}(Q_i, K_i, V_i)\ , \\ &\text{Att}(Q_i, K_i, V_i) = \text{Softmax}(Q_i K_i^T) V_i\ , \end{aligned} \tag{3}$$

where $W_j^Q, W_j^K \in \mathbb{R}^{C_i \times 1}$, $W_j^V \in \mathbb{R}^{C_i \times \frac{C_i}{j}}$ and $W^O \in \mathbb{R}^{C_i \times C_i}$ are projection parameters. j denotes the number

Method	Params(M)	ADE20K			Cityscapes			COCO-Stuff		
		GFLOPs ↓	mIoU (SS/MS) ↑		GFLOPs ↓	mIoU (SS/MS) ↑		GFLOPs ↓	mIoU (SS/MS) ↑	
SegFormer-B0 [30]	3.8	8.4	37.4	38.0	125.5	76.2	78.1	8.4	35.6	-
FeedFormer-B0 [20]	4.5	7.8	39.2	-	107.4	77.9	-	-	-	-
SegNeXt-T [11]	4.3	6.6	41.1	42.2	50.5	79.8	81.4	6.6	38.7	39.1
MetaSeg-T (Ours)	4.7	**5.5**	**42.4**	**43.4**	**47.9**	**80.1**	**81.5**	**5.5**	**39.7**	**40.2**
SegFormer-B2 [30]	27.5	62.4	46.5	47.5	717.1	81.0	82.2	62.4	44.6	-
MaskFormer [6]	42.0	55.0	46.7	48.8	-	-	-	-	-	-
FeedFormer-B2 [20]	29.1	42.7	48.0	-	522.7	81.5	-	-	-	-
SegNeXt-B [11]	27.6	34.9	48.5	**49.9**	275.7	82.6	83.8	34.9	45.8	46.3
MetaSeg-B (Ours)	29.6	**30.4**	**48.5**	49.4	**251.1**	**82.7**	**83.8**	**30.4**	**45.8**	**46.3**

Table 1. Comparison of our MetaSeg with previous state-of-the-arts methods on ADE20K, Cityscapes and COCO-Stuff. GFLOPs is calculated with 512×512 resolutions for ADE20K and COCO-Stuff, 2048×1024 resolutions for Cityscapes. Compared to previous state-of-the-arts methods, our MetaSeg model displays great effectiveness and efficiency.

of attention heads. AvgPool is the average pooling of scale $r_i \in \{2, 4, 8\}$ at each stage, respectively. Compared to SRA [26] that is a previous efficient self-attention method, the computational complexity of our CRA is as:

$$N' = \frac{N}{r_i}\,, \Omega(SRA) = (N')^2C + (N')^2C\,, \\ \Omega(CRA) = (N')^2 1 + (N')^2C\,, \tag{4}$$

where N denotes the number of pixel tokens. In eq.(4), the left and right terms indicate the computations of the query-key operation and the computations of the attention weight-value operation, respectively. By reducing the computation of the query-key operation by C times, our CRA reduces the total computation of the attention operation by about twice.

3.2.2 Channel MLP

The channel MLP is used to consolidate the features processed with our token mixer. Channel MLP consists of the two 1×1 convolution layers with a GELU activation layer. The operation is defined as follows:

$$\mathrm{MLP}(x) = \mathrm{Conv}_{1\times1}(\mathrm{GELU}(\mathrm{Conv}_{1\times1}(x)))\,, \tag{5}$$

where $\mathrm{Conv}_{1\times1}$ denotes the 1×1 convolution layer.

4. Experiment

4.1. Experimental Settings

Datasets. We conducted experiments on four publicly available datasets, ADE20K [34], Cityscapes [7], COCO-Stuff [2], and Synapse [1]. ADE20K is a challenging scene parsing dataset composed of 20,210/2,000/3,352 images for training, validation, and testing with 150 semantic categories. Cityscapes is an urban driving scene dataset that contains 5,000 images finely annotated with 19 categories.

Method	Params (M)	GFLOPs ↓	mIoU (%) ↑	FPS ↑
SegFormer-B0 [30]	3.8	51.8	74.2	25.5
FeedFormer-B0 [20]	4.5	41.6 (-19.7%)	75.5	28.9 (+13.3%)
SegNeXt-T [11]	4.3	29.3 (-43.4%)	77.8	30.2 (+18.4%)
MetaSeg-T (Ours)	4.7	**26.2 (-49.4%)**	**78.4**	**33.6 (+31.8%)**

Table 2. FPS comparison with recent state-of-the-art methods at the input size of 1536×768 using a RTX3090 GPU on Cityscapes.

It composed of 2,975/500/1,525 images in training, validation, and testing. COCO-Stuff is also a challenging dataset, which contains 172 semantic categories and 164,062 images. Synapse is an abdominal organ dataset that consists 30 Computerized Tomography (CT) scans with 3779 axial contrast-enhanced abdominal CT images. Following the experimental settings of TransUNet [4], we split the Synapse dataset into 18 scans for training, and 12 for validation.

Implementation details. The mmsegmentation codebase was used to train our model on 4 RTX 3090 GPUs. We used MSCAN [11] as a backbone network. Our model with MSCAN-T and MSCAN-B backbones were each named MetaSeg-T, MetaSeg-B, and our decoder was randomly initialized. For semantic segmentation evaluation, we adopted the mean Intersection over Union (mIoU) for ADE20K, Cityscapes, and COCO-Stuff datasets, and the Dice Similarity Score (DSC) for Synapse dataset. During the training, we applied the commonly used data augmentation such as random horizontal flipping, random scaling from 0.5 to 2.0 ratios and random cropping with the size of 512×512, 1024×1024, and 512×512 for ADE20K, Cityscapes, and COCO-Stuff datasets, respectively. For Synapse dataset, we used random rotation and flipping for data augmentation with the size of 224×224. We trained our models using AdamW optimizer for 160K iterations on ADE20K and Cityscapes, 160K iterations on COCO-Stuff, and 30K iterations on Synapse. The batch size was 16 for ADE20K and COCO-Stuff, 8 for Cityscapes, and 24 for Synapse. The

Method	DSC (%) ↑
V-Net [16]	68.81
DARR [9]	69.77
UNet [17]	70.11
R50+ViT [8]	71.29
AttnUNet [19]	71.70
R50+UNet [4]	74.68
R50+AttnUNet [4]	75.57
TransUNet [4]	77.48
MT-UNet [24]	78.59
SwinUNet [3]	79.13
HiFormer [12]	80.69
MetaSeg-B (Ours)	**82.78**

Table 3. Comparison with the previous state-of-the-art methods on Synapse dataset.

poly LR schedule with a factor of 1.0 and an initial learning rate of 6e-5 were used.

4.2. Comparison with State-of-the-Art Methods

ADE20K, Cityscapes, and COCO-Stuff datasets. In Table 1, we compared our MetaSeg performance with previous state-of-the-art methods on ADE20K, Cityscapes, and COCO-Stuff datasets. This comparison includes the number of the parameters, Floating Point Operations (FLOPs), and mIoU under both the single scale (SS) and multi-scale (MS) flip inference strategies. As shown in the Table 1, MetaSeg-T showed significant performance of 42.4% mIoU with only 4.7M parameters and 5.5 GFLOPs for ADE20K. Compared to SegNeXt-T that uses the same backbone [11], our MetaSeg-T achieved 1.3% higher mIoU and 16.7% lower GFLOPs on ADE20K. Moreover, our MetaSeg-T showed 0.3% and 1.0% higher mIoU with 5.2% and 16.7% lower GFLOPs on Cityscapes and COCO-Stuff, respectively. Our larger model, MetaSeg-B, also achieved competitive performance compared to previous state-of-the-art models. MetaSeg-B showed 48.5% mIoU with 12.9% less computations compared to SegNeXt-B on ADE20K. Furthermore, our MetaSeg-B achieved 82.7% and 45.8% mIoU with 8.9% and 12.9% less GFLOPs on Cityscapes and COCO-Stuff, respectively. These results demonstrated that our MetaSeg effectively captures the local to global contexts by leveraging the MetaFormer architecture up to the decoder with an efficient token mixer, our CRA.

Speed Benchmark Comparison. In Table 2, we present the speed benchmark comparisons without any additional accelerating techniques. For fair comparison, we measured Frames Per Second (FPS) of a whole single image of 1536×768 on Cityscapes using a single RTX3090 GPU. Compared to previous methods, our method achieved superior FPS with a higher mIoU score. This result demonstrates that a decrease in FLOPs of our method can lead to improvements in processing speed within the GPU.

Backbone	Method	Params(M)	ADE20K GFLOPs ↓	ADE20K mIoU (%) ↑
ConvNeXt [15]	UperNet [29]	60.2	234.7	46.1
	MetaSeg (Ours)	**37.2**	**31.0**	**46.1**
MobileNetV2 [18]	DeepLabV3 [5]	18.7	75.4	34.1
	MetaSeg (Ours)	**3.4**	**4.6**	**34.7**

Table 4. Ablation study on the effect of our proposed decoder for other CNN-based backbones on ADE20K validation set.

Stage2	Stage3	Stage4	Params(M)	ADE20K GFLOPs ↓	ADE20K mIoU (%) ↑
✓	✓	✓	**4.7**	**5.5**	**42.4**
✓	✓		4.3	5.4	40.4
	✓	✓	4.7	5.4	41.6
✓			4.0	5.3	40.4
	✓		4.2	5.3	41.0
		✓	4.5	5.3	41.4

Table 5. Ablation study for applying our proposed Global Meta Block to different stages.

Synapse dataset. In Table 3, we compared our MetaSeg with the previous methods on Synapse dataset using DSC (%). For a fair comparison, we utilized MetaSeg-B in the medical image segmentation task by considering the similar model size with the previous methods. As shown in Table 3, our MetaSeg-B sets the new state-of-the-art result with 82.78% DSC. This result showed a 2.09% higher DSC compared to HiFormer [12]. This indicates that our MetaSeg is effective even for the medical image segmentation task. Therefore, we demonstrated the high capabilities of our MetaSeg for application fields.

4.3. Ablation Study

Effectiveness of MetaSeg Decoder for Various CNN-based Backbones. In Table 4, we experimented with other CNN-based backbones to evaluate the effect of our MetaSeg decoder. In semantic segmentation, ConvNeXt [15] adopts UperNet [29] as its decoder and MobileNetV2 [18] adopts DeepLabV3 [5] as its decoder. For these CNN-based backbones, our decoder showed competitive performance with significant computational reduction of 86% and 93.9%. This indicates that our MetaSeg decoder is an efficient and effective architecture for various CNN-based backbone by enhancing the visual representation from encoder features.

Effectiveness of Global Meta Block. In Table 5, we verified the effectiveness of applying GMB in the decoder. We conducted experiments on various cases of applying or non-applying GMB to each Stage{2,3,4}. Following [11], we excluded the features from the first stage of the encoder in this experiment since they contain too much low-level information which degrades the segmentation performance. The results show that applying GMB to

Token Mixer	Params (M)	ADE20K	
		GFLOPs ↓	mIoU (%) ↑
AvgPool	4.4	5.4	40.7
DW Conv	4.4	5.4	40.4
Conv	5.3	5.8	41.1
SRA [26]	5.7	5.6	42.4
CRA (Ours)	4.7	**5.5**	**42.4**

Table 6. Ablation on the effect of our CRA by applying various token mixers to our Global Meta Block of the decoder. For a fair comparison, we utilized the same backbone, MSCAN-T [11].

Model	Token Mixer	Params (M)	FLOPs ↓		mIoU (%) ↑
			Attention (M)	Total (G)	
MetaSeg-T	SRA [26]	5.7	62.9	5.6	42.4
	CRA (Ours)	**4.7**	**32.4** (-48.5%)	**5.5**	**42.4**
MetaSeg-B	SRA [26]	33.7	125.8	31.1	48.0
	CRA (Ours)	**29.6**	**63.9** (-49.2%)	**30.4**	**48.5**

Table 7. Comparison our CRA with SRA [26] when applied to MetaSeg-T and MetaSeg-B as a token mixer on ADE20K.

Stage{2,3,4} is most effective structure compared to other cases. Especially, compared to Stage{3,4}, applying GMB to Stage{2,3,4} achieved 0.8% higher mIoU performance even though the parameters and GFLOPs are almost the same. This result indicates that capturing the global contexts through the GMB from all features extracted by the encoder Stage{2,3,4} is effective in improving the semantic segmentation performance.

Effectiveness of Global Modeling Token Mixer in Decoder. In Table 6, we conducted an experiment on applying various token mixers to our proposed meta block-based decoder. Through this experiment, we verify which token mixer is the most effective and efficient structure for the decoder when using MSCAN-T, a CNN-based backbone. The global context modeling token mixer (e.g. SRA and our CRA) showed the better mIoU performance compared to the local context modeling token mixer (e.g. pooling, depth-wise convolution and conventional convolution). This result demonstrates the importance of considering the global contexts in the decoder when using a CNN-based backbone.

Efficiency of Channel Reduction Attention. In Table 7, we focus on the parameter size and computational costs of our channel reduction self-attention (CRA) and the spatial reduction self-attention (SRA) [26] to compare which method is more efficient in terms of capturing global contexts. SRA is a widely used self-attention method that reduces the spatial resolution of the key-value by treating the token as a vector. In contrast, our CRA scalarizes each query and key token by reducing the channel dimension of the query and key into the one dimension. As shown in Table 7, our CRA reduces the computations of the query-key operation by a factor of C times, leading to a total computation reduction for the attention operation that is about twice as much as SRA. For a more detailed comparison of computations as described in eq.(4), we calculated the sum of the computations only for the attention operations in all stages of the decoder. As shown in Table 7, the attention operation of our CRA has 48% and 49% less FLOPs than the SRA on MetaSeg-T and MetaSeg-B, respectively. This indicates that our CRA is more efficient than the previous attention methods, as well as capturing the global context effectively.

Visualization of Features. In Fig. 4, we visualized the prediction map and the attention score map of our MetaSeg-T. The attention score map is the similarity score between

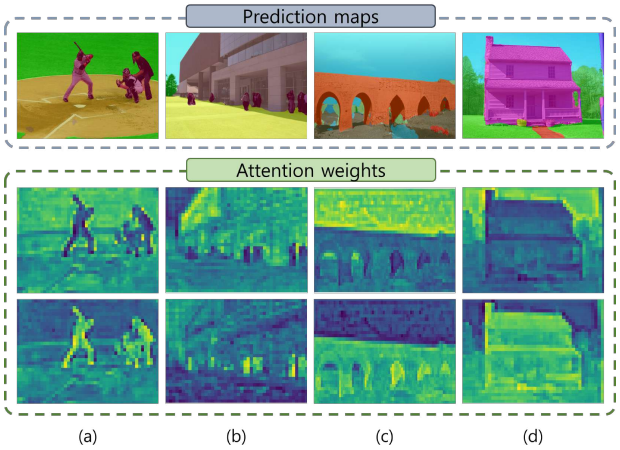


Figure 4. Visualization of our prediction maps and our attention score maps on ADE20K.

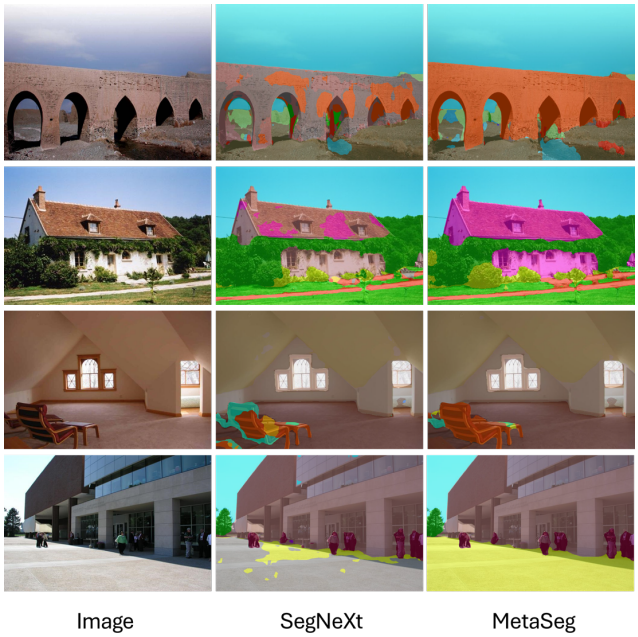


Figure 5. Qualitative results on ADE20K dataset. Compared to SegNeXt [11], our MetaSeg predicts more detailed for various categories.

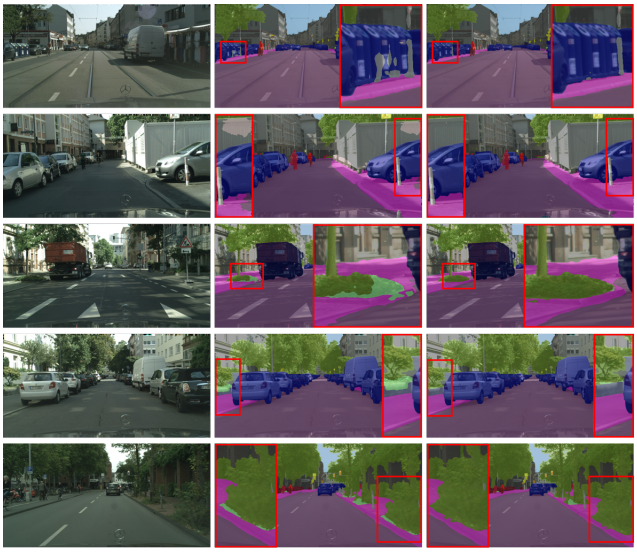


Figure 6. Qualitative results on Cityscapes dataset. The predictions of our MetaSeg are more precise than those of SegNeXt [11].

the query and key, which are applied our channel reduction attention method. As shown in Fig. 4 (a) and (b), the attention score maps showed significant similarity for people who are far apart. In Fig. 4 (c) and (d), the similarities of the large regions, such as a bridge and a house, were also captured clearly. These results indicate that our CRA can capture the meaningful similarity scores for extracting the global context features, even though the channel dimension of each pixel token has been reduced to the one dimension. By considering the globality well, our final prediction maps showed accurate segmentation results for the distant objects and the large regions.

4.4. Qualitative Results

In Figs. 5 and 6, we showed segmentation results of our MetaSeg and SegNext [11] on ADE20K and Cityscapes, respectively. Compared to SegNext, our MetaSeg better recognized the object details near the boundaries. This indicates that our model captures more useful visual contexts by leveraging the capacity of the MetaFormer architecture up to the decoder. In addition, our method segmented the large regions (e.g. road and bridge) more precisely. Furthermore, our model showed accurate predictions for far apart objects (e.g. person and house) that belong to the same category. These results indicate that our CRA can sufficiently consider the global contexts with the computational efficiency. In Fig. 7, we compared our predictions with HiFormer [12] on Synapse dataset. Our method predicted more accurately for the detailed regions. This indicates that our MetaSeg is effective for the application domain.

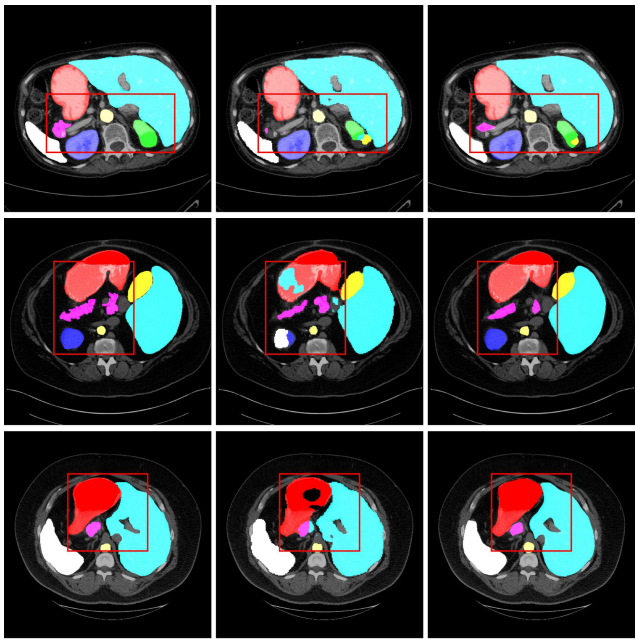


Figure 7. Qualitative results on Synapse dataset. Compared to HiFormer [12], the more elaborately predicted regions are highlighted with a red rectangle.

5. Conclusion

This paper proposed MetaSeg, a novel and powerful semantic segmentation network that effectively captures the local to global contexts by leveraging the MetaFormer architecture up to the decoder. Our MetaSeg showed that the capacity of the MetaFormer can be extended to the decoder as well as the backbone. In addition, we proposed a novel attention module for efficient semantic segmentation, Channel Reduction Attention (CRA) module, which can efficiently consider the globality by reducing the channel dimension of the query and key into the one dimension for low computational costs in the self-attention operation. Experiments demonstrated the effectiveness and efficiency of our method on three public semantic segmentation datasets and a medical image segmentation dataset for application.

Acknowledgements

This research was supported by Samsung Electronics Co., Ltd(IO201218-08232-01) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00260091) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge. 2, 5

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 5

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 6

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 5, 6

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1, 5

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5

[8] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 6

[9] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3d multi-organ segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 656–666. Springer, 2020. 6

[10] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 3

[11] 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. *arXiv preprint arXiv:2209.08575*, 2022. 1, 2, 3, 5, 6, 7, 8

[12] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 6, 8

[13] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022. 1, 2

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 6

[16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[17] 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. 6

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3, 6

[19] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019. 6

[20] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2263–2271, 2023. 1, 5

[21] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2

[22] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[24] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2390–2394. IEEE, 2022. 6

[25] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 1, 2

[26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 1, 2, 3, 5, 7

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1

[28] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1, 2

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

[30] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 2, 3, 5

[31] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 1, 2

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022. 2, 3

[33] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5