SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

XIE et al.

2021

1. Background

✓ Many state-of-the-art semantic segmentation frameworks are variants of popular architectures for image classification.

→ Designing backbone architectures has remained an active area in semantic segmentation.

✓ Witnessing the great success in NLP, there has been a recent surge of interest to introduce Transformers to vision tasks. (ViT)

1. Background

✓ ViT has two important limitations

1) ViT outputs single-scale low resolution features instead of multi-scale ones.
2) It has very high computational cost on large images

✓ Transformer Backborn Models: PVT, Swin Transformer, SETR

→ These methods mainly consider the design of the Transformer encoder, neglecting the contribution of the decoder for further improvements.

2. Introduction

✓ **This paper introduces "SegFormer"**

➢ A novel **positional-encoding-free** and **hierarchical** Transformer encoder.

➢ A **lightweight All-MLP decoder** design that yields a powerful representation without complex and computationally demanding modules.

➢ SegFormer sets **new a state-of-the-art** in terms of efficiency, accuracy and robustness in three publicly available semantic segmentation datasets.

3. Related Work

✓ Semantic Segmentation

✓ Transformer backbones

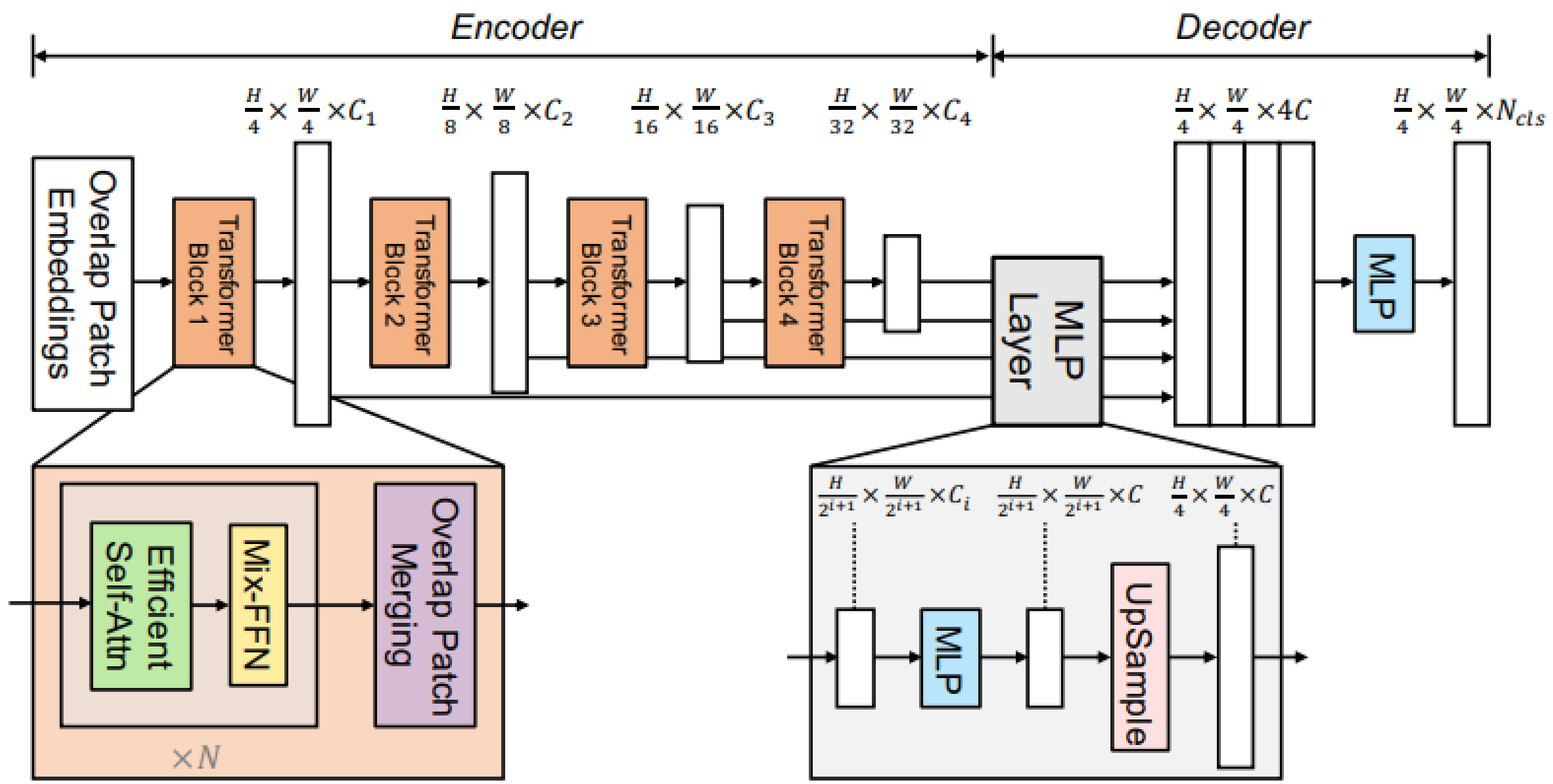
= ViT, CPVT, TNT, CrossViT, PVT

✓ Transformers for specific tasks

= DETR(Object Detection)

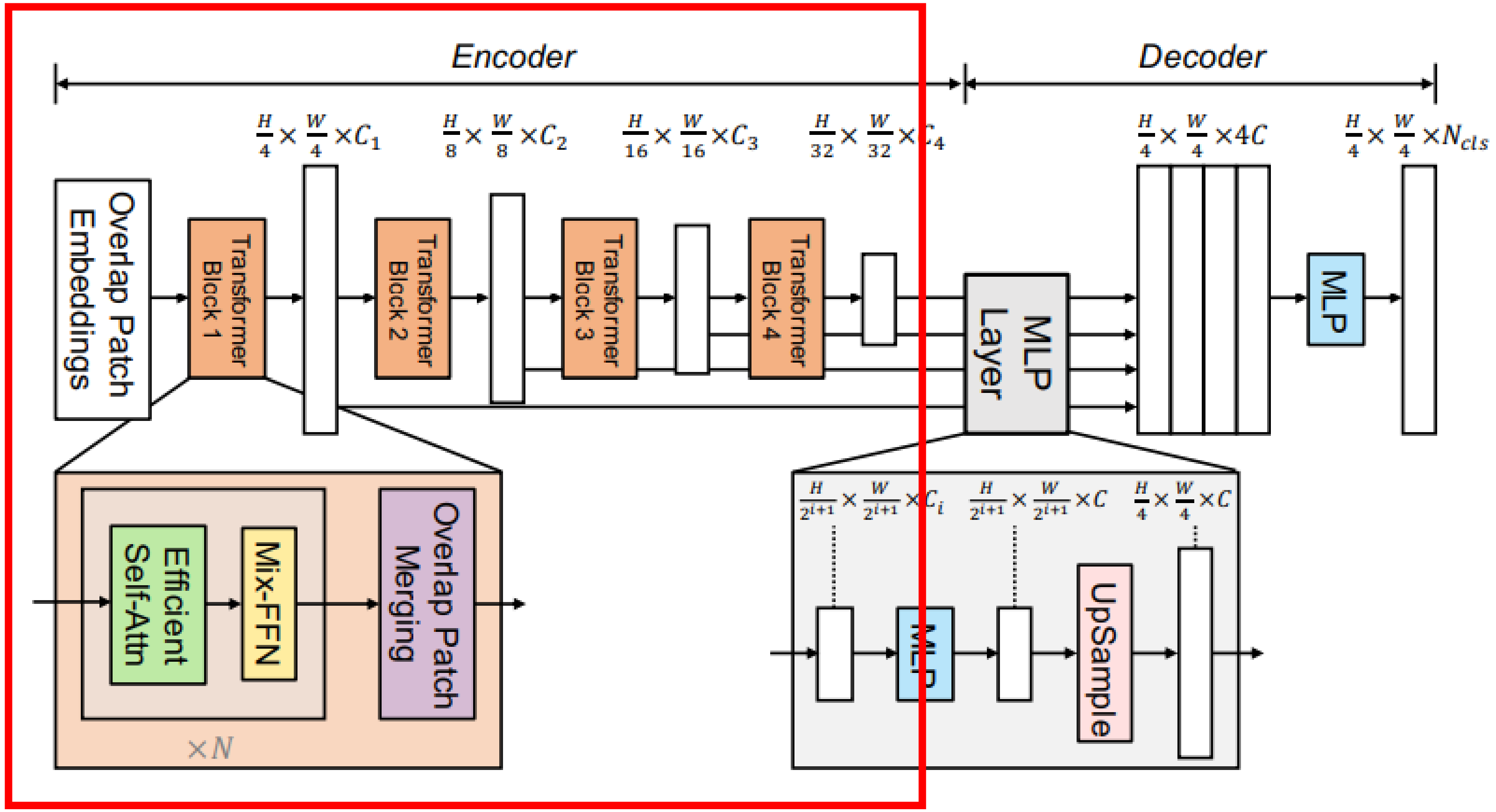
4. Method

❖ Structure: 2 main Model



4. Method

❖ Hierarchical Transformer Encoder



4. Method

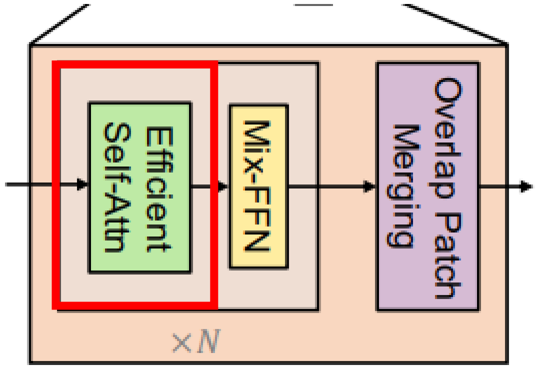
❖ Efficient Self-Attention

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_{head}}}\right)V.$$

$$\hat{K} = \text{Reshape}\left(\frac{N}{R}, C \cdot R\right)(K)$$

$$K = \text{Linear}(C \cdot R, C)(\hat{K}),$$

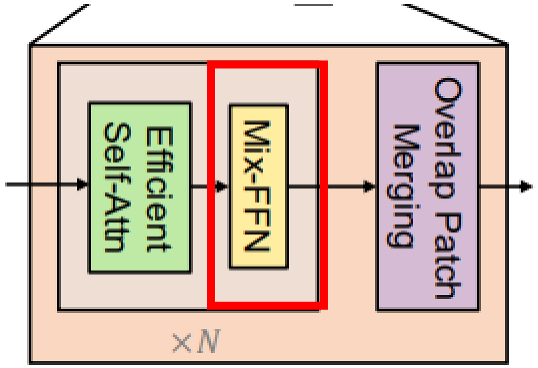
Complexity = $O(N^2) \gg O\left(\frac{N^2}{R}\right)$



4. Method

❖ Mix-FFN(Positional-Encoding-Free Design)

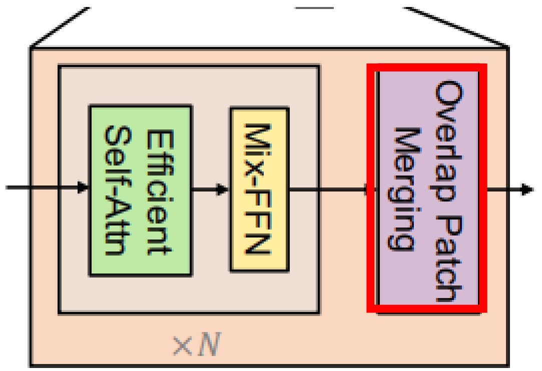
$$\mathbf{x}_{out} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(\mathbf{x}_{in})))) + \mathbf{x}_{in},$$



- ✓ Consider the effect of zero padding to the leak location information by directly using a 3 × 3 Conv in the feed-forward network.
- ✓ Mix-FFN mixes a 3 × 3 convolution and an MLP into each FFN.
- ✓ Show 3 × 3 Conv is sufficient to provide positional information for Transformers.
- ✓ Use depth-wise convolutions for reducing the number of parameters and improving efficiency.

4. Method

❖ Overlapped Patch Merging

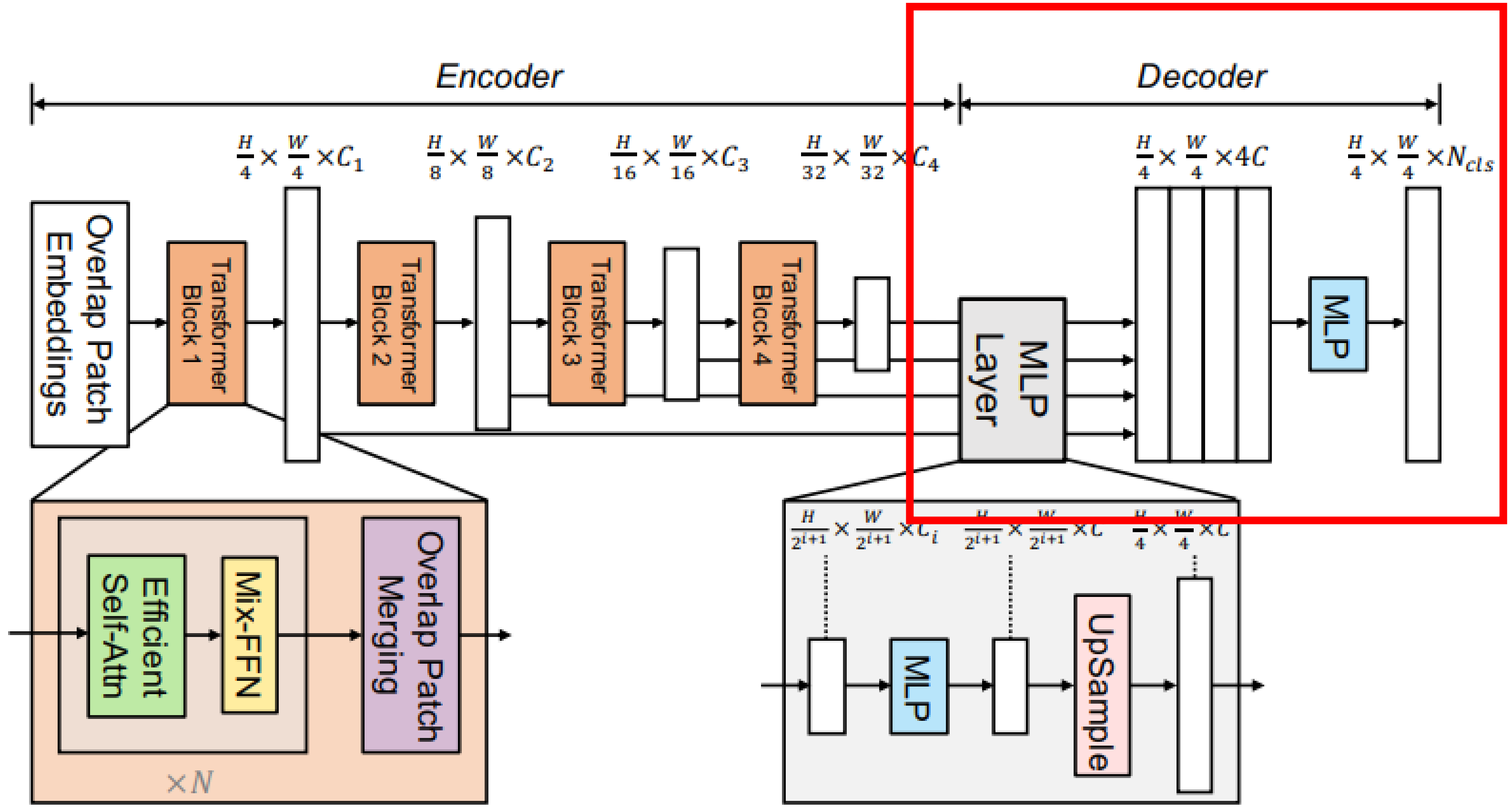


✓ Using Overlapped patch merging,
we can shrink our hierarchical features.

✓ Define patch_size(K), Stride(S) and padding_size(P) to
perform overlapping patch merging to produces features with the same size as the non-overlapping process.

4. Method

❖ Lightweight All-MLP Decoder



4. Method

❖ Lightweight All-MLP Decoder

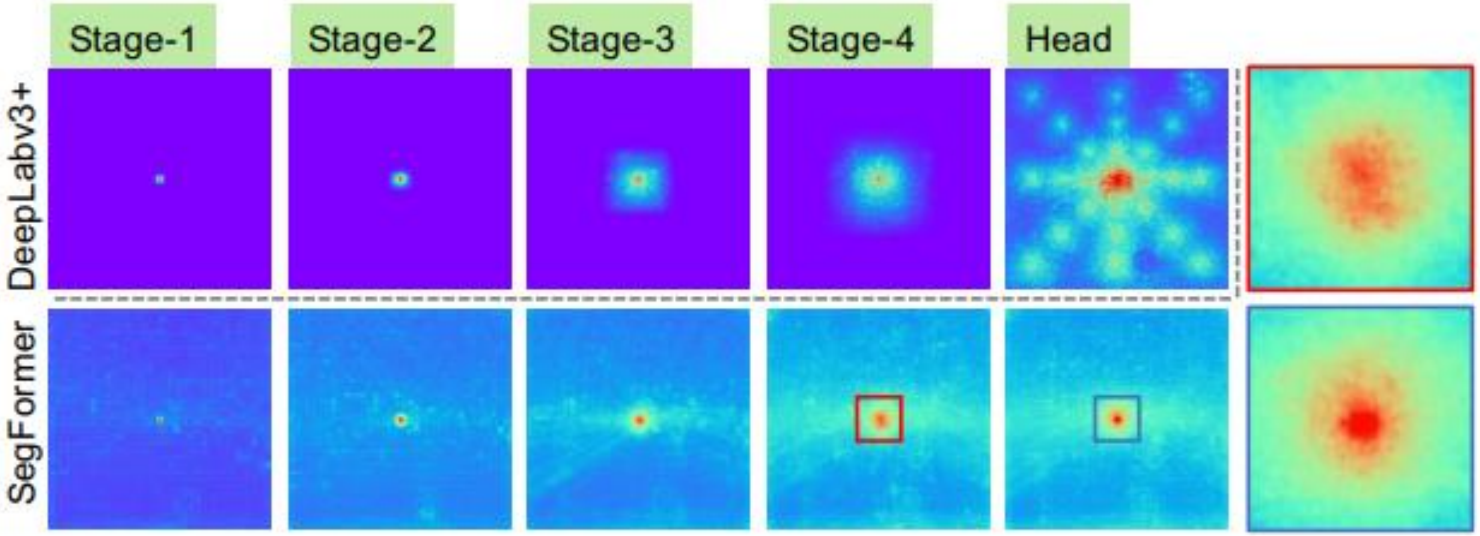
✓ The proposed All-MLP decoder consists of four main steps.

$$
\begin{aligned}
\hat{F}_i &= \text{Linear}(C_i, C)(F_i), \forall i \\
\hat{F}_i &= \text{Upsample}(\frac{W}{4} \times \frac{W}{4})(\hat{F}_i), \forall i \\
F &= \text{Linear}(4C, C)(\text{Concat}(\hat{F}_i)), \forall i \\
M &= \text{Linear}(C, N_{cls})(F),
\end{aligned}
$$

4. Method

❖ Effective Receptive Field Analysis

✓ Visualize ERFs of the four encoder stages and the decoder heads for both DeepLabv3+ and SegFormer.



5. Experiments

❖ Implementation details

- ✓ Datasets: Citycapes, ADE20K, COCO-Stuff
- ✓ Model
  - Encoder: pre-train the encoder on the Imagenet-1K dataset
  - Decoder: randomly initialize the decoder.
- ✓ Argumentation
  - ✓ Resize, Horizontal flipping, Random Cropping
- ✓ LR Schedule
- ✓ Metrics: mIoU

5. Experiments

Table 1: Ablation studies related to model size, encoder and decoder design.

(a) Accuracy, parameters and flops as a function of the model size on the three datasets. "SS" and "MS" means single/multi-scale test.

Encoder Model Size	Params Encoder	Params Decoder	ADE20K Flops ↓	ADE20K mIoU(SS/MS) ↑	Cityscapes Flops ↓	Cityscapes mIoU(SS/MS) ↑	COCO-Stuff Flops ↓	COCO-Stuff mIoU(SS) ↑
MiT-B0	3.4	0.4	8.4	37.4 / 38.0	125.5	76.2 / 78.1	8.4	35.6
MiT-B1	13.1	0.6	15.9	42.2 / 43.1	243.7	78.5 / 80.0	15.9	40.2
MiT-B2	24.2	3.3	62.4	46.5 / 47.5	717.1	81.0 / 82.2	62.4	44.6
MiT-B3	44.0	3.3	79.0	49.4 / 50.0	962.9	81.7 / 83.3	79.0	45.5
MiT-B4	60.8	3.3	95.7	50.3 / 51.1	1240.6	82.3 / 83.9	95.7	46.5
MiT-B5	81.4	3.3	183.3	51.0 / 51.8	1460.4	82.4 / 84.0	111.6	46.7

(b) Accuracy as a function of the MLP dimension C in the decoder on ADE20K.

C	Flops ↓	Params ↓	mIoU ↑
256	25.7	24.7	44.9
512	39.8	25.8	45.0
768	62.4	27.5	45.4
1024	93.6	29.6	45.2
2048	304.4	43.4	45.6

(c) Mix-FFN vs. positional encoding (PE) for different test resolution on Cityscapes.

Inf Res	Enc Type	mIoU ↑
768×768	PE	77.3
1024×2048	PE	74.0
768×768	Mix-FFN	80.5
1024×2048	Mix-FFN	79.8

(d) Accuracy on ADE20K of CNN and Transformer encoder with MLP decoder. "S4" means stage-4 feature.

Encoder	Flops ↓	Params ↓	mIoU ↑
ResNet50 (S1-4)	69.2	29.0	34.7
ResNet101 (S1-4)	88.7	47.9	38.7
ResNeXt101 (S1-4)	127.5	86.8	39.8
MiT-B2 (S4)	**22.3**	**24.7**	43.1
MiT-B2 (S1-4)	62.4	27.7	45.4
MiT-B3 (S1-4)	79.0	47.3	**48.6**

서울여자대학교 소프트웨어융합학과 김세희

5. Experiments

Table 2: **Comparison to state of the art methods on ADE20K and Cityscapes.** SegFormer has significant advantages on #Params (M), #Flops, #Speed and #Accuracy. Note that for SegFormer-B0 we scale the short side of image to {1024, 768, 640, 512} to get speed-accuracy tradeoffs.

	Method	Encoder	Params ↓	ADE20K			Cityscapes		
				Flops ↓	FPS ↑	mIoU ↑	Flops ↓	FPS ↑	mIoU ↑
Real-Time	FCN [1]	MobileNetV2	9.8	39.6	64.4	19.7	317.1	14.2	61.5
	ICNet [11]	-	-	-	-	-	-	30.3	67.7
	PSPNet [15]	MobileNetV2	13.7	52.9	57.7	29.6	423.4	11.2	70.2
	DeepLabV3+ [18]	MobileNetV2	15.4	69.4	43.1	34.0	555.4	8.4	75.2
	SegFormer (Ours)	MiT-B0	**3.8**	**8.4**	**50.5**	**37.4**	125.5	15.2	**76.2**
				-	-	-	51.7	26.3	75.3
				-	-	-	31.5	37.1	73.7
				-	-	-	**17.7**	**47.6**	71.9
Non Real-Time	FCN [1]	ResNet-101	68.6	275.7	14.8	41.4	2203.3	1.2	76.6
	EncNet [22]	ResNet-101	**55.1**	218.8	14.9	44.7	1748.0	1.3	76.9
	PSPNet [15]	ResNet-101	68.1	256.4	15.3	44.4	2048.9	1.2	78.5
	CCNet [39]	ResNet-101	68.9	278.4	14.1	45.2	2224.8	1.0	80.2
	DeeplabV3+ [18]	ResNet-101	62.7	255.1	14.1	44.1	2032.3	1.2	80.9
	OCRNet [21]	HRNet-W48	70.5	164.8	**17.0**	45.6	1296.8	**4.2**	81.1
	GSCNN [33]	WideResNet38	-	-	-	-	-	-	80.8
	Axial-DeepLab [72]	AxialResNet-XL	-	-	-	-	2446.8	-	81.1
	Dynamic Routing [73]	Dynamic-L33-PSP	-	-	-	-	**270.0**	-	80.7
	Auto-Deeplab [48]	NAS-F48-ASPP	-	-	-	44.0	695.0	-	80.3
	SETR [7]	ViT-Large	318.3	-	5.4	50.2	-	0.5	82.2
	SegFormer (Ours)	MiT-B4	64.1	**95.7**	15.4	51.1	1240.6	3.0	83.8
	SegFormer (Ours)	MiT-B5	84.7	183.3	9.8	**51.8**	1447.6	2.5	**84.0**

5. Experiments

Table 3: Ablation study of different Transformer encoders and different decoders. All the model are trained on ADE20K with 160K iterations.

Encoder	Decoder	mIoU	FPS	Decoder GFlops	Decoder Params (M)
MiT-B2	UperNet (Swin)	46.5	14.2	210.7	29.7
MiT-B2	MLA (SETR)	46.2	9.5	87.7	4.2
MiT-B2	MLP (Ours)	46.5	21.4	42.1	3.3
MiT-B5	UperNet (Swin)	50.7	5.3	210.7	29.7
MiT-B5	MLA (SETR)	50.9	3.8	87.7	4.2
MiT-B5	MLP (Ours)	51.0	9.8	42.1	3.3
Swin-T	MLP (Ours)	43.4	20.6	42.8	3.6
Swin-T	UperNet (Swin)	44.5	15.4	211.3	31.4
ViT-L	MLP (Ours)	47.7	4.7	0.6	0.6
ViT-L	MLA (SETR)	47.7	4.6	1.8	3.7

Table 4: **Comparison to state of the art methods on Cityscapes test set.** IM-1K, IM-22K, Coarse and MV refer to the ImageNet-1K, ImageNet-22K, Cityscapes coarse set and Mapillary Vistas.

Method	Encoder	Extra Data	mIoU
PSPNet [15]	ResNet-101	IM-1K	78.4
PSANet [41]	ResNet-101	IM-1K	80.1
CCNet [39]	ResNet-101	IM-1K	81.9
OCNet [19]	ResNet-101	IM-1K	80.1
Axial-DeepLab [72]	AxiaiResNet-XL	IM-1K	79.9
SETR [7]	ViT	IM-22K	81.0
SETR [7]	ViT	IM-22K, Coarse	81.6
SegFormer	MiT-B5	IM-1K	82.2
SegFormer	MiT-B5	IM-1K, MV	**83.1**

5. Experiments

Table 5: **Main results on Cityscapes-C.** "DLv3+", "MBv2", "R" and "X" refer to DeepLabv3+, MobileNetv2, ResNet and Xception. The mIoUs of compared methods are reported from [75].

Method	Clean	Blur				Noise				Digital				Weather			
		Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spatt	Fog	Frost
DLv3+ (MBv2)	72.0	53.5	49.0	45.3	49.1	6.4	7.0	6.6	16.6	51.7	46.7	32.4	27.2	13.7	38.9	47.4	17.3
DLv3+ (R50)	76.6	58.5	56.6	47.2	57.7	6.5	7.2	10.0	31.1	58.2	54.7	41.3	27.4	12.0	42.0	55.9	22.8
DLv3+ (R101)	77.1	59.1	56.3	47.7	57.3	13.2	13.9	16.3	36.9	59.2	54.5	41.5	37.4	11.9	47.8	55.1	22.7
DLv3+ (X41)	77.8	61.6	54.9	51.0	54.7	17.0	17.3	21.6	43.7	63.6	56.9	51.7	38.5	18.2	46.6	57.6	20.6
DLv3+ (X65)	78.4	63.9	59.1	52.8	59.2	15.0	10.6	19.8	42.4	65.9	59.1	46.1	31.4	19.3	50.7	63.6	23.8
DLv3+ (X71)	78.6	64.1	60.9	52.0	60.4	14.9	10.8	19.4	41.2	68.0	58.7	47.1	40.2	18.8	50.4	64.1	20.2
ICNet	65.9	45.8	44.6	47.4	44.7	8.4	8.4	10.6	27.9	41.0	33.1	27.5	34.0	6.3	30.5	27.3	11.0
FCN8s	66.7	42.7	31.1	37.0	34.1	6.7	5.7	7.8	24.9	53.3	39.0	36.0	21.2	11.3	31.6	37.6	19.7
DilatedNet	68.6	44.4	36.3	32.5	38.4	15.6	14.0	18.4	32.7	52.7	32.6	38.1	29.1	12.5	32.3	34.7	19.2
ResNet-38	77.5	54.6	45.1	43.3	47.2	13.7	16.0	18.2	38.3	60.0	50.6	46.9	14.7	13.5	45.9	52.9	22.2
PSPNet	78.8	59.8	53.2	44.4	53.9	11.0	15.4	15.4	34.2	60.4	51.8	30.6	21.4	8.4	42.7	34.4	16.2
GSCNN	80.9	58.9	58.4	41.9	60.1	5.5	2.6	6.8	24.7	75.9	61.9	70.7	12.0	12.4	47.3	67.9	32.6
SETR-DeiT	78.9	64.9	65.1	59.1	65.3	54.7	60.5	51.9	69.4	74.9	69.6	74.9	58.5	**44.3**	64.8	68.2	39.1
SegFormer-B5	**82.4**	**69.1**	**68.6**	**64.1**	**69.8**	**57.8**	**63.4**	**52.3**	**72.8**	**81.0**	**77.7**	**80.1**	**58.8**	40.7	**68.4**	**78.5**	**49.9**

6. Conclusion

- ✓ In this paper, we present SegFormer, a simple, clean yet powerful semantic segmentation method.
- ✓ SegFormer contains a positional-encoding-free, hierarchical Transformer encoder and a lightweight All-MLP decoder.
- ✓ SegFormer not only achieves new state of the art results on common datasets, but also shows strong zero-shot robustness.