U-MixFormer: UNet-like Transformer with Mix-Attention for Efficient Semantic Segmentation

Seul-Ki Yeom
Nota AI GmbH
Mariendorfer Damm 1, 12099 Berlin, Germany
skyeom@nota.ai

Julian von Klitzing
Nota AI GmbH
Mariendorfer Damm 1, 12099 Berlin, Germany
julian.von.klitzing@campus.tu-berlin.de

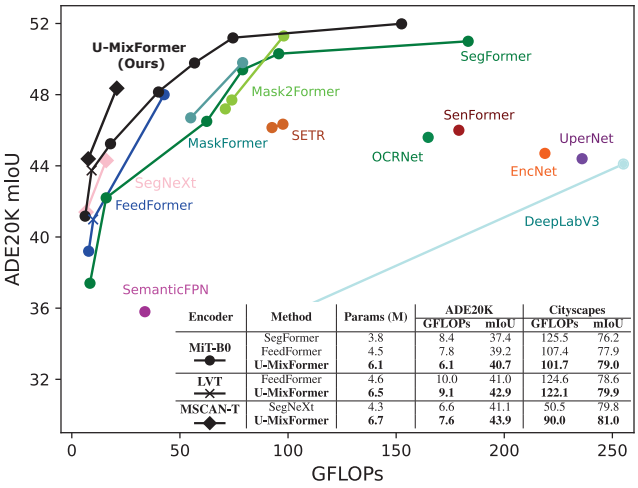
Abstract

Semantic segmentation has witnessed remarkable advancements with the adaptation of the Transformer architecture. Parallel to the strides made by the Transformer, CNN-based U-Net has seen significant progress, especially in high-resolution medical imaging and remote sensing. This dual success inspired us to merge both strengths, leading to the inception of a U-Net-based vision transformer decoder tailored for efficient contextual encoding. Here, we propose a novel transformer decoder, U-MixFormer, built upon the U-Net structure, designed for efficient semantic segmentation. Our approach distinguishes itself from the previous transformer methods by leveraging lateral connections between the encoder and decoder stages as feature queries for the attention modules, apart from the traditional reliance on skip connections. Moreover, we innovatively mix hierarchical feature maps from various encoder and decoder stages to form a unified representation for keys and values, giving rise to our unique mix-attention *module.*

Our approach demonstrates state-of-the-art performance across various configurations. Extensive experiments show that U-MixFormer outperforms SegFormer, FeedFormer, and SegNeXt by a large margin. For example, U-MixFormer-B0 surpasses SegFormer-B0 and FeedFormer-B0 with 3.8% and 2.0% higher mIoU and 27.3% and 21.8% less computation and outperforms SegNext with 3.3% higher mIoU with MSCAN-T encoder on ADE20K. Code available at https://github.com/julian-klitzing/u-mixformer.

1. Introduction

Semantic segmentation, a fundamental downstream task in computer vision, has consistently received increasing attention in industry and academia. The importance of semantic segmentation is highlighted by its widespread applications in real-world scenarios such as autonomous driving [25] and medical diagnosis [4]. Despite these advancements, achieving precise pixel-wise predictions remain challenging due to the need to balance global and local contexts.



Encoder	Method	Params (M)	ADE20K GFLOPs	ADE20K mIoU	Cityscapes GFLOPs	Cityscapes mIoU
MiT-B0	SegFormer	3.8	8.4	37.4	125.5	76.2
	FeedFormer	4.5	7.8	39.2	107.4	77.9
	U-MixFormer	**6.1**	**6.1**	**40.7**	**101.7**	**79.0**
LVT	FeedFormer	4.6	10.0	41.0	124.6	78.6
	U-MixFormer	**6.5**	**9.1**	**42.9**	**122.1**	**79.9**
MSCAN-T	SegNeXt	4.3	6.6	41.1	50.5	79.8
	U-MixFormer	**6.7**	**7.6**	**43.9**	**90.0**	**81.0**

Figure 1. Performance vs. computational efficiency on ADE20K (single-scale inference). U-MixFormer outperforms previous methods in all configurations.

The introduction of the fully convolutional network (FCN) [18] popularized the encoder-decoder structure, in which the encoder extracts high-level semantics and the decoder combines them with spatial details. While variants [1, 6] have improved this approach, traditional CNNs struggle to capture long-range context. This limitation has prompted interest in vision transformer-based methods for segmentation.

Transformer [24], initially designed for natural language processing, has been adapted for vision tasks with notable success thanks to the self-attention mechanism to capture global relationships within input sequences. Inspired by its success, Dosovitskiy *et al*. adapted it for vision tasks, leading to the Vision Transformer (ViT), which interprets im-

ages as sequences of embedded patches and processes them using the Transformer encoder [10]. This approach yielded remarkable results on ImageNet. Since the introduction of the ViT, numerous studies have been performed into its adaptation for semantic segmentation. The primary objectives have been two-fold: refining the encoder and crafting decoders that adeptly utilize features from the encoder stage.

There has been a notable shift towards utilizing the transformer's decoder structure in vision tasks. DETR pioneered this approach and integrated the transformer encoder-decoder framework into detection and segmentation [3]. Following DETR, Segmenter [23], MaskFormer [8], and Mask2Former [7] introduced decoders for mask prediction with global class labels, emphasizing high-level features. More recently, FeedFormer [22] proposed a decoder design that decodes high-level encoder features using only the lowest-level encoder feature. Despite the ongoing progress on transformer-based decoders for segmentation, these methods often rely on computationally intensive feature configurations within their attention mechanisms. Moreover, these methods exhibit inefficiencies in the propagation of feature maps across the decoder stages.

Traditionally, the U-Net architecture [21], characterized by its symmetric encoder-decoder structure, has been a favored choice for semantic segmentation, particularly in the medical field. This favor stems from U-Net's characteristics of effectively capturing and propagating hierarchical features across stages, a capability critical for boundary refinement and multi-scale feature integration.

Building upon these strengths, in this paper, we propose *U-MixFormer*, a novel UNet-like transformer decoder. U-MixFormer introduces a *mix-attention* module that integrates multi-stage features as keys and values, gradually propagating and refining them across decoder stages. This module successively propagates and refines features across decoder stages, effectively managing dependencies to capture context, refine boundaries, and integrate hierarchical representations with the global context modeling of Transformers. To the best of our knowledge, it is the first work to synergize the inherent strengths of U-Net with the transformative capabilities of Vision Transformers, particularly through a novel attention module for effectively harmonizing queries, keys, and values for semantic segmentation.

Our contributions are summarized as follows:

- **Novel Decoder Architecture with U-Net** We propose a novel powerful transformer-decoder architecture motivated by the U-Net for efficient semantic segmentation. Capitalizing on U-Net's proficiency in capturing and propagating hierarchical features, our design distinctively uses the lateral connections of a transformer encoder as query features. This approach ensures a harmonious fusion of high-level semantics and low-level structures.

- **Optimized Feature Synthesis for Enhanced Contextual Understanding** To improve the efficiency of our UNet-like transformer architecture, we mix and update multiple encoder and decoder outputs as integrated features for keys and values, resulting in our proposed *mix-attention* mechanism. This approach has not only rich feature representation for each decoder stage but also boosts contextual understanding.

- **Compatibility with Diverse Encoders** We demonstrate the compatibility of U-MixFormer combined with the existing popular encoders of both transformer-based (MiT [28] and LVT [29]) and CNN-based (MSCAN [11]) encoders.

- **Empirical Benchmarking** As shown in Figure 1, U-MixFormer achieves a new state-of-the-art in terms of computational cost as well as accuracy among semantic segmentation methods. It consistently outperforms light-weight, middle-weight, and even heavy-weight encoders. This superiority is demonstrated for the ADE20K and Cityscapes datasets, with notable performance on the challenging Cityscape-C dataset.

2. Related Work

2.1. Encoder Architectures

SETR [32] was the first architecture to adopt ViT as an encoder for semantic segmentation. Because ViT only divides the input image into patches, SETR produces single-scaled encoder features. PVT [27] and Swin Transformer [17] repeatedly group feature maps into new non-overlapping patches between encoder stages, thereby hierarchically generating multi-scale encoder features. Both methods also enhance the efficiency of the self-attention module by either reducing the spatial dimensions of keys and values (PVT) or grouping patches with shifted windows (Swin Transformer). SegFormer [28] reuses PVT's efficiency strategy while removing positional encodings and embedding feature maps into overlapping patches. In contrast to the previously mentioned methods, the encoders of SegNeXt [11] and LVT [29] employs convolutional attention mechanisms.

2.2. Decoder Architectures

DETR [3] was the first method to deploy a transformer decoder for semantic segmentation. Subsequent works [7, 8, 23] adapted DETR but also rely on object-learnable queries, which are computationally expensive, particularly when combined with multi-scale encoder features. In contrast,

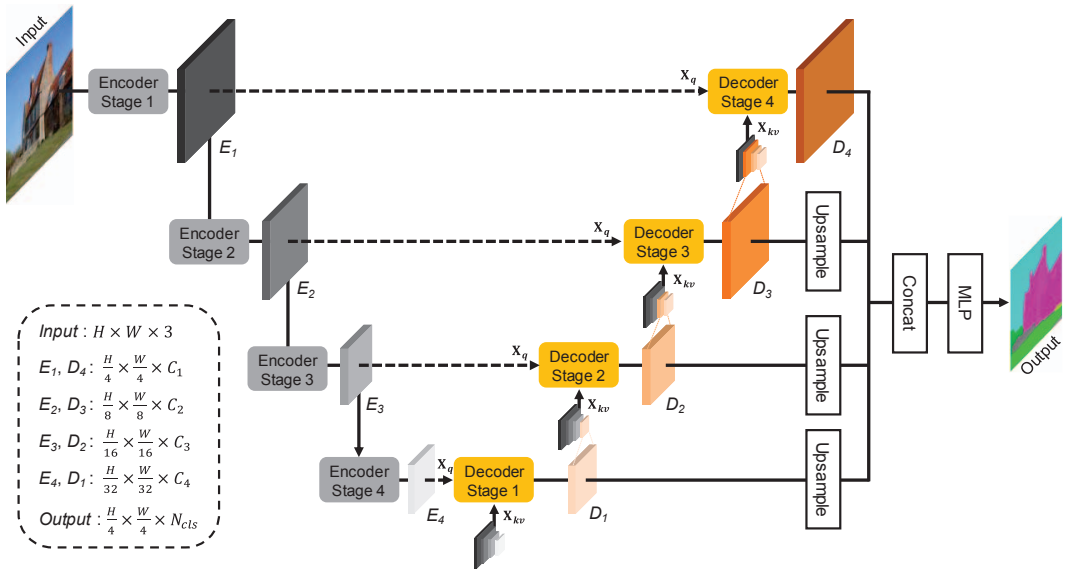


Figure 2. U-MixFormer architecture: Encoder (left) extracts multi-resolution feature maps from input image. U-MixFormer decoder (right) fuses lateral encoder outputs (dashed lines) as $\mathbf{X_q}$ with previous decoder stage outputs by incorporating them into $\mathbf{X_{kv}}$ using our mixed-attention mechanism. The combined feature $\mathbf{X}_{kv}$ used in mix-attention allows the query to find matches across all different stages, i.e., degrees of contextual granularity, which leads to enhanced feature refinement. Feature maps from all decoder stages are then concatenated, and a MLP predicts the output.

FeedFormer [22] directly utilizes the features from the encoder stages as feature queries, leading to enhanced efficiency. FeedFormer decodes the high-level encoder features (used as features for queries) with the lowest-level encoder feature (used as features for keys and values). However, this setup processes the feature maps individually, without incremental propagation of feature maps between decoder stages, thereby missing the opportunity for more incremental refinement to improve object boundary detection. Additionally, other recent MLP or CNN-based decoders [11, 28] also lack incremental propagation of decoder features.

2.3. UNet-like Transformer

In both the medical and remote sensing domains, efforts have been made to transition the UNet architecture from a CNN-based framework to a transformer-based one. TransUNet [5] marked the first successful endeavor to incorporate the Transformer into medical image segmentation, using ViT in conjunction with their CNN encoder. Other hybrid approaches are presented in [12,20,26]. Cao *et al*. introduced Swin-UNet, the first fully transformer-based UNet-like architecture [2]. This design features heavyweight Swin Transformer stages for both the encoder and decoder, preserving the lateral connections between them as skip connections. *In contrast to Swin-UNet, our approach employs lighter-weight decoder stages, rendering it suitable for a wider range of downstream tasks. Furthermore, we interpret the lateral connections as features for queries instead of as skip connections and incorporate a unique attention mechanism.*

3. Proposed Method

This section introduces U-MixFormer, a novel UNet-like transformer decoder architecture for semantic segmentation. In general, our decoder is composed of as many stages $i \in \{1, ..., N\}$ as there are encoder stages. For clarity, Figure 2 provides a visual overview of this architecture, exemplified with a four-stage ($N = 4$) hierarchical encoder such as MiT, LVT, or MSCAN [11, 28, 29].

First, the encoder processes an input image with $H \times W \times 3$. The four stages $i \in \{1, ..., 4\}$ yield hierarchical, multi-resolution features $\mathbf{E}_i$ with $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$. Second, our decoder stages i sequentially generate refined features $\mathbf{D}_{4-i+1}$ by performing *mix-attention* where features for queries $\mathbf{X}_q^i$ equal the respective lateral encoder feature map. The features for keys and values $\mathbf{X}_{kv}^i$ are given by a mix of encoder and decoder stages. Notably, our decoder mirrors the dimensions of the encoder stage outputs. Third, the decoder features are upsampled using bilinear interpolation to match the height and width of $\mathbf{D}_1$. Finally, the concatenated features are processed by an MLP to predict the segmentation map with $H/4 \times W/4 \times 3$.

3.1. Mix-Attention

Attention modules used in transformer blocks compute the scaled dot-product attention for queries $\mathbf{Q}$, keys $\mathbf{K}$ and

values $\mathbf{V}$ as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V} \quad (1)$$

where d_k is the embedding dimension of the keys and $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are obtained from linear projections of selected features.

Central to our method is the selection of features $\mathbf{X_{kv}}$ that are to be projected to keys and values, which results in our proposed mix-attention mechanism. A comparison between traditional self-, cross-, and the novel *mix-attention* is illustrated in Figure 3.

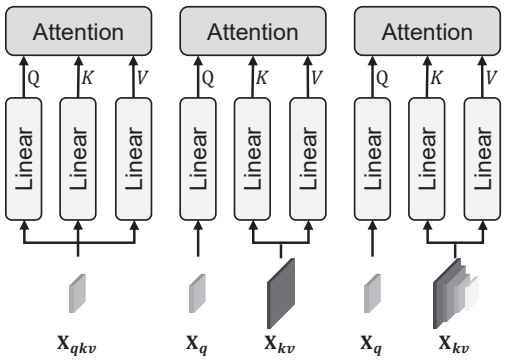


Figure 3. Comparison of three different attention modules: Self- (left), cross- (middle) and mix-attention (right).

In self-attention, the features used to generate queries, keys and values are identical ($\mathbf{X_{qkv}}$) and originate from the same source, i.e. the same encoder/decoder stage. Cross-attention employs two distinct features, $\mathbf{X_q}$ and $\mathbf{X_{kv}}$, each derived from a *single* unique source. In contrast, our mix-attention mechanism utilizes a mixed feature for $\mathbf{X_{kv}}$ sourced from *multiple* multi-scale stages. This concept allows the query to find matches across all different stages, i.e. degrees of contextual granularity, thereby facilitating enhanced feature refinement. The efficacy of this method is validated by our experiments in the Ablation Studies section.

The selection of a feature set $\mathcal{F}^i$ for a decoder stage i is formalized piecewise as follows:

$$\mathcal{F}^i = \begin{cases} \{\mathbf{E}_j\}_{j=1}^{N} & \text{if } i = 1 \\ \{\mathbf{E}_j\}_{j=1}^{N-i+1} \cup \{\mathbf{D}_j\}_{j=1}^{i-1} & \text{otherwise} \end{cases} \quad (2)$$

where for the first decoder stage ($i = 1$) all encoder features are selected. For subsequent stages, previously computed decoder stage outputs are propagated by replacing their lateral encoder counterparts in $\mathcal{F}^i$.

To align the spatial dimensions of the features in $\mathcal{F}^i$, we adapt the *spatial reduction* procedure introduced by PVT [27] :

$$\begin{aligned} \hat{\mathbf{F}}_j^i &= AvgPool(pr_j, pr_j)(\mathcal{F}_j^i), \forall j \in \{1, ..., N-1\} \\ \hat{\mathbf{F}}_j^i &= Linear(C_j, C_j)(\hat{\mathbf{F}}_j^i), \forall j \in \{1, ..., N-1\} \end{aligned} \quad (3)$$

where $\mathcal{F}_j^i$ denotes the j-th element of feature set $\mathcal{F}^i$, and pr_j is the pooling ratio which aligns the size with the smallest feature map $\mathcal{F}_N^i$. The operations AvgPool and Linear are configured as AvgPool($kernelSize$, $stride$)($\cdot$) and Linear(C_{in}, C_{out})($\cdot$), respectively.

The spatially aligned features are concatenated along the channel dimension to form a mixed feature $\mathbf{X}_{kv}^i$ for keys and values.

$$\mathbf{X}_{kv}^i = Concat(\{\hat{\mathbf{F}}_j^i\}_{j=1}^{N-1} \cup \{\mathcal{F}_N^i\}) \quad (4)$$

3.2. Decoder Stage

We adapt the traditional transformer decoder block, by discarding the self-attention module as recommended by FeedFormer [22]. Additionally, we replace the cross-attention module with our proposed mix-attention module. The resulting structure is depicted in Figure 4.

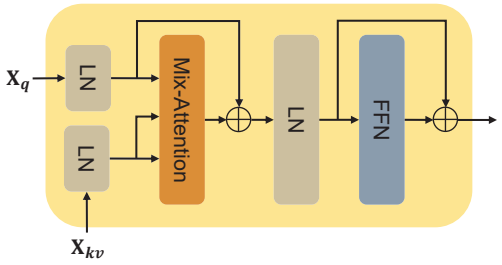


Figure 4. Structure of decoder stage employing our mix-attention mechanism and taking features for queries $\mathbf{X}_q$ and a mixed feature for keys and values $\mathbf{X}_{kv}$ as input.

Using layer normalization (LN) and a feedforward network (FFN), the output for a $DecoderStage_i$ is computed as follows:

$$\begin{aligned} \mathbf{A}_i &= LN(MixAtt.(LN(\mathbf{X}_{kv}^i, \mathbf{X}_q^i)) + LN(\mathbf{X}_q^i)) \\ DecoderStage_i &= \mathbf{D}_{N-i+1} = FFN(\mathbf{A}_i) + \mathbf{A}_i \end{aligned} \quad (5)$$

where $MixAtt.$ denotes our proposed mix-attention.

3.3. Relationship to UNet Architectures

We introduced U-MixFormer as a UNet-like architecture. However, we want to underline the main differences between our approach and other UNet-like variants.

- Because we view the lateral connections as features for queries, our decoder feature maps implicitly upsample spatial resolution between decoder stages in a data-driven way. It arises from the rules of matrix multiplication in the attention module, and our UNet-like design choice to propagate $\mathbf{X}_q$ through the lateral connections i.e. the output's dimension of a decoder stage i is determined by $\mathbf{X}_q^i$'s dimension.

Table 1. Performance comparison with the state-of-the-art methods on ADE20K and Cityscapes. The values in brackets indicate the improvement (+) of U-MixFormer over its respective counterparts with the same encoders for mIoU metrics.

	Method	Encoder	ADE20K		Cityscapes		Params (M) ↓
			mIoU ↑	GFLOPs ↓	mIoU ↑	GFLOPs ↓	
Light-weight	FCN [18]	MobileNetV2	19.7	39.6	61.5	317.1	9.8
	PSPNet [31]	MobileNetV2	29.6	52.9	70.2	423.4	13.7
	DeepLabV3+ [6]	MobileNetV2	34.0	69.4	75.2	125.5	15.4
	SwiftNetRN [19]	ResNet-18	-	-	75.4	104.0	11.8
	Semantic FPN [15]	ConvMLP-S	35.8	33.8	-	-	12.8
	SegFormer [28]	MiT-B0	37.4	8.4	76.2	125.5	3.8
	FeedFormer [22]	MiT-B0	39.2	7.8	77.9	107.4	4.5
	FeedFormer [22]	LVT	41.0	10.0	78.6	124.6	4.6
	SegNeXt [11]	MSCAN-T	41.1	6.6	79.8	50.5	4.3
	U-MixFormer [Ours]	MiT-B0	**41.2 (+3.8)**	6.1	**79.0 (+2.8)**	101.7	6.1
		LVT	**43.7 (+2.7)**	9.1	**79.9 (+1.3)**	122.1	6.5
		MSCAN-T	**44.4 (+3.3)**	7.6	**81.0 (+1.2)**	90.0	6.7
Middle-weight	CCNet [13]	ResNet-101	43.7	278.4	79.5	2224.8	68.9
	DeepLabV3+ [6]	ResNet-101	44.1	255.1	80.9	2032.3	52.7
	Auto-DeepLab [16]	Auto-DeepLab-L	-	-	80.3	695.0	44.4
	OCRNet [30]	HRNet-W48	45.6	164.8	81.1	1296.8	70.5
	Seg-S-/16 [23]	ViT-S	45.4	31.8	-	-	22.0
	MaskFormer [8]	Swin-T	46.7	55.0	-	-	42.0
	Mask2Former [7]	Swin-T	47.7	74.0	82.1	-	47.0
	Mask2Former [7]	ResNet-101	47.8	90.0	-	-	63.0
	SegFormer [28]	MiT-B1	42.2	15.9	78.5	243.7	13.7
	SegNeXt [11]	MSCAN-S	44.3	15.9	81.3	124.6	13.9
	SegFormer [28]	MiT-B2	46.5	62.4	81.0	717.1	27.5
	FeedFormer [22]	MiT-B2	48.0	42.7	81.5	522.7	29.1
	U-MixFormer [Ours]	MiT-B1	**45.2 (+3.0)**	17.8	**79.9 (+1.4)**	246.8	24.0
		MiT-B2	**48.2 (+1.7)**	40.0	**81.7 (+0.7)**	515.0	35.8
		MSCAN-S	**48.4 (+4.1)**	20.8	**81.8 (+0.5)**	154.0	24.3

- Our approach uses all decoder stages to predict the segmentation map, not only the final one.

- The feature map of the last decoder stage yields a resolution of $H/4$, $W/4$, whereas others restore the original spatial resolution H, W.

4. Experiments

4.1. Experimental Settings

4.1.1 Datasets

Experiments are conducted on two popular benchmark datasets: ADE20K [33] and Cityscapes [9]. ADE20K is a rigorous scene parsing benchmark highlighting 150 intricate semantic concepts, split into 20,210 images for training and 2,000 for validation. Cityscapes incorporates 19 densely annotated object categories from urban imagery, aggregating 5,000 images of a high resolution of 2048×1024. It also introduces 19,998 roughly annotated images for enhanced model training.

4.1.2 Implementation Details

To evaluate the versatility of U-MixFormer across various encoder complexities, we incorporated three distinct encoder backbones: the Mix Transformer (MiT) [28], Light Vision Transformer (LVT) [29], and the multi-scale convolutional attention-based encoder (MSCAN) [11]. In detail, we utilized MiT-B0, LVT, and MSCAN-T for our light-weight variants and MiT-B1/2 and MSCAN-S for the middle-weight architectures, while the heavier variants included MiT-B3/4/5. The embedding dimensions were 128 in the final MLP phase for light-weight models and 768 for others. Section A.1 of the supplementary material provides additional information about the training and evaluation setting.

4.2. Experimental Results

We compare our results with existing semantic segmentation methods on the ADE20K and Cityscapes datasets. Table 1 and 2 showcases our results, including the number of parameters, Floating Point Operations (FLOPs), and mIoU across both datasets. As shown in Figure 1, we plot the performance-computation curves of different methods

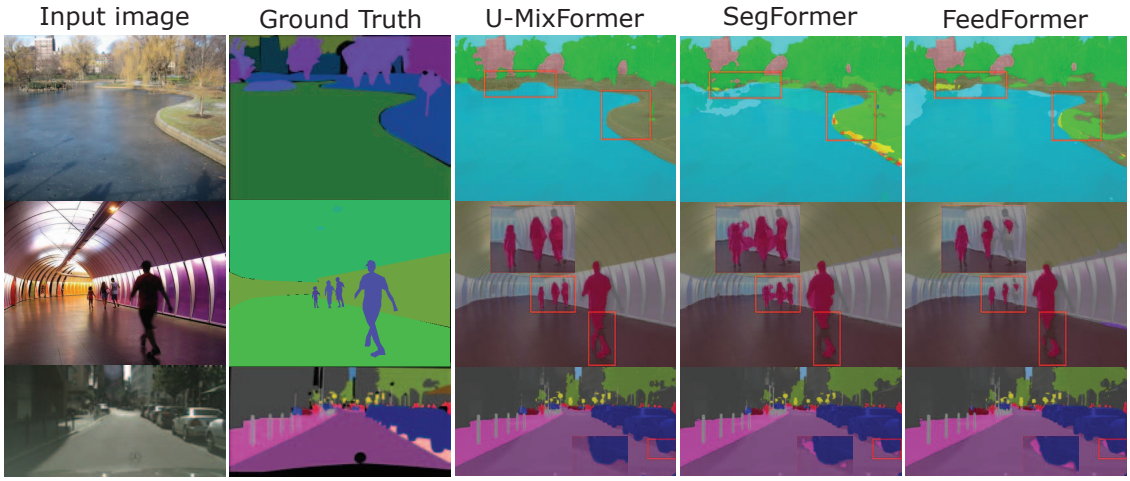


Figure 5. Qualitative analysis on ADE20K and Cityscapes datasets: We select SegFormer and FeedFormer as benchmarks since they share the same encoder. Our observations indicate that U-MixFormer outperforms these methods, particularly in segmenting complex object details such as the boundaries between the lake and floor and across human objects.

Table 2. Performance comparison for heavyweight encoders on ADE20K.

Method	Encoder	mIoU ↑	GFLOPs ↓	Params (M) ↓
Seg-B-Mask/16	ViT-B	48.7	129.4	102.4
Seg-L-Mask/16	ViT-L	51.8	399.9	333.2
SETR-PUP	ViT-L	48.6	425.9	318.3
SETR-MLA	ViT-L	48.6	368.4	310.5
MaskFormer	Swin-S	49.8	79.0	63.0
SegFormer	MiT-B3	49.4	79.0	47.3
SegFormer	MiT-B4	50.3	95.7	64.1
SegFormer	MiT-B5	51.0	183.3	84.7
U-MixFormer	MiT-B3	49.8	56.8	55.7
U-MixFormer	MiT-B4	50.4	73.4	72.4
U-MixFormer+	MiT-B4	51.2	74.5	72.4
U-MixFormer	MiT-B5	51.9	149.5	93.0
U-MixFormer+	MiT-B5	52.0	152.0	93.0

on the Cityscapes and ADE20K validation set.

4.2.1 Light- and Middle-weight Models

In Table 1, the upper section shows the performance of the light-weight models. As shown in the table on ADE20K, our light-weight U-MixFormer-B0 establishes 41.2% mIoU with 6.1M parameters and 6.1 GFLOPs, outperforming all other light-weight counterparts in terms of FLOPs and mIoU demonstrating a better trade-off of performance-computation. Notably, compared to SegFormer and FeedFormer, which employ the same encoder (MiT-B0), U-MixFormer achieved mIoU improvement of 3.8% and 2.0% while reducing computation by 27.3% and 21.8%. The performance disparity is even more pronounced on Cityscapes, where our model achieves 79.0% mIoU with only 101.7 GFLOPs, indicating mIoU increases of 2.8% and 1.1% and computation reductions of 18.9% and 5.3% compared to SegFormer-B0 and FeedFormer-B0, respectively. When utilizing LVT, our model's performance observes further enhancements with 2.7% and 1.3% of increased mIoU across datasets. Furthermore, our U-MixFormer with MSCAN-T, a recent encoder from SegNeXt, also delivers standout results: 44.4% and 81.0% mIoU on ADE20K and Cityscapes, respectively, using 6.7M parameters.

The latter section of Table 1 shifts the focus to middle-weight models, where our approach continues to demonstrate superior results, maintaining its edge over competitors.

4.2.2 Heavy-weight Models

As detailed in Table 2, U-MixFormer outperforms SegFormer when paired with the same heavy encoders, specifically MiT-B3/4/5. For instance, on ADE20K, U-MixFormer-B3 yields 49.8% mIoU with only 56.8 GFLOPs. This shows a 0.4% improvement in mIoU and a 28.1% reduction in computation compared to SegFormer-B3. We additionally hypothesized that enlarging the model size (changing from MiT-B0 to MiT-B5) would allow for richer contextual information extraction from the encoder stage, potentially boosting performance. For this reason, we trained and evaluated heavy-weight model variants, MiT-B4 and MiT-B5, introducing a method to extract additional keys and values from the encoder's 3rd stage midpoint where stack lots of attention blocks. We have termed this enhanced variant, U-MixFormer+. For MiT-B4 and MiT-B5 configurations, we extracted 5 and 6 keys and values, respectively, to facilitate mix-attention. As a result, we observed a reasonable performance improvement of 0.8% for MiT-B4 and 0.1% for MiT-B5 when integrating more con-

Table 3. Ablation study showing the effectiveness of mix-attention and U-Net like architecture.

Methods	Params (M)	GFLOPs	mIoU
Baseline - FeedFormer (Cross-Attention)	4.5	7.8	39.2
Mix-Attention	6.0	5.7	39.9 (+0.7)
Cross-Attention + U-Net	5.0	6.1	40.8 (+0.9)
Mix-Attention + U-Net (proposed method)	**6.1**	**6.1**	**41.2 (+0.4)**

textual data from the encoder, with only a marginal increase in computational demands.

4.3. Qualitative Results

Figure 5 presents the qualitative results of U-MixFormer, FeedFormer, and SegFormer, all utilizing identical encoders on the ADE20K and Cityscapes datasets. U-MixFormer excels in segmenting intricate object details and challenging regions more clearly than other approaches. It can significantly identify semantically relevant areas and object details, underlying its ability to learn contextual feature representations from the multi-stage encoder for efficient segmentation.

4.4. Ablation Studies

4.4.1 Effectiveness of Mix-Attention and U-Net like Architecture

In Table 3, we systematically evaluate model performance for various design choices. To guarantee fairness in our comparison, all models are trained and evaluated under a uniform random seed. We designate FeedFormer, which is based on conventional cross-attention, as our baseline. Integrating contextual information from multiple encoder stages via the *mix-attention* module results in a 0.7% mIoU improvement while reducing computational costs. Adapting a U-Net transformer decoder without mix-attention boosts mIoU by 0.9% with a slight FLOPs increase(+0.4). Remarkably, by applying the mix-attention module for the U-Net architecture, the model's performance increases to 41.2%, signifying a substantial enhancement over the traditional cross-attention in U-Net like configurations.

4.4.2 Robustness in Image Corruption

In safety-critical domains such as autonomous driving and intelligent transportation systems, the robustness of image segmentation is paramount. In this respect, we assessed the robustness of U-MixFormer against corruption and disturbances. Following [14], we introduced Cityscapes-C, an augmented version of the Cityscapes *val*, encompassing 16 algorithmic corruptions spanning noise, blur, weather, and digital categories. We compare our U-MixFormer with SegFormer and FeedFormer, both sharing the same encoder. The findings, shown in Table 4, highlight the superior robustness of U-MixFormer. In particular, U-MixFormer consistently outperformed the baseline methods across all corruption types, demonstrating exceptional robustness in noise (e.g., Gaussian and shot noise) and challenging weather conditions (e.g., snowy and frosty environments), with improvements of up to 33.3%. Notably, it achieved significant improvements across all corruption categories, with a remarkable margin of up to 20.0% and 33.3% against shot noise and 21.8% and 19.2% in snowy conditions. These results demonstrate the robustness of U-MixFormer, making it an ideal choice for applications where safety and reliability are crucial.

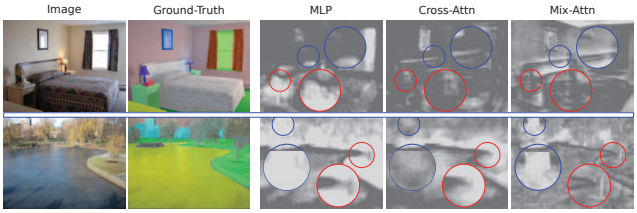


Figure 6. Qualitative comparison between MLP, Cross attention, and the proposed mix-attention approach on ADE20K. We can clearly see the border segment of objects (Top row: wall/bed (blue) and ramp/box/bed (red), bottom row: building/background (blue) and lake/floor (red)), which can lead to boost semantic segmentation performance.

4.4.3 Effectiveness of Mix-Attention

To validate the efficacy of the mix-attention module based on U-Net microscopically, we conducted an ablation study by extracting feature maps from identical locations compared to the common study (MLP and Cross-attention) in Figure 6. We observed a distinct enhancement in our approach. Specifically, the feature map depicted object details with greater precision and clearly delineated the boundaries between objects. This observation demonstrates that the proposed method can significantly segment and capture distinguishable visual details locally and globally.

4.4.4 Effectiveness of Decoding head with the same encoder

We aimed to benchmark our methodology against the state-of-the-art using a consistent encoder, specifically adopting MiT and MSCAN. These multi-stage design encoders have received significant attention for their efficiency and innovative design. As highlighted by results in SegNeXt, both MiT and MSCAN achieved higher mean Intersection over Union (mIoU) scores and reduced computational overhead,

Table 4. The average mIoU values for both the clean and corrupted variants of the Cityscapes validation set were computed for three semantic segmentation methods, all of which utilize the same encoder (MiT). mIoU is averaged across all applicable severity levels except for the noise corruption category that takes into account for the first three out of the five severity levels.

Method	Clean	Blur				Noise				Digital				Weather			
		Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spat	Fog	Frost
SegFormer	76.2	59.3	59.8	48.7	60.0	26.2	27.5	31.1	52.0	73.2	66.6	72.0	38.3	21.5	53.2	67.2	31.8
FeedFormer	77.9	59.5	60.0	50.3	59.4	21.3	21.5	25.6	47.4	74.0	66.7	73.6	39.1	22.2	52.0	65.9	32.1
U-MixFormer	**79.0**	**62.4**	**61.7**	**51.7**	**62.1**	**32.8**	**34.4**	**38.4**	**56.7**	**76.2**	**67.2**	**74.8**	**42.7**	**27.5**	**56.1**	**68.0**	**33.9**

Table 5. Performance comparison with MSCAN encoders on ADE20K and Cityscapes, sorted in ascending order by mIoU.

Method	Encoder	ADE20K		Cityscapes		Params (M) ↓
		mIoU ↑	GFLOPs ↓	mIoU ↑	GFLOPs ↓	
SegNeXt [11]	MSCAN-T	41.1	6.6	79.8	50.5	4.3
	MSCAN-S	44.3	15.9	81.3	124.6	13.9
	MSCAN-B	48.5	34.9	82.6	275.7	27.6
U-MixFormer [Ours]	MSCAN-T	**44.4 (+3.3)**	7.6	**81.0 (+1.2)**	90.0	6.7
	MSCAN-S	**48.4 (+4.1)**	20.8	**81.8 (+0.5)**	154.0	24.3
	MSCAN-B	**49.9 (+1.4)**	34.1	**83.2 (+0.6)**	259.7	37.2

evident from their fewer FLOPs. This analysis is vital to highlight the advantages of our method compared to these established encoders. As depicted in Table 1 and 2, our U-MixFormer consistently outperforms across various model complexities from B0 to B5 in both mIoU and FLOPs. Table 5 further highlights U-MixFormer's superiority, achieving a 3.3%, 4.1%, and 1.4% mIoU increase on ADE20K over the SegNeXt with the same encoders (i.e. MSCAN-T, -S, and -B, respectively). Furthermore, MSCAN-S closely rivals the performance of the next heavier MSCAN encoder (i.e. SegNeXt with MSCAN-B) with smaller number of parameters as well as smaller FLOPs. These findings demonstrate U-MixFormer is a promising decoder architecture in semantic segmentation.

4.4.5 Limitation and Future Works

Despite the competitive results regarding the computational cost and mIoU of our U-MixFormer, certain limitations need to be addressed. We tested the inference time of a single 2048 $\times$ 1024 image using a single A100 GPU under the *mmsegmentation* benchmark setup. As evident from Table 6, U-MixFormer tends to have a slower inference time than other light-weight models. The latency can be attributed to the inherent structure of the U-Net, which necessitates the preservation of information through lateral (or residual) connections. While essential for capturing hierarchical features, these connections introduce overheads during the inference phase. To address this limitation, we aim to explore model compression techniques such as pruning and knowledge distillation in our future work. These approaches are anticipated to potentially improve the inference speed while retaining the accuracy benefits of U-MixFormer.

Table 6. Comparison of inference times with light-weight models

Method	Inf. time (ms)	mIoU	GFLOPs	Params (M)
PSPNet	**26.7**	70.2	423.4	13.7
DeepLabV3+	36.0	75.2	125.5	15.4
SegFormer	44.8	76.2	125.5	**3.8**
FeedFormer	54.4	77.9	107.4	4.5
U-MixFormer	55.4	**79.0**	**101.7**	6.1
U-MixFormer	68.4	**81.0**	**90.0**	6.7

5. Conclusion

In this paper, we present U-MixFormer, built upon the U-Net structure designed for semantic segmentation. U-MixFormer starts with the most contextual encoder feature map and progressively incorporates finer details, building upon U-Net's capability to capture and propagate hierarchical features. Our mix-attention design emphasizes components of merged feature maps, aligning them with increasingly granular lateral encoder features. This ensures a harmonious fusion of high-level contextual information with intricate low-level details, which is pivotal for precise segmentation. While our model performs well on the multiple datasets, future work will focus on evaluating and enhancing its generalization capability across different data domains to ensure robustness. We demonstrate the superiority of our U-MixFormer across diverse encoders on popular benchmark datasets.

6. Acknowledgments

This work was supported by the Korea Institute for Advancement of Technology (KIAT) (No. RS-2024-00468747, Development of AI and Lightweight Technology for Embedding Multisensory Intelligence Modules).

References

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

[2] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–218, 2023. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 2

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, 2020. 1

[5] 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. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1, 5

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 2, 5

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875, 2021. 2, 5

[9] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[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. In *Advances in Neural Information Processing Systems*, pages 1140–1156, 2022. 2, 3, 5, 8

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1748–1758, 2022. 3

[13] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 603–612, 2019. 5

[14] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8825–8835, 2020. 7

[15] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. *arXiv preprint arXiv:2109.04454*, 2021. 5

[16] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 82–92, 2019. 5

[17] 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 (ICCV)*, pages 10012–10022, October 2021. 2

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 5

[19] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12607–12616, 2019. 5

[20] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging (MLMI)*, pages 267–276, 2021. 3

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 2

[22] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2263–2271, 2023. 2, 3, 4, 5

[23] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272, 2021. 2, 5

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[25] Li Wang, Dong Li, Han Liu, Jinzhang Peng, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation in autonomous driving. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2487–2494, 2022. 1

[26] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. 3

[27] 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 (ICCV)*, pages 568–578, 2021. 2, 4

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

[29] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11998–12008, 2022. 2, 3, 5

[30] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12351, pages 173–190, 2020. 5

[31] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017. 5

[32] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 2

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 5