EfficientFCN: Holistically-guided Decoding for Semantic Segmentation

Jianbo Liu[1], Junjun He[2], Jiawei Zhang[3], Jimmy S. Ren[3], and Hongsheng Li[1]

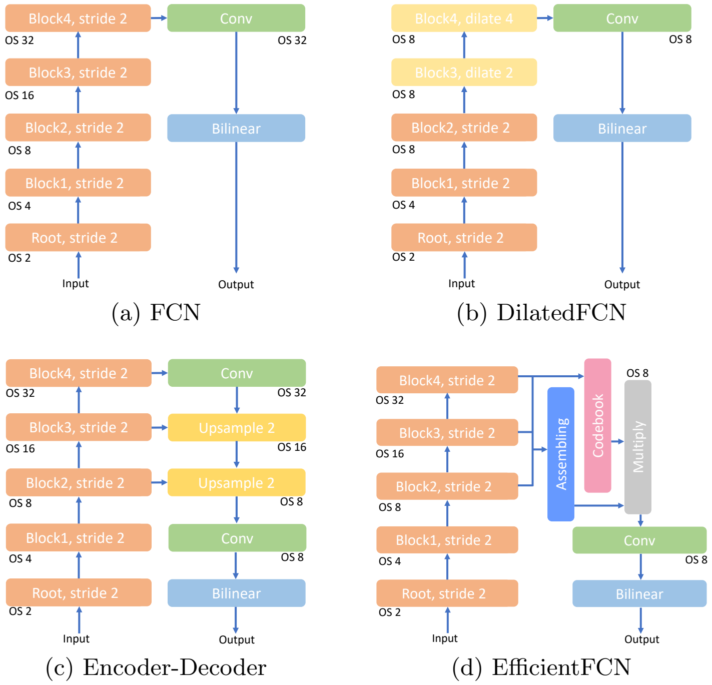
[1] CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
[2] Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] SenseTime Research
{liujianbo@link, hsli@ee}.cuhk.edu.hk

Abstract. Both performance and efficiency are important to semantic segmentation. State-of-the-art semantic segmentation algorithms are mostly based on dilated Fully Convolutional Networks (dilatedFCN), which adopt dilated convolutions in the backbone networks to extract high-resolution feature maps for achieving high-performance segmentation performance. However, due to many convolution operations are conducted on the high-resolution feature maps, such dilatedFCN-based methods result in large computational complexity and memory consumption. To balance the performance and efficiency, there also exist encoder-decoder structures that gradually recover the spatial information by combining multi-level feature maps from the encoder. However, the performances of existing encoder-decoder methods are far from comparable with the dilatedFCN-based methods. In this paper, we propose the EfficientFCN, whose backbone is a common ImageNet pretrained network without any dilated convolution. A holistically-guided decoder is introduced to obtain the high-resolution semantic-rich feature maps via the multi-scale features from the encoder. The decoding task is converted to novel codebook generation and codeword assembly task, which takes advantages of the high-level and low-level features from the encoder. Such a framework achieves comparable or even better performance than state-of-the-art methods with only 1/3 of the computational cost. Extensive experiments on PASCAL Context, PASCAL VOC, ADE20K validate the effectiveness of the proposed EfficientFCN.

Keywords: Semantic Segmentation, Encoder-decoder, Dilated Convolution, Holistic Features

1 Introduction

Semantic segmentation or scene parsing is the task of assigning one of the predefined class labels to each pixel of an input image. It is a fundamental yet challenging task in computer vision. The Fully Convolutional Network (FCN) [15], as shown in Fig. 1(a), for the first time demonstrates the success of exploiting a fully convolutional network in semantic segmentation, which adopts a DCNN as



(a) FCN

(b) DilatedFCN

(c) Encoder-Decoder

(d) EfficientFCN

Fig. 1: Different architectures for semantic segmentation. (a) the original FCN with output stride (OS)=32. (b). DilatedFCN based methods sacrifice efficiency and exploit the dilated convolution with stride 2 and 4 in the last two stages to generate high-resolution feature maps. (c)Encoder-Decoder methods employ the U-Net structure to recover the high-resolution feature maps. (d) Our proposed EfficientFCN with codebook generation and codeword assembly for high-resolution feature upsampling in semantic segmentation.

the feature encoder (*i.e.*, ResNet[9]) to extract high-level semantic feature maps and then applies a convolution layer to generate the dense prediction. For the semantic segmentation, high-resolution feature maps are critical for achieving accurate segmentation performance since they contain fine-grained structural information to delineate detailed boundaries of various foreground regions. In addition, due to the lack of large-scale training data on semantic segmentation, transferring the weights pre-trained on ImageNet can greatly improve the segmentation performance. Therefore, most state-of-the-art semantic segmentation methods adopt classification networks as the backbone to take full advantages of ImageNet pre-training. The resolution of feature maps in the original classification model is reduced with consecutive pooling and strided convolution operations to learn high-level feature representations. The output stride of the final feature map is 32 (OS=32), where the fine-grained structural information is discarded. Such low-resolution feature maps cannot fully meet the requirements of semantic segmentation where detailed spatial information is needed. To tackle this problem, many works exploit dilated convolution (or atrous convolution) to enlarge the receptive field (RF) while maintaining the resolution of high-level feature maps. State-of-the-art dilatedFCN based methods[24,2,25,8,26] (shown in Fig. 1(b)) have demonstrated that removing the downsampling operation and replacing convolution with the dilated convolution in the later blocks can achieve superior performance, resulting in final feature maps of output stride 8 (OS=8). Despite the superior performance and no extra parameters introduced by dilated convolution, the high-resolution feature representations require high computational complexity and memory consumption. For instance, for an input image with 512×512 and the ResNet101 as the backbone encoder, the computational complexity of the encoder increases from 44.6 GFlops to 223.6 GFlops when adopting the dilated convolution with the strides 2 and 4 into the last two blocks.

Alternatively, as shown in Fig. 1(c), the encoder-decoder based methods (*e.g.* [18]) exploit using a decoder to gradually upsample and generate the high-resolution feature maps by aggregating multi-level feature representations from the backbone (or the encoder). These encoder-decoder based methods can obtain high-resolution feature representations efficiently. However, on one hand, the fine-grained structural details are already lost in the topmost high-level feature maps of OS=32. Even with the skip connections, lower-level high-resolution feature maps cannot provide abstractive enough features for achieving high-performance segmentation. On the other hand, existing decoders mainly utilize the bilinear upsampling or deconvolution operations to increase the resolution of the high-level feature maps. These operations are conducted in a local manner. The feature vector at each location of the upsampled feature maps is recovered from a limited receptive filed. Thus, although the encoder-decoder models are generally faster and more memory friendly than dilatedFCN based methods, their performances generally cannot compete with those of the dilatedFCN models.

To tackle the challenges in both types of models, we propose the EfficienFCN (as shown in Fig. 1(d)) with the Holistically-guided Decoder (HGD) to bridge the gap between the dilatedFCN based methods and the encoder-decoder based methods. Our network can adopt any widely used classification model without dilated convolution as the encoder (such as ResNet models) to generate low-resolution high-level feature maps (OS=8). Such an encoder is both computationally and memory efficient than those in DilatedFCN model. Given the multi-level feature maps from the last three blocks of the encoder, the proposed holistically-guided decoder takes the advantages of both high-level but low-resolution (OS=32) and also mid-level high-resolution feature maps (OS=8, OS=16) for achieving high-level feature upsampling with semantic-rich features. Intuitively, the higher-resolution feature maps contain more fine-grained structural information, which is beneficial for spatially guiding the feature upsampling process; the lower-resolution feature maps contain more high-level semantic information, which are more suitable to encode the global context effectively. Our HGD therefore generates a series of holistic codewords in a codebook to summarize different global and high-level aspects of the input image from the low-resolution feature maps (OS=32). Those codewords can be properly assembled in a high-resolution grid to form the upsampled feature maps with rich semantic information. Following this principle, the HGD generates assembly coefficients from the mid-level high-resolution feature maps (OS=8, OS=16) to guide the linear assembly of the holistic codewords at each high-resolution spatial location to achieve feature upsampling. Our proposed EfficientFCN with holistically-guided decoder achieves high segmentation accuracy on three popular public benchmarks, which demonstrate the efficiency and effectiveness of our proposed decoder.

In summary, our contributions are as follows.

- We propose a novel holistically-guided decoder, which can efficiently generate the high-resolution feature maps considering holistic contexts of the input image.
- Because of the light weight and high performance of the proposed holistically-guided decoder, our EfficientFCN can adopt the encoder without any dilated convolution but still achieve superior performance.
- Our EfficientFCN achieves competitive (or better) results compared with the state-of-the-art dilatedFCN based methods on the PASCAL Context, PASCAL VOC, ADE20K datasets, with 1/3 fewer FLOPS.

2 Related Work

In this section, we review recent FCN-based methods for semantic segmentation. Since the successful demonstration of FCN [15] on semantic segmentation, many methods were proposed to improve the performance the FCN-based methods, which mainly include two categories of methods: the dilatedFCN-based methods and the encoder-decoder architectures.

DilatedFCN. The Deeplab V2 [2,3] proposed to exploit dilated convolution in the backbone to learn a high-resolution feature map, which increases the output stride from 32 to 8. However, the dilated convolution in the last two layers of the backbone adds huge extra computation and leaves large memory footprint. Based on the dilated convolution backbone, many works [26,5,6,7] continued to apply different strategies as the segmentation heads to acquire the context-enhanced feature maps. PSPNet [28] utilized the Spatial Pyramid Pooling (SPP) module to increase the receptive field. EncNet [25] proposed an encoding layer to predict a feature re-weighting vector from the global context and selectively high-lights class-dependent feature maps. CFNet [26] exploited an aggregated co-occurrent feature (ACF) module to aggregate the co-occurrent context by the pair-wise similarities in the feature space. Gated-SCNN[20] proposed to use a new gating mechanism to connect the intermediate layers and a new loss function that exploits the duality between the tasks of semantic segmentation and semantic boundary prediction. DANet [5] proposed to use two attention modules with the self-attention mechanism to aggregate features from spatial and channel dimensions respectively. ACNet [6] applied a dilated ResNet as the backbone and combined the encoder-decoder strategy for the observation that the global context from high-level features helps the categorization of some large semantic confused regions, while the local context from lower-level visual features helps to generate sharp boundaries or clear details. DMNet [7] generated a set of dynamic filters of different sizes from the multi-scale neighborhoods for handling the scale variations of objects for semantic segmentation. Although these works further improve the performances on different benchmarks, these proposed heads still adds extra computational costs to the already burdensome encoder.

Encoder-Decoder. Another type of methods focus on efficiently acquire the high-resolution semantic feature maps via the encoder-decoder architectures. Through the upsampling operations and the skip connections, the encoder-decoder architecture [18] can gradually recover the high-resolution feature maps for segmentation. DUsampling [21] designed a data-dependent upsampling module based on fully connected layers for constructing the high-resolution feature maps from the low-resolution feature maps. FastFCN [22] proposed a Joint Pyramid Upsampling (JPU) method via multiple dilated convolution to generate the high-resolution feature maps. One common drawback of these methods is that the feature at each location of the upsampled high-resolution feature maps is constructed via only local feature fusion. Such a property limited limits their performance in semantic segmentation, where global context is important for the final performance.

3 Proposed Method

In this section, We firstly give a thorough analysis of the classical encoder-decoder based methods. Then, to tackle the challenges in the classical encoder-decoder methods, we propose the EfficientFCN, which is based on the traditional ResNet as the encoder backbone network for semantic segmentation. In our EfficientFCN, the Holistically-guided Decoder (HGD) is designed to recover the high-resolution (OS=8) feature maps from three feature maps in the last three blocks from the ResNet encoder backbone.

3.1 Overview

In state-of-the-art DCNN backbones, the high-resolution mid-level feature maps at earlier stages can better encode fine-grained structures, while the low-resolution high-level features at the later stages are generally more discriminative for category prediction. They are essential and complementary information sources for achieving accurate semantic segmentation. To combine the strengths of both features, encoder-decoder structures were developed to reconstruct high-resolution feature maps that have semantic-rich information from the multi-scale feature maps. To generate the final feature map $\tilde{f}_8$ of OS=8 for mask prediction, a conventional three-stage encoder-decoder first upsamples the deepest encoder feature map f_{32} of OS=32 to generate OS=16 feature maps f_{16}. The OS=16 feature maps e_{16} of the same size from the encoder is either directly concatenated as $[f_{16}; e_{16}]$ (U-Net [18]) or summed as $f_{16} + e_{16}$ followed by some 1×1 convolution

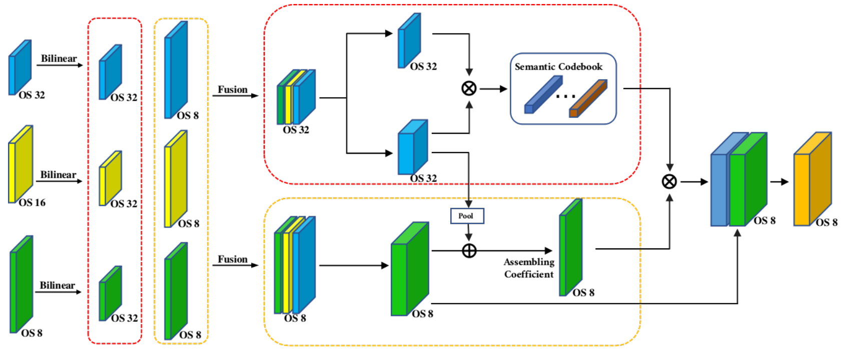


Fig. 2: Illustration of the proposed EfficientFCN model. It consists of three main components. Multi-scale feature fusion fuses multi-scale features to obtain OS=8 and OS=32 multi-scale feature maps. Holistic codebook generation results in a series of holistic codewords summarizing different aspects of the global context. High-resolution feature upsampling can be achieved by codeword assembly.

to generate the upsampled OS=16 feature maps, $\tilde{f}_{16}$. The same upsampling + skip connection procedure repeats again for $\tilde{f}_{16}$ to generate $\tilde{f}_8$. The upsampled features $\tilde{f}_8$ contain both mid-level and high-level information to some extent and can be used to generate the segmentation masks. However, since bilinear upsampling and deconvolution layer in the classical decoder are local operations with limited receptive field. We argue that they are incapable of exploring important global context of the input image, which are crucial for achieving accurate segmentation. Although there were some existing attempts [25,10] on using global context to re-weight the contributions of different channels of the feature maps either in the backbone [25] or in the upsampled feature maps [10]. This strategy only scales each feature channel but maintains the original spatial size and structures. Therefore, it is incapable of generating high-resolution semantic-rich feature maps to improve the recovery of fine-grained structures. To solve this drawback, we propose a novel Holistically-guided Decoder (HGD), which decomposes the feature upsampling task into the generation of a series of holistic codewords from high-level feature maps to capture global contexts, and linearly assembling codewords at each spatial location for semantic-rich feature upsampling. Such a decoder can exploit the global contextual information to effectively guide the feature upsampling process and is able to recover fine-grained details. Based on the proposed HGD, we present the EfficientFCN (as shown in Fig. 1(d)) with an efficient encoder free of dilated convolution for efficient semantic segmentation.

3.2 Holistically-guided Decoder for Semantic-rich Feature Upsampling

To take advantages of both low-resolution high-level feature maps of size OS=32 and the high-resolution mid-level feature maps of sizes OS=8 and OS=16, since the high-level feature maps have already lost most structural details but are semantic-rich to encode categorical information, we argue that recovering detailed structures from them is quite challenging and also non-necessary. Instead, we propose to generate a series of holistic codewords without any spatial order from the high-level feature maps to capture different aspects of the global context. On the other hand, the mid-level high-resolution feature maps have maintained abundant image structural information. But they are from relatively shallower layers and cannot encode accurate enough categorical features for final mask prediction. However, they would still be representative enough for guiding the linear assembly of the semantic-rich codewords for high-resolution feature upsampling. Our proposed Holistically-guided Decoder therefore contains three main components: multi-scale feature fusion, holistic codebook generation, and codeword assembly for high-resolution feature upsampling.

Multi-scale features fusion. Given the multi-scale feature maps from the encoder, although we can directly encode the holistic codewords from the high-level OS=32 feature maps and also directly generate the codeword combination coefficients from the mid-level OS=16 and OS=8 feature maps, we observe the fusion of multi-scale feature maps generally result in better performance. For the OS=8, OS=16, OS=32 feature maps from the encoder, we first adopt separate 1×1 convolutions to compress each of their channels to 512 for reducing the follow-up computational complexity, obtaining e_8, e_{16}, e_{32}, respectively. The multi-scale fused OS=32 feature maps m_{32} are then obtained by downsampling e_8, e_{16} to the size of OS=32 and concatenating them along the channel dimension with e_{32} as $m_{32} = [e_8^{\downarrow}; e_{16}^{\downarrow}; e_{32}] \in \mathbb{R}^{1536 \times (H/32) \times (W/32)}$, where $^{\downarrow}$ represents bilinear downsampling, $[\cdot;\cdot]$ denotes concatenation along the channel dimension, and H and W are the input image's height and width, respectively. We can also obtain the multi-scale fused OS=8 feature maps, $m_8 = [e_8; e_{16}^{\uparrow}; e_{32}^{\uparrow}] \in \mathbb{R}^{1536 \times (H/8) \times (W/8)}$, in a similar manner.

Holistic codebook generation. Although the mutli-scale fused feature maps m_{32} are created to integrate both high-level and mid-level features, their small resolutions make them lose many structural details of the scene. On the other hand, because e_{32} is encoded from the deepest layer, m_{32} is able to encode rich categorical representations of the image. We therefore propose to generate a series of unordered holistic codewords from m_{32} to implicitly model different aspects of the global context. To generate n holistic codewords, a codeword base map $B \in \mathbb{R}^{1024 \times (H/32) \times \ (W/32)}$ and n spatial weighting maps $A \in \mathbb{R}^{n \times (H/32) \times (W/32)}$ are first computed from the fused multi-scale feature maps m_{32} by two separate 1×1 convolutions. For the bases map B, we denote $B(x, y) \in \mathbb{R}^{1024}$ as the 1024-d feature vector at location (x, y); for the spatial weighting maps A, we use $A_i \in \mathbb{R}^{(H/32) \times (W/32)}$ to denote the ith weighting map. To ensure the weighting maps A are properly normalized, the softmax function is adopted to normalize all spatial locations of each channel i (the i-th spatial feature map) as

$$\tilde{A}_i(x, y) = \frac{\exp(A_i(x, y))}{\sum_{p,q} \exp(A_i(p, q))}. \tag{1}$$

The i-th codeword $c_i \in \mathbb{R}^{1024}$ can be obtained as the weighted average of all codeword bases $B(x, y)$, *i.e.*,

$$c_i = \sum_{p,q} \tilde{A}_i(p, q) B(p, q). \tag{2}$$

In other words, each spatial weighting map $\tilde{A}_i$ learns to linearly combine all codeword bases $B(x, y)$ from all spatial locations to form a single codeword, which captures certain aspect of the global context. The n weighting maps eventually result in n holistic codewords $C = [c_1, \cdots, c_n] \in \mathbb{R}^{1024 \times n}$ to encode high-level global features.

Codeword assembly for high-resolution feature upsampling. The holistic codewords can capture various global contexts of the input image. They are perfect ingredients for reconstructing the high-resolution semantic-rich feature maps as they are encoded from the high-level features m_{32}. However, since their structural information have been mostly removed during codeword encoding, we turn to use the OS=8 multi-scale fused features m_8 to predict the linear assembly coefficients of the n codewords at each spatial location for creating a high-resolution feature map. More specifically, we first create a raw codeword assembly guidance feature map $G \in \mathbb{R}^{1024 \times (H/8) \times (W/8)}$ to predict the assembly coefficients at each spatial location, which are obtained by applying a 1×1 convolution on the multi-scale fused features m_8. However, the OS=8 fused features m_8 have no information on the holistic codewords as they are all generated from m_{32}. We therefore consider the general codeword information as the global average vector of the codeword based map $\bar{B} \in \mathbb{R}^{1024}$ and location-wisely add it to the raw assembly guidance feature map to obtain the novel guidance feature map $\bar{G} = G \oplus \bar{B}$, where $\oplus$ represents location-wise addition. Another 1×1 convolution applied on the guidance feature map $\bar{G}$ generates the linear assembly weights of the n codewords $W \in \mathbb{R}^{n \times (H/8) \times (W/8)}$ for all $(H/8) \times (W/8)$ spatial locations. By reshaping the weighting map W as an $n \times (HW/8^2)$ matrix, the holistically-guided upsampled feature $\tilde{f}_8$ can be easily obtained as

$$\tilde{f}_8 = W^\top C. \tag{3}$$

Given the holistically-guided upsampled feature map $\tilde{f}_8$, we reconstruct the final upsampled feature map $\hat{f}_8$ by concatenating the feature map $\tilde{f}_8$ with the guidance feature map G. Such an upsampled feature map $\hat{f}_8$ takes advantages of both m_8 and m_{32}, and contains semantic-rich and also structure-preserved features for achieving accurate segmentation.

Final segmentation mask. Given the upsampled feature map $\hat{f}_8$, a 1×1 convolution can output a segmentation map of OS=8, which is further upsampled back to the original resolution $H \times W$ as the final segmentation mask.

4 Experiments

In this section, we introduce the implementation details, training strategies and evaluation metrics of our experiments. To evaluate our proposed EfficientFCN

Table 1: Comparisons with classical encoder-decoder methods.

Method	Backbone	OS	mIoU%	Parameters (MB)	GFlops (G)
FCN-32s	ResNet101	32	43.3	54.0	44.6
dilatedFCN-8s	dilated-ResNet101	8	47.2	54.0	223.6
UNet-Bilinear	ResNet101	8	49.3	60.7	87.9
UNet-Deconv	ResNet101	8	49.1	62.8	93.2
EfficientFCN	ResNet101	8	55.3	55.8	69.6

model, we conduct comprehensive experiments on three public datasets PASCAL Context [16], PASCAL VOC 2012 [4] and ADE20K [30]. To further evaluate the contributions of individual components in our model, we conduct detailed ablation studies on the PASCAL Context dataset.

4.1 Implementation Details

Network Structure. Different with the dilatedFCN based methods, which remove the stride of the last two blocks of the backbone networks and adopt the dilated convolution with the dilation rates 2 and 4, we use the original ResNet [9] as our encoder backbone network. Thus the size of the output feature maps from the last ResBlock is 32× smaller than that of the input image. After feeding the encoder feature maps into our proposed holistic-guided decoder, the classification is performed on the output upsampled feature map $\hat{f}_8$. The ImageNet [19] pre-trained weights are utilized to initialize the encoder network.
Training Setting. A poly learning rate policy [2] is used in our experiments. We set the initial learning rates as 0.001 for PASCAL Context [16], 0.002 for PASCAL VOC 2012 [4] and ADE20K [30]. The power of poly learning rate policy is set as 0.9. The optimizer is stochastic gradient descent (SGD) [1] with momentum 0.9 and weight decay 0.0001. We train our EfficientFCN for 120 epochs on PASCAL Context, 80 epochs on PASCAL 2012 and 120 epochs on ADE20K, respectively. We set the crop size to 512×512 on PASCAL Context and PASCAL 2012. Since the average image size is larger than other two datasets, we use 576×576 as the crop size on ADE20K. For data augmentation, we only randomly flip the input image and scale it randomly in the range $[0.5, 2.0]$.
Evaluation Metrics. We choose the standard evaluation metrics of pixel accuracy (pixAcc) and mean Intersection of Union (mIoU) as the evaluation metrics in our experiments. Following the best practice [25,8,5], we apply the strategy of averaging the network predictions in multiple scales for evaluation. For each input image, we first randomly resize the input image with a scaling factor sampled uniformly from [0.5, 2.0] and also randomly horizontally flip the image. These predictions are then averaged to generate the final prediction.

4.2 Results on PASCAL Context

The PASCAL Context dataset consists of 4,998 training images and 5,105 testing images for scene parsing. It is a complex and challenging dataset based

Table 2: Results of using different numbers of scales for multi-scale fused feature m_{32} to generate the holistic codewords.

	$\{32\}$	$\{16, 32\}$	$\{8, 16, 32\}$
pixAcc	80.0	80.1	80.3
mIoU	54.8	55.1	55.3

Table 3: Results of using different numbers of scales for multi-scale fused feature m_8 to estimate codeword assembly coefficients.

	$\{8\}$	$\{8, 16\}$	$\{8, 16, 32\}$
pixAcc	78.9	80.0	80.3
mIoU	47.9	52.1	55.3

on PASCAL VOC 2010 with more annotations and fine-grained scene classes, which includes 59 foreground classes and one background class. We take the same experimental settings and evaluation strategies following previous works [25,26,5,6,7]. We first conduct ablation studies on this dataset to demonstrate the effectiveness of each individual module design of our proposed EfficientFCN and then compare our model with state-of-the-art methods. The ablation studies are conducted with a ResNet101 encoder backbone.

Comparison with the classical encoder-decoders. For the classical encoder-decoder based methods, the feature upsampling is achieved via either bilinear interpolation or deconvolution. We implement two classical encoder-decoder based methods, which include the feature upsampling operation (bilinear upsampling or deconvolution) and the skip-connections. To verify the effectiveness of our proposed HGD, these two methods are trained and tested on the PASCAL Context dataset with the same training setting as our model. The results are shown in Table 1. Although the classical encoder-decoder methods have similar computational complexities, their performances are generally far inferior than our EfficientFCN. The key reason is that their upsampled feature maps are recovered in a local manner. The simple bilinear interpolation or deconvolution cannot effectively upsample the OS=32 feature maps even with the skip-connected OS=8 and OS=16 feature maps. In contrast, our proposed HGD can effectively upsample the high-resolution semantic-rich feature maps not only based on the fine-grained structural information in the OS=8 and OS=16 feature maps but also from the holistic semantic information from the OS=32 feature maps.

Multi-scale features fusion. We conduct two experiments on multi-scale features fusion to verify their effects on semantic codebook generation and codeword assembly for feature upsampling. In our holistically-guided decoder, the semantic codewords are generated based on the OS=32 multi-scale fused feature maps m_{32} and the codewords assembly coefficients are predicted from the OS=8 multi-scale fused feature maps m_8. For the codeword generation, we conduct three experiments to generate the semantic codebook from multi-scale fused features with different numbers of scales. As shown in Table 2, when reducing the number of fusion scales from 3 to 2 and from 2 to 1, the performances of our EfficientFCN slightly decrease. The phenomenon is reasonable as the deepest feature maps contain more categorical information than the OS=8 and OS=16 feature maps. For the codeword assembly coefficient estimation, the similar experiments are conducted, where results are shown in Table 3. However, different from the above results, when fewer scales of feature maps are used to form

Table 4: Ablation study of the number of the semantic codewords.

	32	64	128	256	512	1024
pixAcc	79.9	80.1	80.1	80.3	80.3	80.1
mIoU	54.5	54.9	55.0	55.3	55.5	55.1
GFLOPS	67.9	68.1	68.6	69.6	72.1	78.9

Table 5: Segmentation results of state-of-the-art methods on PASCAL Context and ADE20K validation dataset.

Method	**Backbone**	**mIoU%** (PASCAL Context)	**mIoU%** (ADE20K)	**GFlops**
DeepLab-v2 [2]	Dilated-ResNet101-COCO	45.7	-	>223
RefineNet [12]	Dilated-ResNet152	47.3	-	>223
MSCI [11]	Dilated-ResNet152	50.3	-	>223
PSPNet [28]	Dilated-ResNet101	-	43.29	>223
SAC [27]	Dilated-ResNet101	-	44.30	>223
EncNet [25]	Dilated-ResNet101	51.7	44.65	234
DANet [5]	Dilated-ResNet101	52.6	-	>223
APCNet [8]	Dilated-ResNet101	54.7	45.38	245
CFNet [26]	Dilated-ResNet101	54.0	44.89	>223
ACNet [6]	Dilated-ResNet101	54.1	**45.90**	>223
APNB [31]	Dilated-ResNet101	52.8	45.24	>223
DMNet [7]	Dilated-ResNet101	54.4	45.50	242
Ours	ResNet101	**55.3**	45.28	**70**

the multi-scale fused OS=8 feature map m_8, the performances of our Efficient-FCN show significant drops. These results demonstrate that although the OS=8 feature maps contain more fine-grained structural information, the semantic features from higher-level feature maps are essential for guiding the recovery of the semantic-rich high-resolution feature maps.

Number of holistic codewords. We also conduct experiments to survey the effectiveness of the number of codewords in our predicted semantic codebook for feature upsampling. As shown in Table 4, as the number of the semantic codewords increases from 32 to 512, the performance improves 1% in terms of mIoU on PASCAL Context. However, when the number of the semantic codewords further increases from 512 to 1024, the performance has a slight drop, which might be caused by the additional parameters. The larger model capacity might cause model to overfit the training data. In addition, since the assembly coefficients of the semantic codewords are predicted from the OS=8 multi-scale fused feature m_8, the increased number of the semantic codewords also leads to significantly more extra computational cost. Thus, to balance the performance and also the efficiency, we set the number of the holistic codewords as 256 for the PASCAL Context and PASCAL VOC 2012 datasets. Since PASCAL Context only has 60 classes and we observe the number of codewords needed is approximately 4 times than the number of classes. We therefore set the number of codewords as 600 for ADE20K, which has 150 classes.

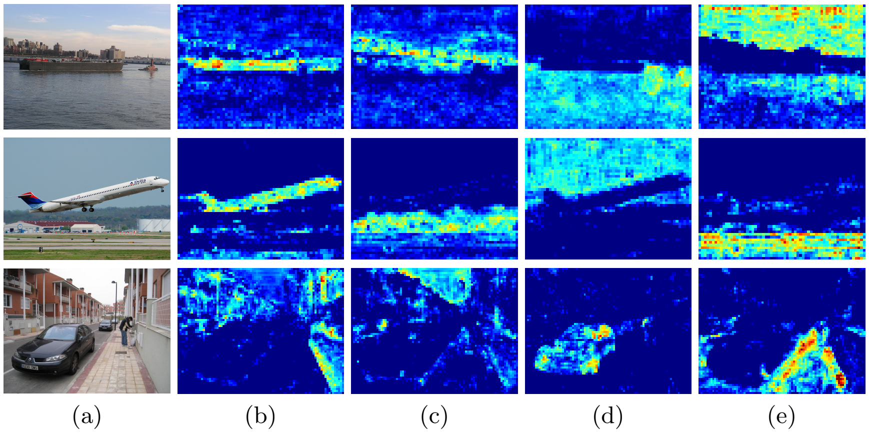


Fig. 3: (a) Input images from the PASCAL Context and ADE20K dataset. (b-e) Different weighting maps $\tilde{A}_i$ for creating the holistic codewords.

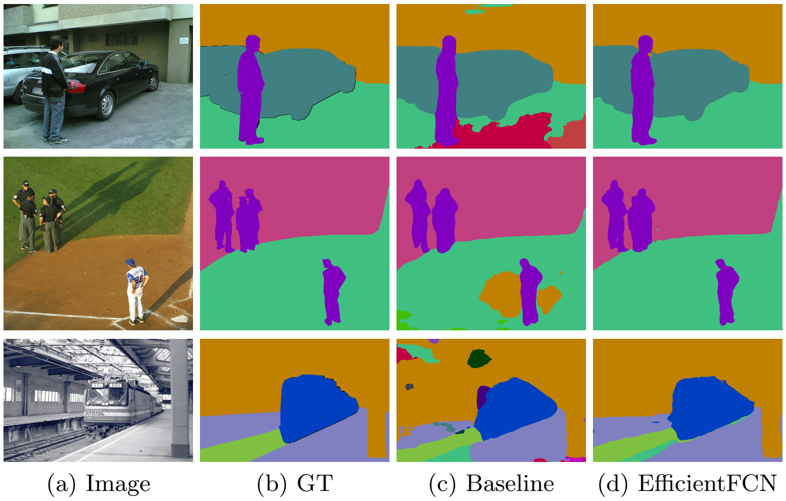


Fig. 4: Visualization results from the PASCAL Context dataset.

Importance of the codeword information transfer for accurate assembly coefficient estimation. The key of our proposed HGD is how to linearly assemble holistic codewords at each spatial location to form high-resolution upsampled feature maps based on the feature maps m_8. In our HGD, although the OS=8 features have well maintained structural image information, we argue that directly using OS=8 features to predict codeword assembly coefficients are less effective since they have no information about the codewords. We propose to transfer the codeword information as the average codeword basis, which is location-wisely added to the OS=8 feature maps. To verify this argument, we

Table 6: Results of each category on PASCAL VOC 2012 test set. Our EfficientFCN obtains 85.4 % without MS COCO dataset pre-training and 87.6% with MS COCO dataset pre-training. (For each columns, the best two entries are filled in gray color.)

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU%
FCN [15]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLabv2 [2]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [29]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [17]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN [14]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [13]	90.6	37.6	80.0	67.8	74.4	92	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
ResNet38 [23]	94.4	72.9	94.9	68.8	78.4	90.6	90.0	92.1	40.1	90.4	71.7	89.9	93.7	91.0	89.1	71.3	90.7	61.3	87.7	78.1	82.5
PSPNet [28]	91.8	71.9	94.7	71.2	75.8	95.2	89.9	95.9	39.3	90.7	71.7	90.5	94.5	88.8	89.6	72.8	89.6	64.0	85.1	76.3	82.6
EncNet [25]	94.1	69.2	**96.3**	76.7	**86.2**	96.3	90.7	94.2	38.8	90.7	73.3	90.0	92.5	88.8	87.9	68.7	92.6	59.0	86.4	73.4	82.9
APCNet [8]	95.8	75.8	84.5	76.0	80.6	96.9	90.0	96.0	**42.0**	93.7	75.4	91.6	95.0	90.5	89.3	75.8	92.8	61.9	88.9	79.6	84.2
CFNet [26]	95.7	71.9	95.0	76.3	82.8	94.8	90.0	95.9	37.1	92.6	73.0	93.4	94.6	89.6	88.4	74.9	**95.2**	63.2	**89.7**	78.2	84.2
DMNet [7]	96.1	**77.3**	94.1	72.8	78.1	**97.1**	**92.7**	96.4	39.8	91.4	75.5	92.7	**95.8**	91.0	90.3	76.6	94.1	62.1	85.5	77.6	84.4
Ours	**96.4**	74.1	92.8	75.6	81.9	96.9	92.6	**97.1**	41.6	**95.4**	72.9	**93.9**	**95.9**	90.6	**90.6**	**77.2**	94.0	67.5	89.3	**79.8**	**85.4**
With COCO Pre-training																					
CRF-RNN [29]	90.4	55.3	88.7	68.4	69.8	88.3	82.4	85.1	32.6	78.5	64.4	79.6	81.9	86.4	81.8	58.6	82.4	53.5	77.4	70.1	74.7
Piecewise [13]	94.1	40.7	84.1	67.8	75.9	93.4	84.3	88.4	42.5	86.4	64.7	85.4	89.0	85.8	86.0	67.5	90.2	63.8	80.9	73.0	78.0
DeepLabv2 [2]	92.6	60.4	91.6	63.4	76.3	95.0	88.4	92.6	32.7	88.5	67.6	89.6	92.1	87.0	87.4	63.3	88.3	60.0	86.8	74.5	79.7
RefineNet[12]	95.0	73.2	93.5	78.1	84.8	95.6	89.8	94.1	43.7	92.0	77.2	90.8	93.4	88.6	88.1	70.1	92.9	64.3	87.7	78.8	84.2
ResNet38[23]	96.2	75.2	95.4	74.4	81.7	93.7	89.9	92.5	48.2	92.0	79.9	90.1	95.5	91.8	91.2	73.0	90.5	65.4	88.7	80.6	84.9
PSPNet [28]	95.8	72.7	95.0	78.9	84.4	94.7	92.0	95.7	43.1	91.0	**80.3**	91.3	96.3	92.3	90.1	71.5	94.4	66.9	88.8	**82.0**	85.4
DeepLabv3[3]	96.4	76.6	92.7	77.8	87.6	96.7	90.2	95.4	47.5	93.4	76.3	91.4	**97.2**	91.0	**92.1**	71.3	90.9	68.9	90.8	79.3	85.7
EncNet[25]	95.3	76.9	94.2	80.2	85.2	96.5	90.8	96.3	47.9	93.9	80.0	92.4	96.6	90.5	91.5	70.8	93.6	66.5	87.7	80.8	85.9
CFNet [26]	**96.7**	79.7	94.3	78.4	83.0	**97.7**	91.6	96.7	**50.1**	95.3	79.6	93.6	**97.2**	94.2	91.7	78.4	**95.4**	**69.6**	90.0	81.4	87.2
Ours	96.6	**80.6**	**96.1**	**82.3**	**87.8**	**97.7**	**94.4**	**97.3**	47.1	**96.3**	77.9	**94.8**	**97.2**	**94.3**	91.1	**81.0**	94.3	61.5	**91.6**	**83.5**	**87.6**

design an experiment that removes the additive information transfer, and only utilizes two 1 × 1 convolutions with the same output channels on the OS=8 feature maps m_8 for directly predicting assembly coefficients. The mIoU of this implementation is 54.2%, which has a clear performance drop if there is no codeword information transfer from the codeword generation branch to the codeword coefficient prediction branch.

Visualization of the weighting maps and example results. To better interpret the obtained holistic codewords, we visualize the weighting maps $\tilde{A}$ for creating the holistic codewords in Fig. 3, where each column shows one weighting map $\tilde{A}_i$ for generating one holistic codeword. Some weighting maps focus on summarizing foreground objects or regions to create holistic codewords, while some other weighting maps pay attention to summarizing background contextual regions or objects as the holistic codewords. The visualization shows that the learned codewords implicitly capture different global contexts from the scenes. In Fig. 4, we also visualize some predictions by the baseline DilatedFCN-8s and by our EfficientFCN, where our model significantly improves the visualized results with the proposed HGD.

Comparison with state-of-the-art methods. To further demonstrate the effectiveness of our proposed EffectiveFCN with the holistically-guided decoder, the comparisons with state-of-the-art methods are shown in Table 5. The dilatedFCN based methods dominate semantic segmentation. However, our work is still able to achieve the best results compared to the dilatedFCN based methods on the PASCAL Context validation set without using any dilated convolution and has significantly less computational cost. Because of the efficient design of our HGD, our EfficientFCN only has 1/3 of the computational cost of state-of-the-arts methods but can still achieve the best performance.

4.3 Results on PASCAL VOC

The original PASCAL VOC 2012 dataset consists of 1,464 images for training, 1,449 for validation, and 1,456 for testing, which is a major benchmark dataset for semantic object segmentation. It includes 20 foreground objects classed and one background class. The augmented training set of 10,582 images, namely train-aug, is adopted as the training set following the previous experimental set in [26]. To further demonstrate the effectiveness of our proposed HGD. We adopt all the best strategies of HGD design and compare it with state-of-the-art methods on the test set of PASCAL-VOC 2012, which is evaluated on the official online server. As shown in Table 6, the dilatedFCN based methods dominate the top performances on the PSCAL VOC benchmark. However, our EfficientFCN with a backbone having no dilated convolution can still achieve the best results among all the ResNet101-based methods.

4.4 Results on ADE20K

The ADE20K dataset consists of 20K images for training, 2K images for validation, and 3K images for testing, which were used for ImageNet Scene Parsing Challenge 2016. This dataset is more complex and challenging with 150 labeled classes and more diverse scenes. As shown in Table 5, our EfficientFCN achieves the competitive performance than the dilatedFCN based methods but has only 1/3 of their computational cost.

5 Conclusions

In this paper, we propose the EfficientFCN model with the holistically-guied decoder for achieving efficient and accurate semantic segmentation. The novel decoder is able to reconstruct the high-resolution semantic-rich feature maps from multi-scale feature maps of the encoder. Because of the superior feature upsampling performance of the HGD, our EfficientFCN, with much fewer parameters and less computational cost, achieves competitive or even better performance compared with state-of-the-art dilatedFCN based methods.

Acknowledgements

This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK 14202217 / 14203118 / 14205615 / 14207814 / 14213616 / 14208417 / 14239816, in part by CUHK Direct Grant.

References

1. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010) 9
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2017) 3, 4, 9, 11, 13
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 4, 13
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 9
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019) 4, 9, 10, 11
6. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 4, 10, 11
7. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3562–3572 (2019) 4, 5, 10, 11, 13
8. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7519–7528 (2019) 3, 9, 11, 13
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2, 9
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018) 6
11. Lin, D., Ji, Y., Lischinski, D., Cohen-Or, D., Huang, H.: Multi-scale context intertwining for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 603–619 (2018) 11
12. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017) 11, 13
13. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2016) 13
14. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1377–1385 (2015) 13
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 1, 4, 13
16. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2014) 9

17. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1520–1528 (2015) 13
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 3, 5
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 9
20. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5229–5238 (2019) 4
21. Tian, Z., He, T., Shen, C., Yan, Y.: Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3126–3135 (2019) 5
22. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816 (2019) 5
23. Wu, Z., Shen, C., Hengel, A.v.d.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016) 13
24. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017) 3
25. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 3, 4, 6, 9, 10, 11, 13
26. Zhang, H., Zhang, H., Wang, C., Xie, J.: Co-occurrent features in semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3, 4, 10, 11, 13, 14
27. Zhang, R., Tang, S., Zhang, Y., Li, J., Yan, S.: Scale-adaptive convolutions for scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2031–2039 (2017) 11
28. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) 4, 11, 13
29. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529–1537 (2015) 13
30. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 9
31. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 593–602 (2019) 11