Cognitive Computation (2018) 10:62–72
https://doi.org/10.1007/s12559-017-9530-0

Very Fast Semantic Image Segmentation Using Hierarchical Dilation and Feature Refining

Qingqun Ning[1] · Jianke Zhu[1,2] · Chun Chen[1]

Received: 7 July 2017 / Accepted: 21 November 2017 / Published online: 5 December 2017

Abstract

With the rapid development of deep learning techniques, semantic image segmentation has been considerably improved recently, which is viewed as the key problem of scene understanding in computer vision. These advances are built upon the capability of complex architectures for deep neural network. In this paper, we present a novel deep neural network architecture designed for semantic image segmentation. In order to improve the segmentation accuracy, we introduce a novel hierarchical dilation block to effectively enlarge the size of receptive field and enable multi-scale processing in fully convolutional neural network. Moreover, we exploit the technique of bypass and intermediate supervision to capture the context information during upsampling and refining coarse features. We have conducted extensive experiments on several popular semantic segmentation testbeds, including Cityscapes, CamVid, Kitti, and Helen facial datasets. The experimental results demonstrate that our proposed approach runs two times faster than the state-of-the-art method. Our full system is able to obtain realtime inference performance on 1080P images using a PC with single GPU. It executes a network forwarding at 200fps in our experiment while retaining high accuracy. Our proposed approach not only runs faster than the existing realtime methods but also performs on par with them.

Keywords Semantic image segmentation · Real-time system · Convolution neural network · Receptive field · Coarse-to-fine

Introduction

Cognitive computing is a research area that helps us to construct cognitive system based upon human cognitive activities [23]. It is expected to solve many outstanding problems in artificial intelligence and computer vision through incorporating and integrating principles from neurobiology, statistics, theoretical computer science and artificial intelligence [10]. Among various difficulties, image understanding is one of the most ordinary problem, which is very closely related to human cognitive activities. It has a series of tasks, including object detection and tracking, action recognition, face recognition, emotion analysis, and scene understanding.

Over the past few decades, cognitive science has shown its effectiveness on scene understanding and image processing. For instance, to classify outdoor scene, Zhao et al. [41] combine biologically inspired features and cortex-like memory patterns. Their cognitive model achieves state-of-the-art performance and significantly reduces the training costs. Inspired by the human visual system, Wang et al. [33] propose a coarse-to-fine pedestrian detection algorithm to actively track pedestrians in real-time.

In this paper, we focus on the problem of semantic image segmentation, which is the basis of scene understanding. The key idea of semantic segmentation is to label each pixel on image and assign it to one known category. It is a cognitive vision-based task that could be solved based on the knowledge on cognitive science. For example, Xie et al. [35] find that cognitive processing at multiple

✉ Jianke Zhu
jkzhu@zju.edu.cn

Qingqun Ning
ningqingqun@zju.edu.cn

Chun Chen
chenc@zju.edu.cn

1 College of Computer Science, Zhejiang University, Zhejiang, China

2 Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Zhejiang, China

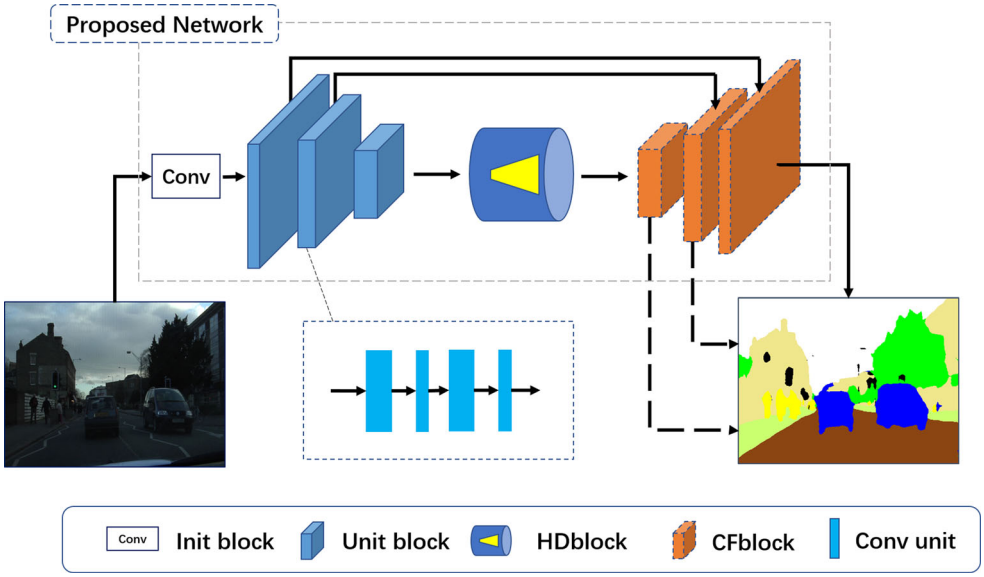

Fig. 1 Overview of our proposed approach. Given an input image, it is first processed by an initial block which contains two convolution units. A convolution unit is a convolution layer followed by an activation layer. It may also contain a batchnorm layer. Then, the output is forwarded to a unit block, which contains four convolution units. The feature map is applied to a hierarchical dilation block (HDblock), and then to a coarse-to-fine block (CFblock). See "Hierarchical Dilation and Coarse-to-Fine Block" for the details. We finally output a label map

scales with contextual information aids perceptual inference tasks. Therefore, they employ multi-scale features and contextual information to solve the problem of semantic image segmentation, in which a multiple adjacency tree model is presented to capture several kinds of regional context. Thus, it can perform exact inference with some simple assumptions. Differently from this approach, we make use of a convolutional neural network (CNN) to tackle this problem. CNN has been proven to be an effective approach to image understanding [11, 16, 29, 34], especially for semantic image segmentation [8, 27]. However, an existing drawnback for CNN-based methods is the feature coarse-to-fine problem. This is mainly due to the successive pooling and subsampling layers that result in feature maps with significantly reduced spatial resolution. Although interpolation [3] and deconvolution layer [27] offer solutions to upsample feature maps, they fail to refine the features simultaneously.

Inspired by multi-scale cognitive mechanisms, we propose to aggregate multiple-scale contextual information upon CNN for semantic image segmentation (Fig. 1). In contrast to previous methods that exploit dilation convolution for multi-scale reasoning in parallel structure or sequential structure, we propose a novel hierarchical dilation block. It not only helps to reduce the depth of CNN, but also increases the variety on fields of view. Thus, our proposed method enables to process image on objects and context at multiple scales. To deal with the problem of coarse-to-fine feature, we introduce a fused block that combines skip connection and intermediate supervision. Therefore, our proposed coarse-to-fine block is capable of acquiring finer feature maps while increasing the spatial resolution. More importantly, our approach is very efficient, which is able to achieve real-time performance on images with the full HD resolution.

Related Work

Semantic image segmentation has been intensively studied for many years. Early methods [28, 35] mainly rely on hand-crafted features in association with traditional machine learning algorithms. These approaches are well-known to be compromised by the limited expressive power of the features.

During past few years, deep learning techniques have shown excellent performance in computer vision. Fully convolutional network (FCN) [27] is the pioneering work that firstly introduces a powerful CNN for the task of semantic segmentation. They replace the conventional fully connected layer with a convolutional one, such that the network output is a spatial map rather than the classification score. However, FCN suffers from some weakness limiting its capability, such as it fails to refine feature and cannot capture the context information effectively.

Inspired from FCN, many research works have been introduced to overcome its drawnback for semantic segmentation. In [3, 36], dilated convolution is proposed to enlarge the receptive field of the network. Noh et al. [21] propose an encoder-decoder structures to deliver spatial information from low layers to high layers. To integrate context information into models, DeepLab models [3] apply Conditional Random Field (CRF) as a pose-processing stage. Zheng et al. [42] fully integrate the CRF

with a FCN and train the whole network in an end-to-end manner. Yu et al. [36] propose a multi-scale context aggregation module. PSPNet [40] exploits the capability of global context information by different-region-based context aggregation through a pyramid pooling module. Some works [7, 26] propose to make use of multi-scale predictions to deal with context knowledge integration. The detailed information can be found in a recent survey [8].

The high accuracy of these methods are all on account of a CNN model with heavy computational cost, which have been pre-trained on ImageNet dataset [6]. Toward fast or even real-time processing, the small network is introduced. SqueezeNet [13] is a low-latency network, which retains accuracy as the well-known AlexNet [16] for image recognition. YOLO [25] is another efficient network architecture for realtime object detection. Additionally, Paszke et al. [22] present an efficient neural network architecture named as ENet. ENet is especially designed for semantic image segmentation, which is built upon various bottleneck blocks.

On the other hand, some seminal works [12, 18, 19, 24, 43, 44] attempt to restrict CNNs into low-precision version by binarizing or quantizing network weights, pruning filters, and enabling sparse weights. Hubara et al. [12] introduce binary weights and activations for neural networks. This will replace most of arithmetic operations with bit-wise ones, which substantially improves power-efficiency. XNOR-NET [24] is another network, in which both the filters and input are binarized. Liu et al .[19] obtain significant speedup by proposing a method to zero out more than 90% of parameters. In contrast to pruning weights, Li et al. [18] propose to directly prune filters for acceleration. Zhou et al. [44] present a method to convert any pre-trained full-precision CNN model into a low-precision version. Recently, Zeng et al. [38] address this by combining the technique of pruning and quantization. All these schemes claim to have less performance drop along with impressing speedup.

Very Fast Semantic Image Segmentation

In this section, we give the details of our network. Firstly, we introduce our proposed hierarchical dilation block to enlarge receptive field. Secondly, we present coarse-to-fine block to deal with the issue of refining features. Finally, an efficient convolutional neural network architecture is proposed to facilitate the real-time performance.

Hierarchical Dilation

To achieve good performance for deep convolution neural network (CNN), increasing the receptive field size of a network is known as an effective technique. Specifically, pooling or subsampling is a universal strategy to increase the size of receptive field. However, excessive subsampling will result in large loss on spatial information for CNN features, which is very important for semantic segmentation. The other scheme is to increase the kernel size of convolution layers, this will directly increase the computational cost significantly [29] and collide with our objective on building an efficient network for semantic segmentation.

To tackle with the above issue, dilated/atrous convolution [3, 36] is a remedy. Dilated convolution is a normal convolution that applies convolution filters with a hole. It is a simple yet effective strategy to enlarge the size of receptive field. There are various mechanisms to make use of the dilated convolution. Traditional structures employ sequential layers with equal or incremental dilation factors or parallel layers with various dilation factors. Figure 2 illustrates different schemes. These approaches successfully increase the network receptive field with limited capability. For example, they are unable to capture the scale variations for objects in images.

To this end, we introduce a novel hierarchical dilation block, named HDblock. The proposed HDblock contains multilevel parallel dilated convolutions and each convolution includes 3×3 convolution kernels with various dilation

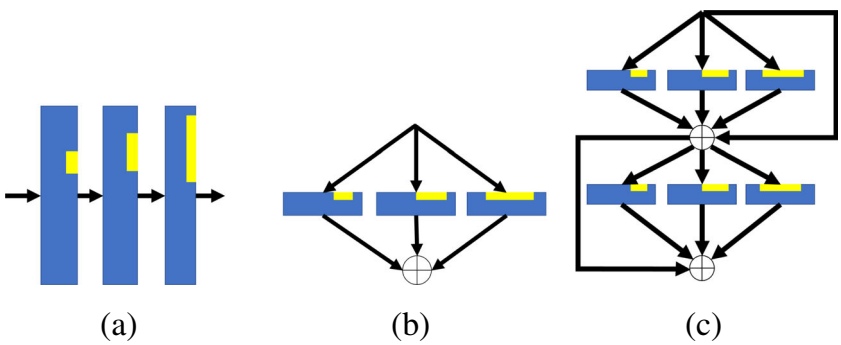

Fig. 2 Different dilation structures to increase the size of receptive field. The inside rectangle indicates the dilation factor of that layer. Panels **a** and **b** are two conventional schemes. Panel **c** is our proposed HDblock that is a hierarchical structure

Table 1 The comparison between a sequential structure of dilated convolution [36] and an example of our HDblock

Methods	Sequential [36]	HDblock
	3×3, 1	3×3, 2
	3×3, 1	3×3, 8
Conv Layers	3×3, 2	3×3, 16
(kernal size,	3×3, 4	3×3, 2
dilated factor)	3×3, 8	3×3, 8
	3×3, 16	3×3, 16
Layer amount	6	2[a]
FoV variety	6	16
Maximal FoV	65×65	65×65

[a]Strictly speaking, this HDblock contains two levels and each level includes three parallel conv layers followed by an element-sum layer

factors. Our HDblock is not a straightforward repeated parallel structure. Importantly, it contains a bypass connection. This hierarchical structure enables us to capture large field-of-view (FoV) in diverser sizes. Suppose that the structure has n levels and each level contains m dilation factors. It is easy to find that simple sequential and parallel structure process n and m kinds of FoV, respectively. However, the size of our proposed HDblock reaches as many as $(m+1)^n$. It is a remarkable large quantity that enables us authentically to capture multi-scale information.

One advantage of our proposed HDblock is that we effectively enlarge the receptive field with less gains in the depth of deep neural networks. For example, the context network architecture introduced in [36] needs six layers to obtain a 65×65 receptive field, while our HDblock achieves that by using a two-level structure with various dilation factors. The detail comparison is shown at Table 1. This is significant as the ability to propagate gradients on deep network is still a concern [32]. The other advantage is that HDblock enables a great variety of FoV for the network such that multi-scale processing is straightly feasible. This not only offers context assimilation on large FoV, but also enables accurate object localization. For example, a small FoV is more appropriate to capture the feature of an eye, while a building needs a large FoV.

Coarse-to-Fine Block

Pooling with downsampling is indispensable part of CNNs. It is essential to reducing the probability of over-fitting and heavy computational cost. However, it will lead to coarse output of deep neural network, which often requires an upsampling process for the task of pixel-wise labeling. Many kinds of upsampling methods have already been proposed. Interpolation layer [3] directly applies bilinear interpolation on the feature maps. Deconvolution layer [21, 27] is another means to obtain upsampling result. It is learnable like normal convolution layer but with fractional stride. Unpooling layer [1, 37] recovers fine prediction by exploiting the recorded locations of the maxima within each pooling region. All these methods are only simply upsampling operation.

Instead of using a straightforward upsampling layer, in this paper, we propose an integrated coarse-to-fine block called CFblock, which aims at upsampling and refining features at the same time. The structure of our proposed CFblock is illustrated in Fig. 3. Firstly, an input feature is processed by a single layer discussed above. Thus, the coarse features are directly enlarged, which are usually being doubled. Practically, we pick a deconvolution layer as upsampling operation without the specific concern. Then,

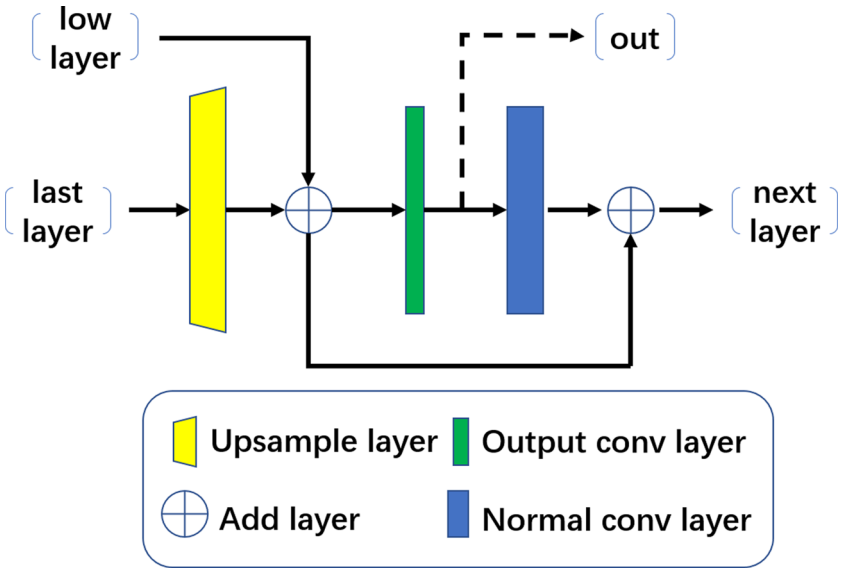

Fig. 3 The structure of our proposed CFblock. The block firstly upsamples the input feature map and merges it with the feature map from lower layer. Then we apply the output to a convolution layer to generate a prediction. This prediction is integrated back and forwarded to the next layer

Table 2 Model details and run-time performance on NVIDIA 1080TI

Model	Parameters	Model size	640 × 360		1280 × 720		1920 × 1080	
			ms	fps	ms	fps	ms	fps
ENet [22]	*0.37*M	*1.4*MB	13.5	74.1	38.7	25.8	80.6	12.4
Baseline	0.81M	3.8MB	*3.4*	*294.1*	*12.1*	*82.6*	*26.4*	*37.9*
Ours	1.67M	6.7MB	4.6	217.4	16.2	61.7	33.9	29.5

The italic entries indicate the best speed performance

we apply two strategies to refine this enlarged feature map. One is a bypass structure. A low layer feature with the same resolution is utilized. To reduce the computational cost as large as possible, we add it to the enlarged coarse feature rather than concating them. Another one is intermediate supervision. The fused output is then processed by a 1×1 convolution layer so as to produce the prediction on score map, which is then forwarded to an intermediate loss layer. It is further employed to supervise the refining process.

The auxiliary loss layer is proven to be beneficial especially for super-deep network [32, 40]. We confirm this point in our empirical study. During testing phase, the auxiliary supervised branches are usually abandoned, as in [40]. We contrarily retain these intermediate predictions and reintegrate them back to the main branch. This enables us to have extra chances to reevaluate the refining process and rectify the generated prediction. The similar strategy is also employed in [20]. Note that they apply this idea to supervise an hourglass block for human pose estimation, while we make use of them inside a block to refine the supervised feature for semantic image segmentation. Finally, the bypass structure is employed again, where the initial refined feature is fused into intermediate prediction by a skip connection. We will show that our proposed CFblock successfully upsample and refine the feature map in the experiment.

Network Architecture

To achieve efficient semantic image segmentation, it is required to trade off between accuracy and speed. One can start from an architecture with very high accuracy, and then strive to speed it up via a variety of mechanisms. Alternatively, a lightweight network architecture can be employed and optimized to boost accuracy. It has the potential advantage that the speed-up techniques can also be applied for further acceleration. In our approach, we choose the second strategy.

The backbone of our network is based on a lightweight architecture called darknet [25], which is originally employed for object detection. They provide several different architectures that have diverse accuracy and speed. In our experiment, we directly use the tiny version.[1]

[1] https://pjreddie.com/darknet/tiny-darknet/

With the proposed HDblock and CFblock, we facilitate our network to achieve real-time performance on semantic image segmentation. To show the efficacy of our proposed approach, we treat the backbone neural network without our presented HDblock and CFblock as the baseline method. We will demonstrate in the experiment that the efficiency of our method is attribute to build HDblock and CFblock upon the lightweight backbone network.

Experiments

In this section, we evaluate our proposed method on four different datasets, including three urban scene understanding datasets Cityscapes [5], CamVid [2], and Kitti [9], and a face parsing dataset Helen [30]. Before presenting the benchmark results, we first provide the details on our implementation and run-time performance evaluation.

Experimental Settings

Implementation

The implementation of our proposed method is based on the deep learning platform Torch7 [4]. Our network is built upon the tiny darknet which is pre-trained on ImageNet [6]. In our experiment, we directly remove the last three layers, since they are designed for classified task. Then our proposed HDblock with two levels and CFblock with two auxiliary losses are appended to the backbone network.

To train a neural network model for semantic segmentation, we employ Adam optimization algorithm [15] and a class weighing scheme to deal with the imbalance class distribution as ENet [22]. The training process converges very quickly, and we train at most 150 epochs for all the datasets. Our initial learning rate is set to 0.001 with a weight decay of 0.0002. Due to the limited GPU memory, we choose different batch sizes for each dataset. Specifically, they are 4, 8, 12, 16 for Kitti, CamVid, Helen, and Cityscapes, respectively.

To make fair comparison, it should be highlighted that we do not make use of any data augmentation techniques,

Table 3 Performance of ENet, baseline, and our proposed approach on Cityscapes val set with resolution 256 × 512

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	Acc.	IoU
ENet [22]	95.8	84.4	85.6	*48.3*	52.9	50.9	*51.1*	*60.3*	91.3	*75.7*	*97.1*	*83.7*	26.9	93.0	*64.8*	*68.5*	29.5	15.3	74.2	65.4	50.2
Baseline	*97.7*	73.7	*91.0*	28.3	35.9	39.3	36.2	53.7	91.0	51.2	94.1	73.4	28.2	92.6	26.3	53.2	16.2	*31.5*	61.4	56.6	46.4
Baseline +HDblock	96.8	85.0	90.7	48.3	37.6	18.8	40.4	39.2	92.5	64.1	94.0	63.4	51.9	91.2	57.6	67.2	39.9	28.9	80.6	62.5	50.6
Baseline +CFblock	96.9	84.9	88.5	39.6	37.2	42.0	40.8	59.3	*95.0*	63.9	96.9	80.8	37.4	92.0	31.8	55.5	41.3	22.5	70.8	61.9	50.5
Ours	97.4	*86.5*	90.5	41.2	42.1	*52.4*	45.6	*60.3*	93.5	71.8	96.2	75.2	*51.9*	*94.6*	40.0	61.6	*49.7*	28.6	*80.6*	*65.8*	*54.5*

The italic values show the results with the highest accuracy

such as random mirroring, resizing and rotating in all our experiments. Also, we do not adopt any post-processing method. All these techniques are expected to further boost the experimental results.

Comparison Methods

We exploit tiny darknet as our baseline method. After removing the last three layers, we append two deconvolution layers to upsample the output score. The stride of each deconvolution is 4. Except this, all the setting for training the baseline method is identical to our proposed approach. Moreover, we compare with ENet in our experiment. The results are obtained with default setting of their original implementation. For the batch size, we pick 4, 10, 10, 10 for Kitti, CamVid, Helen, and Cityscapes, respectively.

Evaluation Metrics

We employ two different metrics to evaluate the quality of semantic segmentation, the mean accuracy (Acc.) over all classes and the mean of class-wise intersection over union (IoU) score. Assume that P_i is the set of pixel predicted as the i-th class, and T_i is the set of pixel belonging to the ith class. Then, we know that $I_i = P_i \cap T_i$ is the set of pixel correctly predicted for the ith class. Let n be the number of class, we can compute the two metrics as below:

$$Acc. = \frac{1}{n}\sum_{i}^{n}\frac{|I_i|}{|T_i|},$$
$$IoU = \frac{1}{n}\sum_{i}^{n}\frac{|I_i|}{|T_i \cup P_i|} \quad (1)$$

Run-Time Performance

We first evaluate the inference time of our model with ENet. To the best of our knowledge, ENet is the fastest neural network architecture designed for semantic segmentation currently. All the running time is obtained on a single NVIDIA 1080Ti GPU using CUDA 8.0 with cuDNN 5.0. Instead of using Torch7, we exploit the deep learning platform Caffe [14] to measure the run-time for fair comparison, since all the methods are implemented by C++. The model structure is identical to the one evaluated on Torch7, except for batchnorm layers which could be merged into convolution layers in front of them as described in the implementation.[2]

[2] https://github.com/TimoSaemann/ENet

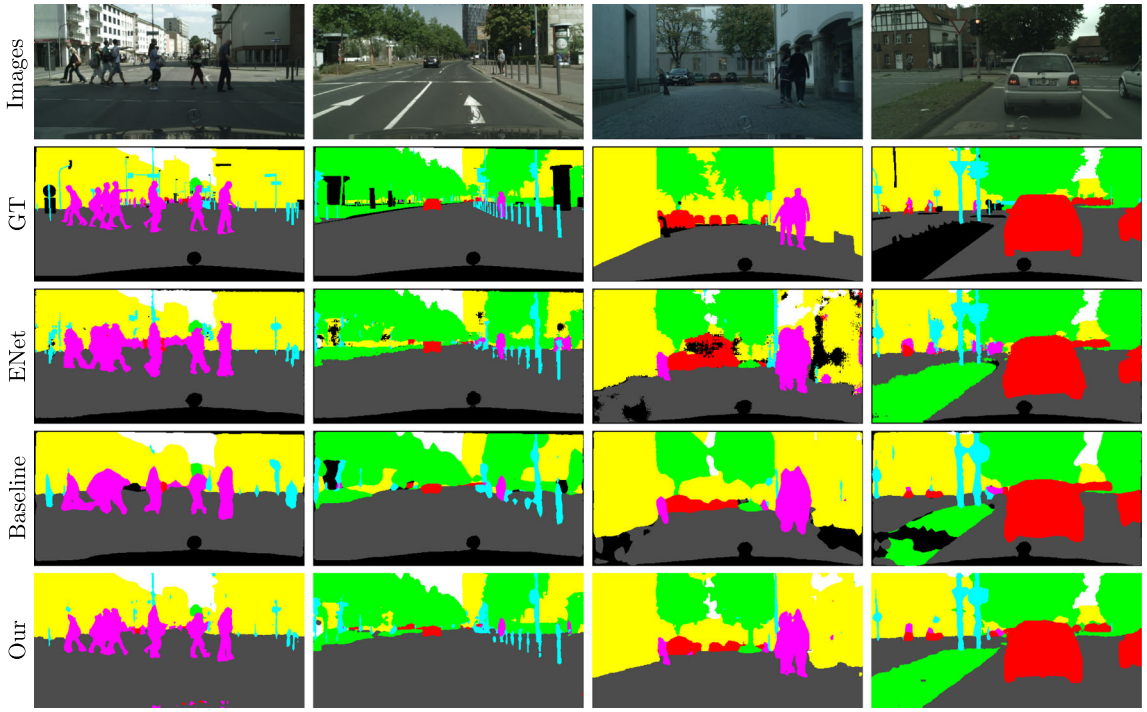

Fig. 4 Comparison results on Cityscapes dataset. Our method generates cleaner and finer prediction, such as the pedestrian in the first column and the road in the third column

The empirical evaluation results are reported in Table 2. For comprehensive comparison, we report results based on various frame resolutions. From the results, we can observe that ENet contains less parameters than the baseline but performs slower. This is due to the heavy computation cost of bottleneck and ENet is a deep structure of bottleneck. Thus, though our model contains about 4× more parameters than ENet [22], the running speed of our proposed method is still at least 2× faster than theirs. Our approach is able to obtain realtime inference performance on 1080P images. In our experiment, it even executes a network forwarding at 200fps. We can also see that the baseline obtains a slightly higher fps. In the following experiments, we will show that the extra time cost contributes to a much higher accuracy. Note that we do not make use of any neural network speedup techniques, such as pruning filters and binarizing weights, which are verified to be nondestructive on accuracy.

Cityscapes Dataset

Cityscapes [5] is a popular dataset for semantic urban scene understanding. Data was captured in 50 cities during several months, daytimes, and good weather conditions. The dataset contains 5000 finely annotated images of resolution 1024 × 2048. The dense annotation contains 30 common class labels of road, pedestrian, building, car, etc. Nineteen of them are selected for evaluation. It is split in 2950, 500, and 1525 images for training, validation, and testing, respectively. The ground truth of testing set is unavailable, and the evaluation is completed via submitting

Table 4 Results on CamVid test set

Method	build.	tree	sky	car	sign	road	ped.	fence	pole	swalk	bike	Acc.	IoU
SegNet [1]	88.8	87.3	92.4	82.1	20.5	*97.2*	57.1	49.3	27.5	84.4	30.7	65.2	55.6
ENet [22]	74.7	77.8	*95.1*	*82.4*	*51.0*	95.1	*67.2*	51.7	35.4	86.7	34.1	68.3	51.3
Baseline	*90.1*	80.5	93.9	80.2	47.9	96.3	52.4	50.3	22.7	84.6	43.5	67.5	57.8
Ours	90.0	86.3	94.4	81.5	48.8	96.7	62.8	*54.6*	28.8	*89.4*	*48.0*	*71.0*	*61.1*

The italic values show the results with the highest accuracy

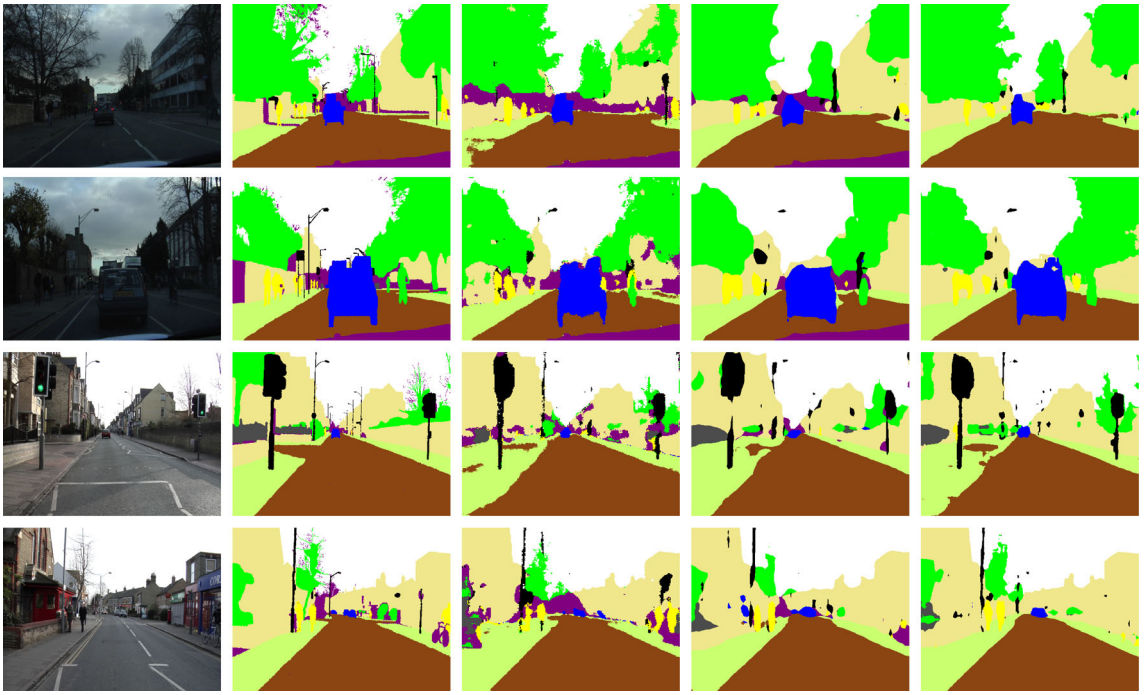

Fig. 5 Visual comparison on CamVid dataset

Table 5 Results on Kitti test set

Method	build.	sky	road	veg.	swalk	car	ped.	cyclist	sign.	fence	Acc.	IoU
ENet [22]	92.5	92.2	82.4	94.6	81.3	79.2	61.1	0.2	22.2	0.2	60.6	50.1
Baseline	91.7	97.8	92.9	97.6	67.6	83.2	63.9	*20.7*	29.6	18.2	66.3	57.6
Ours	*96.3*	*98.6*	86.5	*98.3*	*92.1*	*87.9*	*84.1*	11.0	*36.9*	*23.1*	*71.5*	*62.9*

The italic values show the results with the highest accuracy

Fig. 6 Visual comparison on Kitti dataset

Table 6 Result on Helen test set

Method	FaceSkin	LeftEyebrow	RightEyebrow	LeftEye	RightEye	Nose	UpperLip	InnerMouth	LowerLip	Hair	Acc.	IoU
Enet [22]	94.9	80.9	*83.6*	83.5	84.8	*92.4*	73.5	78.6	83.4	96.3	85.2	73.2
Baseline	93.7	77.0	81.0	75.8	84.0	88.6	77.6	72.0	83.9	96.5	83.0	71.4
Ours	*95.2*	*83.0*	83.1	*87.6*	*85.8*	91.5	*81.9*	*79.9*	*87.3*	96.9	*87.2*	*77.3*

The italic values show the results with the highest accuracy

predictions to the website.[3] In our experiment, we only perform evaluation on the validation set and subsample the resolution to 256 × 512 for fair comparison.

As shown in Table 3, our proposed method outperforms ENet both on Accuracy and IoU. IoU is the recommended metric of the dataset. We achieve 54.5% comparing to 46.4 and 50.2% for baseline and ENet, respectively. We can observe that the baseline can attain the best performance for some classes. In fact, the strong performance of the baseline on some classes is the result of inferior performance of other classes. For instance, the baseline predicts much region of sidewalk as road. Several visual examples are illustrated in Fig. 4.

Ablation Study To show the effectiveness of our proposed method, we conduct ablation experiments with several settings on Cityscapes dataset. We evaluate the performance of baseline method, compared with the performance with and without our proposed HDblock and CFblock. As shown in Table 3, the results of baseline is better than that of baseline both on accuracy and IoU metrics. However, the performance of our proposed approach is improved significantly, which is on par with ENet by taking advantage of our proposed HDblock or CFblock. This demonstrates that our proposed HDblock and CFblock layers are effective for semantic segmentation.

CamVid Dataset

CamVid [2] is a road scene understanding database. It contains 367 images for training, 100 images for validation, and 233 image for testing. To facilitate fair comparison, we do not use the 100 images of validation split as ENet [22] in our experiment. The original frame resolution for this database is 960 × 720. We downsampled all images into 480 × 360 as the reference methods. The images were manually annotated with 32 classes. As suggested in [31], we make use of a subset of 11 classes, including building, tree, sky, car, sign, road, pedestrian, fence, pole, sidewalk and bicyclist.

The detailed results for each category are shown in Table 4. Note that the result of ENet is obtained from the original paper [22]. For a convenience view, we also include the result of SegNet [1] also provided from [22]. Our method achieves an accuracy score of 71.0% and mean IoU score of 61.1%, which are both significant higher than other methods, especially for ENet [22]. Several visual examples are shown in Fig. 5. We find that our method generate more clean and steady prediction than ENet.

Kitti Dataset

Kitti [9] is one of the most popular datasets for autonomous driving. It contains many tasks, such as tracking, object detection, and odometry. It does not officially contain ground truth label for semantic segmentation. We employ a subset of images that are manually annotated by Zhang et al. [39]. It totally includes 252 images, where 140 images are for training and 112 for testing. These images were manually annotated with ten object categories, i.e., building, sky, road, vegetation, sidewalk, car, pedestrian, cyclist, signage, and fence. Moreover, the ground truth contains some regions which are not annotated. which is labeled as void. In our experiment, images are uniformly resized to 368 × 1232 for training and testing. We employ Kitti as a complement dataset as the image resolution is significant different to the former two.

The results are shown in Table 5. It is easy to find that our method outperforms both baseline and ENet at a large margin. Our approach outperforms other methods in almost all categories. We achieve significant higher accuracy on "pedestrian," "sidewalk," and "Fence" categories. We show some qualitative results on Fig. 6. It can be seen that ENet fails to distinguish between pedestrian and cyclist, which is also indicated in Table 5, as its accuracy on "cyclist" is only 0.2%.

Helen Dataset

Helen is a collection of 2330 high resolution face portraits downloaded from *Flickr*. The dataset was originally collected by Le et al. [17]. Moreover, the segment label annotations are provided by Smith et al. [30]. Eleven segment label types for each image are provided, including face skin, left eyebrow, right eyebrow, left eye, right eye, nose, upper lip, inner mouth, lower lip, hair, and background. The dataset is divided into 2000/230/100 image for training, validation and testing, respectively. The

[3] https://www.cityscapes-dataset.com/submit/

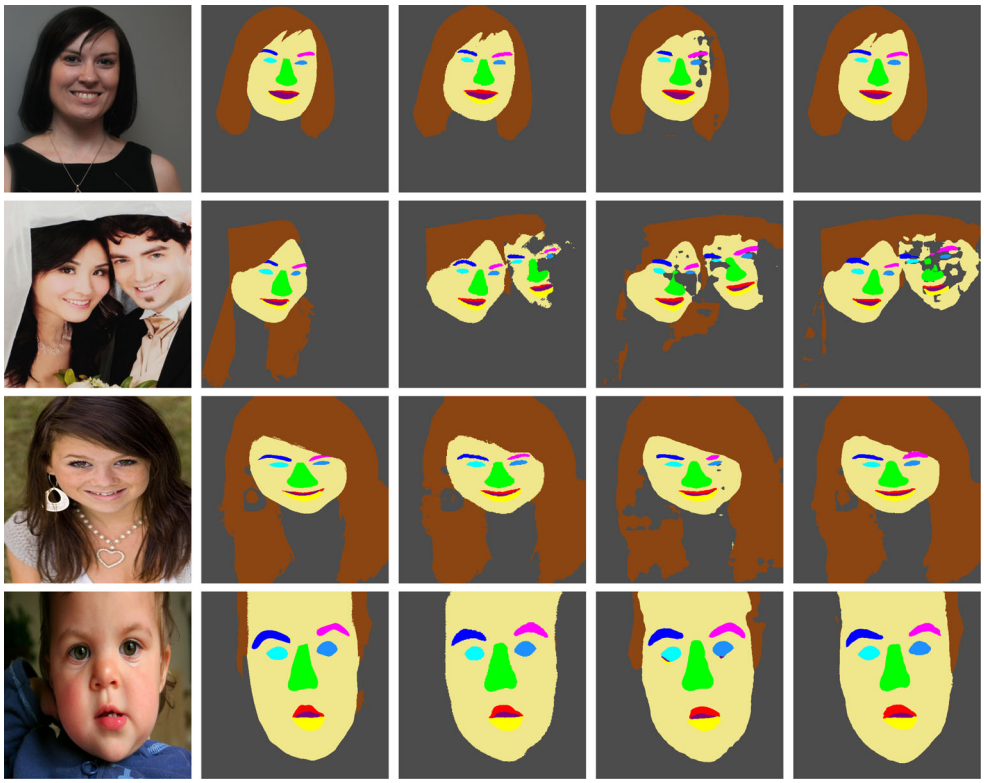

Fig. 7 Visual comparison on Helen dataset

(a) Images (b) Ground truth (c) ENet [22] (d) Baseline (e) Ours

resolutions of each image are varied. So, we resize them into 512×512 in our experiment for convenient comparison.

We evaluate the robustness of our proposed method via Helen. As shown in Table 6, our method still outperforms ENet and baseline method on total different scenario. Several visual examples are illustrated in Fig. 7.

Conclusion

We have proposed an efficient convolution neural network for semantic image segmentation. Inspired by multi-scale cognitive mechanisms, we introduce a hierarchical dilation block to provide various kinds of filed-of-view for deep neural network. This enables us to adopt multi-scale features effectively. According to cognition-based studies on contextual effects, we provide an effective strategy to integrate context information. The experimental results on urban scene understanding benchmark and face parsing dataset demonstrate the efficacy of our proposed approach.

In spite of the benefits of our proposed blocks, our method is still not able to outperform ENet on all the classes. In the future, we consider to use a robust network backbone and combine some speedup techniques.

Acknowledgments This work is supported by the National Key Research and Development Program of China (No. 2016YFB1001501).

Compliance with Ethical Standards

Conflict of Interests Jianke Zhu has received research grants from Alibaba Group.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. 2015. arXiv:1511.00561.
2. Brostow GJ, Fauqueur J, Cipolla R. Semantic object classes in video: A high-definition ground truth database. Pattern Recogn Lett. 2009;30(2):88–97.
3. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 2016. arXiv:1606.00915.
4. Collobert R, Kavukcuoglu K, Farabet C. Torch7: A matlab-like environment for machine learning. In: BigLearn, NIPS Workshop, number EPFL-CONF-192376; 2011.
5. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 3213–3223.

6. Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. p. 248–255. IEEE; 2009.
7. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 2650–2658.
8. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. 2017. arXiv:1704.06857.
9. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. Int J Robot Res. 2013;32(11):1231–1237.
10. Gros C. Cognitive computation with autonomously active neural networks: an emerging field. Cogn Comput. 2009;1(1):77–90.
11. 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; 2016. p. 770–778.
12. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. Binarized neural networks. In: Advances in neural information processing systems; 2016. p. 4107–4115.
13. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. 2016. arXiv:1602.07360.
14. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. 2014. arXiv:1408.5093.
15. Kingma D, Adam JB. A method for stochastic optimization. arXiv preprint. 2014. arXiv:1412.6980.
16. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
17. Le V, Brandt J, Lin Z, Bourdev L, Huang T. Interactive facial feature localization. Comput Vision–ECCV. 2012;2012:679–692.
18. Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient convnets. 2016. arXiv:1608.08710.
19. Liu B, Wang M, Foroosh H, Tappen M, Pensky M. Sparse convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 806–814.
20. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, Springer; 2016. p. 483–499.
21. Noh Hyeonwoo, Hong Seunghoon, Han Bohyung. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 1520–1528.
22. Paszke A, Chaurasia A, Kim S, Culurciello E. Enet: A deep neural network architecture for real-time semantic segmentation. 2016. arXiv:1606.02147.
23. Pylyshyn ZW. Computation cognition: Toward a foundation for cognitive science. Cambridge: The MIT Press; 1986.
24. Rastegari M, Ordonez V, Redmon J, Farhadi A. Xnor-net: Imagenet classification using binary convolutional neural networks. In: European Conference on Computer Vision, Springer; 2016. p. 525–542.
25. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 779–788.
26. Roy A, Todorovic S. A multi-scale cnn for affordance segmentation in rgb images. In: European Conference on Computer Vision, Springer; 2016. p. 186–201.
27. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–651.
28. Shotton J, Johnson M, Cipolla R. Semantic texton forests for image categorization and segmentation. In: IEEE Conference on Computer vision and pattern recognition, 2008. CVPR 2008, IEEE; 2008. p. 1–8.
29. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
30. Smith BM, Li Z, Brandt J, Lin Z, Yang J. Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 3484–3491.
31. Sturgess P, Alahari K, Ladicky L, Torr PHS. Combining appearance and structure from motion features for road scene understanding. In: BMVC 2012-23rd British Machine Vision Conference. BMVA; 2009.
32. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.
33. Wang Y, Zhao Q, Bo W, Wang S, Zhang Y, Guo W, Feng Z. A real-time active pedestrian tracking system inspired by the human visual system. Cogn Comput. 2016;8(1):39–51.
34. Wen G, Hou Z, Li H, Li D, Jiang L, Xun E. Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cogn Comput. 2017;9(5):597–610.
35. Xie J, Lu Y, Zhu L, Chen X. Semantic image segmentation method with multiple adjacency trees and multiscale features. Cogn Comput. 2017;9(2):168–179.
36. Fisher Y, Koltun V. Multi-scale context aggregation by dilated convolutions. 2015. arXiv:1511.07122.
37. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer; 2014. p. 818–833.
38. Zeng Dan, Zhao Fan, Shen Wei, Ge Shiming. Compressing and accelerating neural network for facial point localization. Cognitive Computation. 2017.
39. Zhang R, Candra SA, Vetter K, Zakhor A. Sensor fusion for semantic segmentation of urban scenes. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE; 2015. p. 1850–1857.
40. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. 2016. arXiv:1612.01105.
41. Zhao J, Chun D, Sun H, Liu X, Sun J. Biologically motivated model for outdoor scene classification. Cogn Comput. 2015;7(1):20–33.
42. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Zhizhong S, Dalong D, Huang C, Torr PHS. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 1529–1537.
43. Zhou A, Yao A, Guo Y, Xu L, Chen Y. Incremental network quantization: Towards lossless cnns with low-precision weights. 2017. arXiv:1702.03044.
44. Zhou S, Wu Y, Ni Z, Zhou X, Wen H, Zou Y. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. 2016. arXiv:1606.06160.