Ccfnet: Channel-Communication Factorization ConvNet for Real-Time Semantic Segmentation

J Lin\textsuperscript{1, a} and J Hu\textsuperscript{2, b}

\textsuperscript{1}School of Electronic and Information Engineering, Sun Yat-sen University, 132 East Waihuan Road, Guangzhou Higher Education Mega Center, Guangzhou, Guangdong 510006, China
\textsuperscript{2}School of Data and Computer Science, Sun Yat-sen University, 132 East Waihuan Road, Guangzhou Higher Education Mega Center, Guangzhou, Guangdong 510006, China

\textsuperscript{a}linjling7@sysu.edu.cn; \textsuperscript{b}hujguo@mail.sysu.edu.cn

Abstract. In this paper, we address a challenging real-time semantic segmentation task by proposing a lightweight encoder-decoder architecture. The architecture called channel-communication factorization ConvNet (CCFCNet), in which both efficiency and accuracy are taken into account. The core of our work is channel-communication factorization convolution module (CCFC) and dense dialation-rate block (DDB). CCFC module, where split, channel communication and fuse operations are utilized to greatly save reasoning time and improve the quality of information refinement with a few additional calculations. Meanwhile, four CCFC modules with different dialation rates form a dense dialation-rate block (DDB), which can obtain denser feature refinement and enlarge receptive field to improve the segmentation accuracy of objects with different size in images. The proposed architecture, which contains only 1.25M parameters, achieves a mean IOU of 71.4% on the Cityscapes dataset and can run over to 42FPS on a single GTX 1080Ti GPU. It makes a better trade-off between accuracy and efficiency than state-of-the-art methods with comparable performance.

1. Introduction

As a multi-dimensional signal, image has a wealth of information, but understanding the image requires the help of extremely complex algorithms. As a crucial role in image understanding, semantic segmentation is a sub-direction in the area of computer vision, and its task is to assign category to each pixel of the images [1]. In recent years, segmentation algorithms have been used in areas such as autonomous driving, distance education, and telemedicine, which require very long data processing time. Obviously, the current time-consuming algorithms cannot be implemented on smart mobile platforms where computation resources are greatly limited. Therefore, it is crucial to create an energy-saving, fast and promising accurate segmentation algorithm.

Deep Convolution neural networks (DCNNs) have shown impressive capabilities on refining features with high resolution, and fully convolutional network (FCN) [2], which consists of a series of convolution layers and max-pooling layers, is the first deep convolution neural network in the field of semantic segmentation. However, segmentation based only on shallow output without pooling layers will not be able to use deep semantic information for reasoning, resulting in that although multiple stages of the pooling layer have advanced features, it will also reduce spatial details and feature
resolution. To overcome this problem, an unique skip connection structure in encoder-decoder architecture was proposed [3]. With this structure, high-level semantics and low-level fine-grained information are well integrated. Efforts are also made to build architectures with residual block stacked [4], but the applicability of multi-scale objects in the image is not taken into account in these methods. Several strategies have been proposed to alleviate the problem, such as learnable upsampling [5], pyramid networks [6]. However, these methods are computationally intensive.

Recently, methods have been proposed to balance accuracy and efficiency. For instance, applying factorization convolutions in residual blocks [7,8], popular methods for pursuing extremely tight frames, but these modules can only handle a single dilation rate. Unlike this, FCU and PFCU [9], DSP block [10] and SS-nbt module [11] employ 2D or 3D residual block with even more than one dilated rate. However, they have to divide channels into several groups that may prevent communication between channels of different groups.

To achieve a good trade-off between accuracy and efficiency, we establish a network named channel-communication factorization ConvNet (CCFCNet), which mainly involves three contributions:

- We propose an effective real-time semantic segmentation network with encoder-decoder structure named CCFCNet (as shown in figure 1), which can maintain both high resolution and accuracy on Cityscapes Dataset [1].
- We introduce a factorization module that can efficiently obtain semantic information at a low parameter level.
- We construct a dense dialation-rate block (DDB) with large receptive field and dense feature sampling capability. The block greatly boosts the learning performance of the model.

2. Literature
As mentioned in the introduction, the task of pixel-wise prediction faces two problems: enlarging receptive field and maintaining efficiency and accuracy simultaneously in semantic segmentation.

2.1. Enlarging receptive field
In current researches, multi-scale objects, which help to analysis local and global contexts, are often regarded as the key factor to improve the accuracy of semantic segmentation. Among these researches, utilizing a large convolution kernel such as 5 x 5 convolution kernel [12], is the most direct but computationally expensive way to increase receptive field. Besides, more people prefer to apply atrous convolution. There are some impressive results exhibited to us by DeepLab [13], DenseASPPNet [14]. Following the same spirit, Chen et al [15] embed global features into local contexts and Zhang et al [16] exploit long-range contexts by proposing a scale-adaptive convolution. Further, [17] design a deep feature aggregation network for fusing multi-level features. DMNet [18] exploit several dynamic filters with different sizes to enlarge the receptive field for improving the sensitivity to large-scale objects and small-scale objects. During scene parsing, some of these defects cause the time-consuming inference and much heavy overheads.

2.2. Maintaining efficiency and accuracy simultaneously in semantic segmentation
Lightweight semantic segmentation networks have been demonstrated to solve the problems of low segmentation accuracy and constraint computation resources. Among them, ENet [7] employed encoder-decoder architectures, showing us impressive performance in terms of accuracy and speed. Several multi-branch cascaded networks [19,20] are also employed for a fast execution speed. Although efforts have been devoted to designing efficient and accurate models, most of works have increased speed by sacrificing accuracy. [21] assembled Efficient Spatial Pyramid (ESP) modules into an encoder-decoder structure for semantic segmentation under resource constraints. Convolutional factorization technique [8] have been applied to build tight frames. Besides, MCIM [10] exhibited us an impressive performance of parameter reduction with high accuracy performance.
3. CCFCNet: proposed architecture

![Figure 1](image)

**Figure 1.** An overview of channel-communication factorization ConvNet (CCFCNet) and dense dialation-rate block (DDB). $D_1$ and $D_2$ are dialation rates of two branches in the CCFC modules.

3.1. Channel-communication factorization convolution module

As shown in figure 2, three operations: “split”, “channel communication”, and “fuse” are applied to the residual block named channel-communication factorization convolution (CCFC) module for better efficiency and accuracy. Firstly, we follow the idea of group convolution [22], that is, the idea of splitting the input into two branches. With this idea, the number of network parameters can be greatly reduced. Sequentially, the output of the first branch and the input of the second branch are concated to integrate information from all channels instead of working independently. That’s why we call this operation "channel communication". Besides, two groups of $1 \times 3$ and $3 \times 1$ factorization convolutions are applied instead of $3 \times 3$ convolutions in each branch. The second group of factorization convolutions of each branch are atrous convolutions, which aim at obtain larger receptive field. At last, the two low-dimensional branches are merged together by “fuse” to reorganize the feature maps, and maintain the number of output feature maps. With network performance being improved, this module does not cause parameter redundancy and significantly increase complexity, which fully promotes the flow and reuse of information.

![Figure 2](image)

**Figure 2.** CCFC module. © and  means concating and pixel-wise adding. $D_1$ and $D_2$ represent dialation rates, while $C$ is the number of channels.

![Figure 3](image)

**Figure 3.** Receptive field and features sampling of atrous convolution. (a) One-dimensional atrous convolution with dilation rate of 6. (b) A one-dimensional convolution with larger dilation rate of 6 stacked with a one-dimensional convolution with smaller dilation rate of 3.
3.2. Dense dialation-rate block

It is well known that pixels in an image can be classified by using semantic segmentation algorithms. However, the classifier is based not only on the information of each pixel itself, but also on local information. Thus, convolutional layers are expected to have as large as possible receptive fields for accurately classifying the pixels of multi-scale objects. As mentioned above, there are two branches with different dialation rates in our CCFC module. Accordingly, we propose a dense dialation-rate block (DDB), as shown in figure 1(b). The block is composed of four CCFC modules, each of which has two different dialation rates.

Using stacked atrous convolutions can not only obtain a larger receptive field, but also refine dense features, as shown in figure 3. In our scheme, we use DDB block with dense connection technique, to take advantage of stacking convolutions. And the receptive field of each branch and each CCFC module can be formulated in equation (1) and equation (2) respectively, while equation (3) is the largest receptive field \( R_{\text{max}} \) of the block. Where \( D \) is dialation rate, and \( K \) indicates the kernel size of convolution.

\[
R(D) = (D - 1) \times (K - 1) + 1
\]

\[
R_{\text{max}}^n = R_{\text{branch1}} + R_{\text{branch2}} - 1, (1 \leq n \leq 4)
\]

\[
R_{\text{max}} = R_{\text{max}}^1 + R_{\text{max}}^2 + R_{\text{max}}^3 + R_{\text{max}}^4 - 3
\]

3.3. Architecture design

Our asymmetric encoder-decoder structure is presented in table 1. The encoder consists of a series of CCFC modules, Downsampling Units and dense dialation-rate modules to generate advanced feature maps. It is able to collect higher-level semantic information and reduce the amount of computation by using Downsampling Units. Several separate CCFC modules with a dialation rate of 1 applied in the encoder are designed to capture context. And the decoder can upsample the feature maps to match the input resolution with the help of APN modules and CCFC modules.

Several traditional decoders only acquire knowledge from the encoder, which greatly limits their performance. Therefore, it is necessary to enhance their ability to learn semantic information from spatial features, as mentioned in [23]. Therefore, we modify the efficient decoder in LEDNet [11], that is, adding an Upsampling Unit and two CCFC modules after APN module to boost learning performance of the decoder.

| Stage | Layer | Type                | Size                |
|-------|-------|---------------------|---------------------|
|       | 1     | Downsampling Unit   | 32 x 512 x 256     |
|       | 2-4   | 3 x CCFC module     | 32 x 512 x 256     |
|       | 5     | Downsampling Unit   | 64 x 256 x 128     |
| Encoder | 6-7   | 2 x CCFC module     | 64 x 256 x 128     |
|       | 8     | Downsampling Unit   | 128 x 128 x 64     |
|       | 9-10  | 2 x CCFC module     | 128 x 128 x 64     |
|       | 11    | DDB block           | 128 x 128 x 64     |
|       | 12    | APN                 | C x 128 x 64       |
| Decoder | 13    | Upsampling Unit     | C x 256 x 128     |
|        | 14-15 | 2 x CCFC module     | C x 256 x 128     |
|        | 16    | Upsampling Unit     | C x 1024 x 512    |

Table 1. Composition of our architecture. The last column is the dimension of output feature maps. \( C \) represents the number of classes.
4. Analysis

4.1. Training Protocol
We evaluate our CCFCNet on the CityScapes dataset, which is popular with semantic segmentation researchers, due to its highly variable set of urban scenarios and challenging 19 labeled classes. 5000 fine-annotated 2,048 × 1,024 resolution images are available in the dataset [1], in which 2975 images for training, 500 images for validation and the rest for testing. However, the annotated images of testing are not available. All experiments were conducted on a virtual machine with a single GTX 1080TI GPU. Mean intersection over union (mIOU), speed (fps), model parameters and running time are used as criteria to comprehensively evaluate our model. Our model is trained using Adam optimization function and a batch size of 5. We set the training epoch of 150, weight decay of 0.0001 and momentum of 0.9, and start training with a learning rate of 5 x10^{-4}.

4.2. Experimental results
We implement our network based on PyTorch framework and train the encoder and decoder in separate steps. We first train the encoder and then connect a decoder to the pre-trained encoder to implement the architecture. There are two training schemes: “single dataset training”, that is, the encoder and decoder are both trained on Cityscapes; another solution is “pre-training”, which first initialize the encoder weights with ImageNet [21], and then attach the rest convolutional layers behind the pre-trained encoder. Both schemes have been adopted during our work. In order to determine the dialation rates of each module in DDB block, we have tried a variety of schemes of hybrid dialation rates (as show in table 2), and the final scheme (as shown in figure 1) presented us the best results.

Table 2. The results of using different hybrid dilation rates strategy of DDB block in our model without ImageNet pre-training.

| HDRS       | atrous convolution | mIOU(%) |
|------------|--------------------|---------|
| {(1,1), (1,1), (1,1), (1,1)} | ✓ | 69.03  |
| {(2,2), (4,4), (6,6), (8,8)} | ✓ | 69.21  |
| {(3,7), (7,11), (11,15), (15,17)} | ✓ | 69.72  |
| {(2,3), (5,8), (9,11), (13,17)} | ✓ | 70.1   |

Table 3. The comparison between our method and state-of-the-art method on cityscapes dataset. “Cla” and “Cat” are class and category mIoU scores, and CCFCNet(pre) means the encoder of our network was pretrained on ImageNet.

| Network      | Param (M) | Time (ms) | Speed (Fps) | Cla (%) | Cat (%) |
|--------------|-----------|-----------|-------------|---------|---------|
| ENet [7]     | 0.36      | 34        | 31          | 58.3    | 80.4    |
| SegNet [23]  | 29.5      | 67        | 15          | 57.0    | 79.1    |
| ESPNet [24]  | 0.4       | 9         | 112         | 64.8    | 82.2    |
| ICNet [19]   | 7.80      | 33        | 30          | 69.5    | 86.4    |
| LEDNet [11]  | 0.94      | 14        | 71          | 70.6    | 87.1    |
| ERFNet [8]   | 2.1       | 24        | 41.7        | 69.7    | 87.3    |
| CCFCNet(pre) | 1.25      | 23.8      | 42          | 71.4    | 87.4    |
Table 4. Accuracy result on Cityscapes test dataset. "Our" is our ImageNet-pretrained model.

| Method | Roa | Sid | Pol | Wal | TLI | B u i | Fen | TSI | Veg | Rid | Sky | Ped | Ter | Car | Tra | Bus | Tru | Bic | Mot |
|--------|-----|-----|-----|-----|-----|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ENet   | 96.3| 74.2| 43.4| 32.2| 34.1| 7.5   | 0   | 33.2| 44.0| 88.6| 38.4| 90.6| 65.5| 61.4| 90.6| 48.1| 50.5| 36.9| 55.4| 30.8|
| SegNet | 94.6| 73.2| 35.7| 28.4| 39.8| 8.4   | 0   | 32.0| 45.1| 87.0| 42.8| 91.8| 62.8| 63.8| 89.3| 44.1| 43.1| 38.1| 51.9| 35.8|
| ERFNet | 97.0| 77.5| 45.0| 35.0| 35.6| 7.6   | 2   | 36.1| 46.3| 90.8| 40.9| 92.6| 67.0| 63.2| 92.3| 50.1| 52.5| 38.1| 57.2| 41.8|
| ENe   | 97.1| 79.2| 61.5| 43.2| 60.4| 8.9   | 7   | 48.9| 63.4| 91.5| 56.1| 93.5| 74.6| 68.3| 92.6| 51.3| 72.7| 51.3| 70.5| 53.6|
| LIDNet | 98.1| 79.5| 62.8| 47.7| 61.3| 9.1   | 6   | 49.9| 72.8| 92.6| 53.7| 94.9| 76.2| 61.2| 90.9| 52.7| 64.0| 64.4| 71.6| 44.4|
| BRNet | 97.9| 82.1| 59.0| 45.2| 62.6| 9.0   | 7   | 50.4| 68.4| 91.9| 59.8| 94.2| 78.5| 69.4| 93.4| 53.7| 60.8| 52.3| 64.2| 49.9|
| Ours  | 97.8| 81.4| 61.5| 47.9| 62.0| 90.5  | 5   | 72.9| 93.6| 57.0| 95.1| 76.6| 70.0| 94.0| 53.9| 66.6| 62.3| 69.3| 47.8 |

We compare the results of our network and several state-of-the-art real-time semantic segmentation networks. As depicted in table 3, CCFCNet contains only 1.25M parameters and can process 42 frames per second. Compared with some other methods, our ImageNet-pretrain network performs the best in both class and category scores. Its class score is even 25% higher than SegNet and it also performs better in speed. In table 4, we list the results for 19 categories on the Cityscapes test set. In these categories, our architecture has achieved good segmentation accuracy, especially in the categories: “wall”, “fence”, “vegetation”, “terrain”, “sky”, “car” and “truck”. It is demonstrated that the pixels of multi-scale objects in images are generally well classified by our method. There are some visual examples of the segmentation results of CCFCNet in figure 4.

Figure 4. Some visual examples on Cityscapes database. The three lines of images are original images from the training or validation set of the database, segmentation results and ground truths.

5. Conclusions
In conclusion, we have proposed an encoder-decoder real-time semantic segmentation network named CCFCNet. There are two highlights in our architecture: channel-communication factorization convolution module (CCFC) and dense dilation-rate block (DDB). Utilizing the former can strike a balance between efficiency and accuracy to build a lightweight network, while employing the latter can effectively promote the refinement of dense features with large receptive field. Both theoretical analysis and experimental results have demonstrated that our method has a good performance on multi-scale objects. Our scheme makes a better trade-off between accuracy and efficiency than state-of-the-art methods with a speed of 42 fps and mean IOU of 71.4%.

References
[1] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and...
Schiele B 2016 The cityscapes dataset for semantic urban scene understanding Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 3213-23
[2] Long J, Shelhamer E and Darrell T 2015 Fully convolutional networks for semantic segmentation Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 3431-40
[3] Chen L C, Papandreou G, Kokkinos I and Murphy K and Yuille A L 2017 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4) 834-48
[4] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation Int. Conf. on Medical Image Computing and Computer-assisted Intervention pp 234-241
[5] Yuan Y and Wang J 2018 Ocnet: Object context network for scene parsing Preprint arXiv/1809.00916
[6] Zhao H, Shi J, Qi X, Wang X and Jia J 2017 Pyramid scene parsing network Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 2881-90
[7] Paszke A, Chaurasia A, Kim S and Culurciello E 2016 Enet: A deep neural network architecture for real-time semantic segmentation Preprint arXiv/1606.02147
[8] Romero E, Alvarez J M, Bergasa L M and Arroyo R 2017 Erfnet: Efficient residual factorized convnet for real-time semantic segmentation IEEE Transactions on Intelligent Transportation Systems 19(1) 263-72
[9] Wang Y, Zhou Q, Xiong J, Wu X and Jin X 2019 ESNet: An Efficient Symmetric Network for Real-Time Semantic Segmentation Chinese Conf. on Pattern Recognition and Computer Vision (PRCV) pp 41-52
[10] Jiang B, Tu W, Yang C and Yuan J 2019 Context-Integrated and Feature-Refined Network for Lightweight Urban Scene Parsing Preprint arXiv/1907.11474
[11] Wang Y, Zhou Q, Liu J, Xiong J, Gao G, Wu X and Latecki L J LEDNet: A 2019 Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation Preprint arXiv/1905.02423
[12] Lecun Y et al 1998 Gradient-based learning applied to document recognition Procs. of the IEEE 86(11) 2278-324
[13] Chen L C, Papandreou G, Schroff F and Adam H 2017 Rethinking atrous convolution for semantic image segmentation Preprint arXiv/1706.05587
[14] Yang M, Yu K, Zhang C, Li Z and Yang K 2018 Denseaspp for semantic segmentation in street scenes Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 3684-92
[15] Chen L C, Zhu Y, Papandreou George, Schroff F and Adam H 2018 Encoder-decoder with atrous separable convolution for semantic image segmentation Procs. of the European Conf. on Computer Vision (ECCV) pp 801-18
[16] Zhang R, Tang S,Zhang Y,Li J and Yan S 2017 Scale-adaptive convolutions for scene parsing Procs. of the IEEE Int. Conf. on Computer Vision pp 2031-9
[17] Li H, Xiong P, Fan H and Sun J 2019 Dfanet: Deep feature aggregation for real-time semantic segmentation Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 9522-31
[18] He J, Deng Z and Qiao Y 2019 Dynamic Multi-scale Filters for Semantic Segmentation Procs. of the IEEE Int. Conf. on Computer Vision pp 3562-72
[19] Zhao H, Qi X, Shen X, Shi J and Jia J 2018 Icnet for real-time semantic segmentation on high-resolution images Procs. of the European Conf. on Computer Vision (ECCV) pp 405-420
[20] Yu C, Wang J, Peng C, Gao C, Yu G and Sang N 2018 Bisenet: Bilateral segmentation network for real-time semantic segmentation Procs. of the European Conf. on Computer Vision (ECCV) pp 325-41
[21] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks Advances in Neural Information Processing Systems pp 1097-105
[22] Das A, Kandan S, Yogamani S and Krizek P 2019 Design of Real-time Semantic Segmentation Decoder for Automated Driving Preprint arXiv/1901.06580

[23] Badrinarayanan V, Kendall A and Cipolla R 2017 Segnet: A deep convolutional encoder-decoder architecture for image segmentation IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12) 2481-2495

[24] Mehta S, Rastegari M, Caspi A, Shapiro L and Hajishirzi H 2018 Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation Procs. of the European Conf. on Computer Vision (ECCV) pp 552-68