SCIENCE CHINA
Information Sciences

• NEWS & VIEWS •

December 2019, Vol. 62 227101:1–227101:2
https://doi.org/10.1007/s11432-019-2685-1

# An open-source project for real-time image semantic segmentation

Quan ZHOU[1,2*], Yu WANG[1,2], Jia LIU[1,2], Xin JIN[3,4] & Longin Jan LATECKI[5]

[1]*National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing* 210003*, China;*
[2]*Key Laboratory of Ministry of Education for Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing* 210003*, China;*
[3]*Beijing Electronic Science and Technology Institute, Beijing* 100070*, China;*
[4]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191*, China;*
[5]*Department of Computer and Information Sciences, Temple University, Philadelphia* PA 19122*, USA*

Received 22 July 2019/Revised 23 August 2019/Accepted 7 October 2019/Published online 28 October 2019

**Citation** Zhou Q, Wang Y, Liu J, et al. An open-source project for real-time image semantic segmentation. Sci China Inf Sci, 2019, 62(12): 227101, https://doi.org/10.1007/s11432-019-2685-1

Recent years have witnessed great progress of deep convolutional neural networks (DCNNs) for solving scene understanding tasks [1–3]. These advances prefer to construct deeper and larger network to achieve higher accuracy, yet with the sacrifice of implementing efficiency. In the context of many real-world scenarios, such as augmented reality, robotics, and self-driving, the computationally cheap networks are often required to carry out real-time estimation and decision. Therefore, those accurate networks requiring enormous resources are not suitable for the mobile devices (e.g., drones, robots, and smartphones), which have limited energy overhead, restrictive memory constraints, and reduced computational capabilities. Recently, it is widely accepted that pursuing the best performance in limited computational budgets has become a primary trend in computer vision. To this end, this essay introduces an open-source project of a lightweight encoder-decoder network (EDN) for the task of real-time image semantic segmentation.

*Network overview.* The entire architecture of LEDNet is shown in Figure 1, which composes of encoder and decoder counterpart. Similar to most EDNs, the encoder employs convolution and pooling operation to abstract high-level features. Inspired by the convolution factorization principle [4], however, the core component of the encoder is a novel residual module, called ss-nbt, that adopts a split-transform-merge strategy, approaching the representational power of large and dense convolution layers. In contrast to ShuffleNet [5] that adopts depthwise and $1 \times 1$ group convolution, our ss-nbt employs factorized convolution to avoid using pointwise convolution, saving a large number of computational costs. In addition, the residual layer of ShuffleNet [5] only performs convolution on half number of input feature channels. Conversely, our split-transform-merge strategy allows both split branches to undergo a set of factorization convolution to enhance network representation ability. Note the downsampling is postponed in the encoder, adding a bit of computational burden, but helping to gather more context. To improve segmentation performance, we utilize dilated convolutions to enlarge receptive field. In contrast to the approaches that extend field-of-view using larger kernel sizes, this technique is more effective in terms of computational cost and size of models.

On the other hand, unlike most EDNs that sequentially enlarge feature resolution using decovolution operation, the decoder of LEDNet adopts an attention mechanism [6] to reweight convolutional feature responses, in which an attention pyramid module (APN) is employed to model the interde-

* Corresponding author (email: quan.zhou@njupt.edu.cn)

 

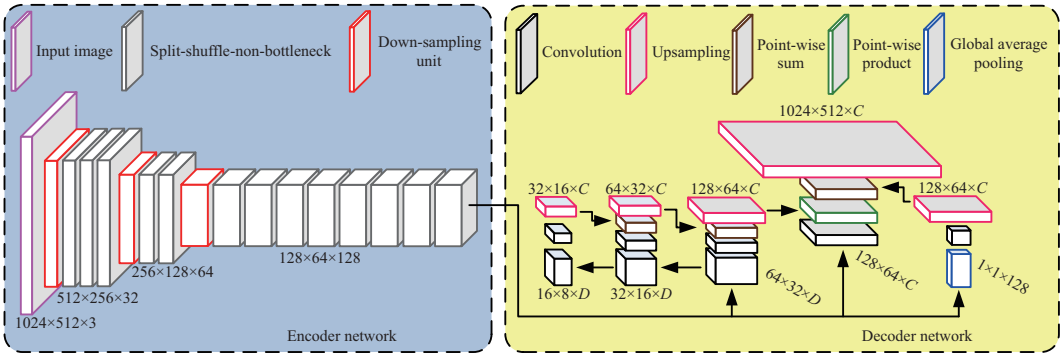


**Figure 1** (Color online) The overall asymmetric architecture of the proposed LEDNet. The encoder employs an FCN-like network, while an attention pyramid network is adopted in the decoder.

pendencies between features within different spatial locations and different channels. The pyramid structure fuses information of different scales step-by-step, where the context cues of neighbor scales are integrated more precisely. Hereafter, the abstracted attention weights are pixel-wisely multiplied with the convolutional features, derived from the output of encoder using a $1 \times 1$ convolution. To further boost accuracy, a global average pooling branch is added to integrate global prior attention. Finally, an upsampling operation is implemented to recover the resolution. Benefiting from APN, the decoder of LEDNet can capture multi-scale context cues, and produce pixel-channel-level attention for convolutional features.

*Usage.* LEDNet is an open-source project for the task of real-time image semantic segmentation. One may first build up implementing environment with at least PyTorch 0.4.1, Cuda 9.0, Cudnn 7.1, and Python 3.6. If one wants to train the model, the Cityscapes dataset [7] should be first downloaded, which includes 5000 finely annotated ground truth and over 20000 coarsely annotated images collected from 50 different European cities. This dataset is divided into three parts, where 2975, 500, and 1525 images are used respectively for training, validation, and testing. After the file path of training data and model hyper-parameters is correctly set, one may perform "main.py" to train LEDNet only using finely annotated training data. If one wants to train the model with additional coarsely annotated images, the training data should be reloaded and the whole LEDNet framework should be rebuilt. In addition, we provide the code to train encoder of LEDNet using Imagenet dataset, where one can first run "lednet_imagenet.py", and then fine-tune the trained model using training images. After the network is well trained, LEDNet can be evaluated using "eval_cityscapes_server.py" in terms of intersection-over-union (mIoU), inference time (FPS) and model size (number of parameters). In this project, we release two lightweight versions of LEDNet, where the difference, as shown in Figure 1, is the number of feature channels $D$ in APN of the decoder. One may choose to train heavier LEDNet when $D = 128$, or lighter version when $D = 20$. The lighter version yields 70.6% class mIoU and 87.1% category mIoU, respectively, and has only 0.94 M parameters with 71 FPS inference speed using a single GTX 1080Ti GPU. The heavier version has nearly 3× larger model size than lighter one, but achieves 1.6% improvement in terms of class mIoU. Compared with recent state-of-the-art lightweight networks [5, 8–10], the proposed LEDNet achieves superior performance in terms of segmentation accuracy and implementing efficiency trade-off.

*Access method.* LEDNet can be downloaded from https://github.com/xiaoyufenfei/LEDNet.

## References


1 Geng Q C, Zhou Z, Cao X C. Survey of recent progress in semantic image segmentation with CNNs. Sci China Inf Sci, 2018, 61: 051101
2 Li X L, Shi J H, Dong Y S, et al. A survey on scene image classification (in Chinese). Sci Sin Inform, 2015, 45: 827–848
3 Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network. In: Proceedings of CVPR, 2016. 6230–6239
4 Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of CVPR, 2016. 2818–2826
5 Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of ECCV, 2018. 122–138
6 Hu J, Shen L, Sun G, et al. Squeeze-and-excitation networks. In: Proceedings of CVPR, 2018. 7132–7141
7 Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of CVPR, 2016. 3213–3223
8 Howard A G, Zhu M L, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. 2017. ArXiv: 1704.04861
9 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495
10 Paszke A, Chaurasia A, Kim S, et al. Enet: a deep neural network architecture for real-time semantic segmentation. 2016. ArXiv: 1606.02147