Supplementary Materials of When Humans Meet Machines: Towards Efficient Segmentation Networks

Peike Li
peike.li@student.uts.edu.au
Xuanyi Dong
xuanyi.dong@student.uts.edu.au
Xin Yu
xin.yu@uts.edu.au
Yi Yang
yi.yang@uts.edu.au

ReLER Lab
Australian Artificial Intelligence Institute
University of Technology Sydney
Sydney, AUSTRALIA

1 Architecture Analysis

We visualize the detailed network configuration of HMSeg and TinyHMSeg in Figure 1. Since our searching space is based on the inverted residual blocks, we list the searched architecture structure layer-wisely. *IR-kxk-e* denotes the inverted residual block with kernel size k and expansion ratio e.

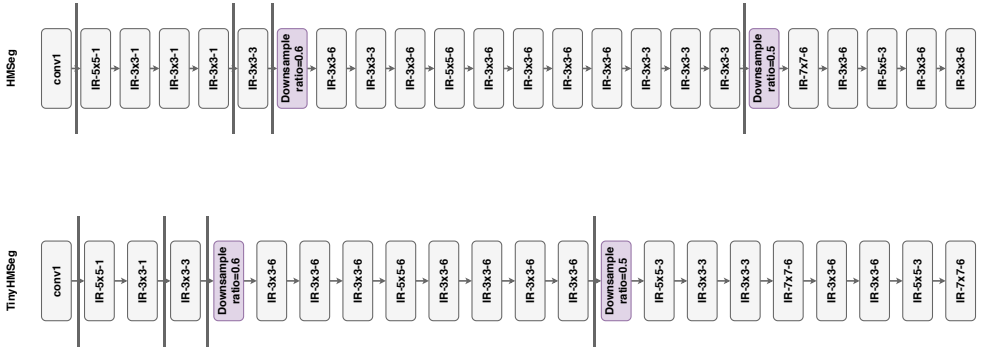

Figure 1: Network structure of MobileNetV2, HMSeg and TinyHMSeg. *IR-kxk-e* denotes the inverted residual block with kernel size k and expansion ratio e. Stage 1-5 are successively separated by the gray lines.

2 Preliminary Computational Cost Analysis.

In most CNN-based semantic segmentation prior works [3, 10], convolutional layers are widely employed. Thus we first analyze the computational complexity of this basic convolutional operator. Given an input feature map $U \in \mathbb{R}^{c_i \times h \times w}$ as an input tensor, a convolution operator ϕ with a $k \times k$ filter is applied to obtain an output feature map $V \in \mathbb{R}^{c_o \times \frac{h}{s} \times \frac{h}{s}}$, where c, h, w, s denote the number of channels, height, width and stride, respectively. For the conventional convolution operator, the total computational cost is $O(\phi) \approx c_i c_o k^2 hw/s^2$ and model parameters are calculated as $S(\phi) \approx c_i c_o k^2$. In most layers of Resnet-like CNNs, the input channel numbers are the same as the output ones. Thus, the computational cost of the layer grows quadratically as feature resolution or channel numbers increase.

3 Comparison with state-of-the-art methods.

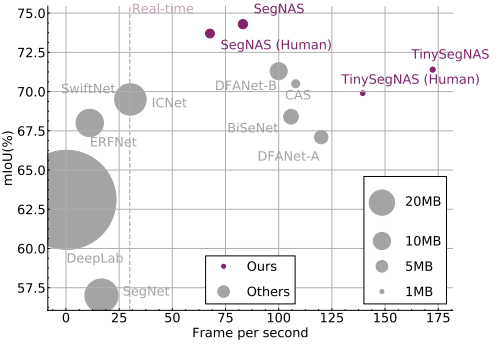

Figure 2: Comprehensive network comparisons on the Cityscapes dataset, including inference speed (frame per second), network performance (mIoU) and model size (number of parameters). State-of-the-art methods, ICNet [10], ERFNet [7], SwiftNet [6], BiSeNet [8], DFANet [4], and CAS [9], are compared. Two classical networks SegNet [1], DeepLab [3] are also included. Our human-machine collaboration designed model HMSeg achieves a better trade-off between speed and accuracy.

4 More implementation details about optimizing the architecture.

Without using a proxy task, we search the architecture directly on the Cityscapes training set. The training set is divided into two parts, *i.e.*, the `train` and `val` sets. The original validation set is not used in the searching procedure to avoid over-fitting. We search the model for 320 epochs. We optimize the weights via SGD with a cosine decayed learning rate from 0.01 and optimize the architecture parameters via Adam with a learning rate of 0.001. For HMSeg model, we set the target latency time as 12ms. To achieve more efficiency in TinyHMSeg, we reduce all the channel numbers by a factor of 2 and set the target value to 6ms.

5 More details about experiments on CamVid

CamVid [2] is a relative small-scale dataset for urban scene understanding. The dataset is extracted from five video sequences from a driving vehicle. It contains 701 images, which are split into 367, 101 and 233 images as the training, validation and testing sets respectively. As the size of this dataset is faraway enough to training from scratch, we adopt the model trained on Cityscapes as a pre-train model and further fine-tune for 120 epochs. The input resolution of images is 540x720 pixels and SGD is used with a base learning rate of 0.01 and weight-decay of 5e-4. We show some visualization results in Figure 3.

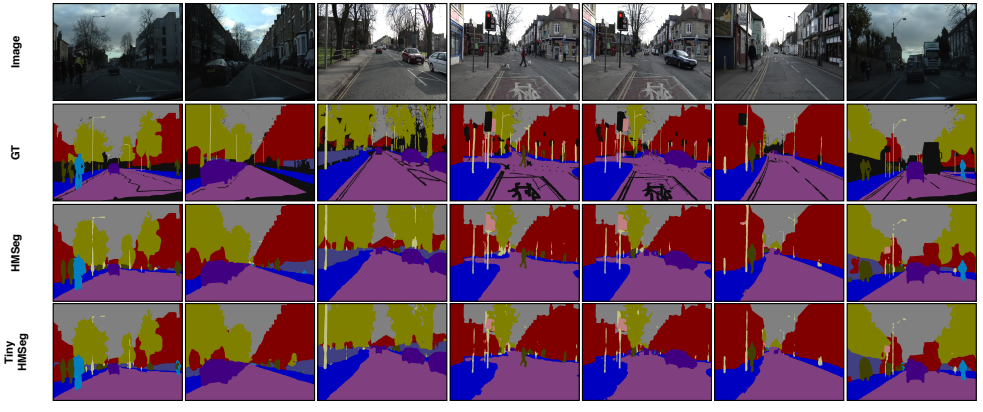

Figure 3: Visualization results of our HMSeg and TinyHMSeg on CamVid validation set. Better zoom in to see details.

6 More details about experiments on LIP

LIP [5] is one of the largest datasets for the single person human parsing task. The dataset contains 50462 images in total, split into 30462, 10000, and 10000 images as the training, validation and testing sets. We train the network for 150 epochs. Here, we show some visualization results in Figure 4.

Figure 4: Visualization results of our HMSeg and TinyHMSeg on LIP validation set. Better zoom in to see details.

References

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

[2] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

[4] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *CVPR*, pages 9522–9531, 2019.

[5] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

[6] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *CVPR*, pages 12607–12616, 2019.

[7] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017.

[8] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018.

[9] Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, and Tao Mei. Customizable architecture search for semantic segmentation. In *CVPR*, pages 11641–11650, 2019.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[11] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, pages 405–420, 2018.