Semantic scene understanding on mobile device with illumination invariance for the visually impaired

Chengyou Xu, Kaiwei Wang*, Kailun Yang, Ruiqi Cheng and Jian Bai
State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310027, China

ABSTRACT

For Visually Impaired People (VIP), it's very difficult to perceive their surroundings. To address this problem, we propose a scene understanding system to aid VIP in indoor and outdoor environments. Semantic segmentation performance is generally sensitive to the environment and illumination changes, including the change between indoor and outdoor environments and the change across different weather conditions. Meanwhile, most existing methods have paid more attention on either the accuracy or the efficiency, instead of the balance between both of them. In the proposed system, the training dataset is preprocessed by using an illumination-invariant transformation to weaken the impact of illumination changes and improve the robustness of the semantic segmentation network. Regarding the structure of semantic segmentation network, the lightweight networks such as MobileNetV2 and ShuffleNet V2 are employed as the backbone of DeepLabv3+ to improve the accuracy with little increasing of computation, which is suitable for mobile assistance device. We evaluate the robustness of the segmentation model across different environments on the Gardens Point Walking dataset, and demonstrate the extremely positive effect of the illumination-invariant pre-transformation in challenging real-world domain. The network trained on computer achieves a relatively high accuracy on ADE20K relabeled into 20 classes. The frame rate of the proposed system is up to 83 FPS on a 1080Ti GPU.

Keywords: Semantic segmentation, robustness, illumination invariance, scene understanding, mobile device.

1. INTRODUCTION

According to the latest data from World Health Organization (WHO), there are 253 million people estimated to be visually impaired globally, and 36 million are blind[1]. Without the sense of sight, it's very difficult for visually impaired people (VIP) to perceive their surroundings, so that the quality of their life is seriously affected. Semantically interpreting traversable areas and detecting obstacles are most of the important tasks to assist the navigation of VIP[2]. To address these problems and help the VIP understand the scenes more conveniently, a lightweight, precise and fast semantic segmentation network is needed to be developed on mobile device. In this paper, we propose a scene understanding system, which is an improved semantic segmentation network based on DeepLabv3+. The system can be deployed on mobile devices, to aid VIP in indoor and outdoor environments.

Semantic segmentation used in scene understanding is a practical and challenging task in computer vision. Compared with object detection, semantic segmentation assigns a label to every pixel, and recognizes more object classes in an image[3]. Focusing on improving the accuracy, many semantic segmentation networks[4, 5, 6] use well-designed backbone. However, robustness is a significant character in practical application of semantic segmentation. Extreme illumination variance often happens when scene changes between indoors and outdoors, challenging the robustness of the segmentation network in the real world[7, 8]. Illumination invariant preprocessing is a promising approach to confirm a stable result under the changing illumination conditions. Moreover, many segmentation networks ignore the characteristic of real-time inference, which is more crucial in our VIP assistance situation. To obtain a balance between accuracy and speed, DeepLabv3+ [9] with lightweight network as the backbone achieves relatively good results in both precision and speed. DeepLabv3+ also uses the Atrous Spatial Pyramid Pooling (ASPP) [10] to receive different scales of feature maps. But in fact, ASPP is not fully utilized yet when the backbone is a lightweight network. Regarding lightweight networks, classical networks MobileNetV2[11] and ShuffleNet V2[12] have achieved good accuracy without much computation and parameters. Combination of DeepLabv3+ with ASPP and lightweight network as backbone may reach both better accuracy and real-time performance.

In this paper, illumination invariant preprocessing is carried out during the training of the models. We adopt the DeepLabv3+ as the structure of the network, and use the advanced lightweight networks MobileNetV2, ShuffleNet V2 as

the encoder. ASPP is also applied in the DeepLab structure. The performance evaluation on the Gardens Point Walking dataset proves that illumination invariance preprocessing is effective to improve the robustness of segmentation when scene changes between indoors and outdoors. Meanwhile, the speed and accuracy are improved to meet the needs of faster semantic segmentation, and be suitable to be developed on mobile device for assisting VIP. Our contributions can be listed as follows:

1. We improve the robustness of the semantic segmentation when scene changes between indoors and outdoors by adding the illumination invariance preprocessing.

2. We train a lightweight semantic segmentation model, realizing 50% mean Intersection over Union (mIoU) by using DeepLabv3+ with MobileNetV2 on 20-class ADE20K dataset, which can be used on Android mobile phone to realize a semantic scene understanding for VIP assistance.

2. RELATED WORK

2.1 Illumination invariance processing

Illumination-invariant transform is used in shadow removal， scene classification and segmentation, There are main kinds of methods to remove or reduce the impact of the illumination changing. A kind of method is intrinsic image decomposition. The concept of intrinsic image was introduced in[13]. A given RGB image can be decomposed into two intrinsic images: reflectance image, which refers to its material-dependent properties, and shading, which refers to its light-dependent properties. Reflectance image does not change with the chanve of to light condition, so it's widely used in illumination invariance processing. According to the intrinsic image theory, many approaches[14, 15, 16, 17] converted RGB color space into one channel space representing the illumination variance. The method[18] combined it with HSV color space and improved the results of CNN semantic segmentation. With the success of deep CNN, some researches obtained the intrinsic image from input image by training deep CNNs. Paper[19] developed a deep Retinex-Net learned from low-light dataset, to address low-light enhancement. There are also methods that combine the illumination invariance decomposition and semantic segmentation together with one single network, such as[20], which regarded that the two tasks have some relationships and proposed an end-to-end CNN architecture to achieve the joint learning. Even though many of these illumination invariance processing methods are useful, a simple and effective method is needed for our application.

2.2 Semantic segmentation

Semantic segmentation based on Convolutional Neural Networks (CNNs), especially Fully Convolutional Networks (FCNs) [21] has greatly improved the performance of segmentation tasks compared with methods using traditional hand-crafted features. Inspired by FCN, many more complex model variants have been proposed to get more precise and faster, without too much increase of computation. In terms of structure, many classic networks adopted encoder-decoder structure, such as U-Net[22], SegNet[23], using encoder to extract features and decoder to restore pixel position. Furthermore, influenced by ResNet, architectures like ENet[4], ERFNet[5] and ERF-PSPNet[24] paid attention to convolution blocks design to improve real-time performance. On the other hand, spatial pyramid pooling was exploited in DeepLab[25] and PSPNet[6] to obtain multi scale information.

DeepLab series are designed to increase the accuracy without raising too much computation costs. DeepLabv3+ adopted a novel encoder-decoder structure, Atrous Spatial Pyramid Pooling (ASPP)[10] and depthwise separable convolution, then made it possible to trade-off precision and runtime by changing a parameter named output stride.

To meet the needs of running the model on mobile devices for applications, lightweight feature extracting networks have the advantages of less parameters and smaller models. A series of works on lightweight architecture design have been proposed in recent years, such as Xception, MobileNetV1[26] and V2[11], ShuffleNet V1[27] and V2[12], etc. MobileNetV2 applies depthwise convolution to reduce the computation and uses inverted residuals, linear bottlenecks to ensure the quality of feature extracting. ShuffleNet V2 uses channel split and shuffle to improve the accuracy without large amount of parameters, making it a proper architecture for the network backbone of real-time semantic scene understanding system.

2.3 Network used on mobile device

With the continuous optimization of network structure, CNN models is getting smaller with less parameters, which makes it possible to develop CNN networks on mobile devices with limited computational ability, such as mobile phones

and tablets. Recently, many networks like MobileNet have been developed on IOS and Android, achieving excellent performance in object recognition, semantic segmentation and so on. For example, Google has put the real-time segmentation of human face and background on smart phone[28]; MobileNetV2 has also been used on mobile phone with semantic segmentation structure such as DeepLab for indoor navigation[29], etc.

3. METHOD

This section describes the illumination invariance preprocessing progress and the architecture of our semantic scene understanding system. Images are preprocessed into illumination invariant images and fed into the network. The network is based on the advanced semantic segmentation DeepLabv3+, with lightweight network (MobilenNetV2, ShuffleNet V2) as the encoder. After the encoder loads the pre-trained weights on ImageNet dataset, the model is trained on the adjusted ADE20K dataset. In addition, the preprocessing method is applied to improve the robustness of the model under different light conditions.

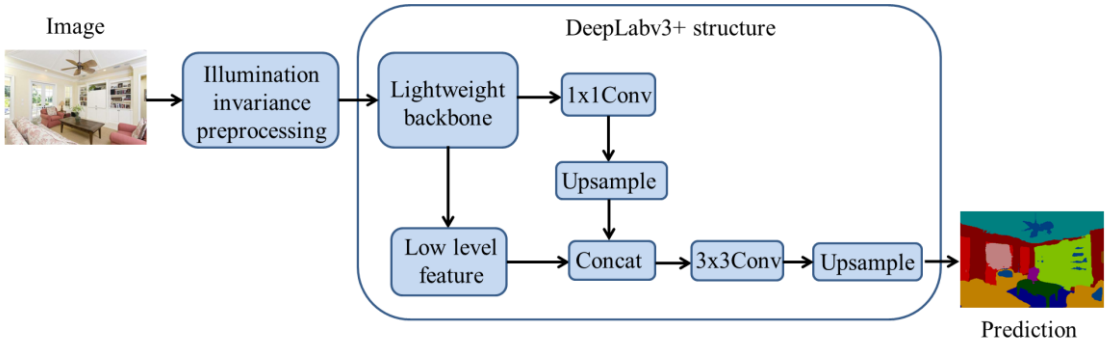

Figure1. The general view of our semantic scene understanding process.

3.1 Illumination invariance preprocessing

For purpose of improving robustness of semantic segmentation in indoor and outdoor scenes, we add the illumination invariance preprocessing before feeding the input images to the network. To choose a simple and useful preprocess method, we compare all the methods, then follow the one proposed in [15] and convert a 3-channel RGB space into a one-dimensional channel I, which is only related to the materials of objects. The response of a linear image sensor R is decided by spectral sensitivity $F(\lambda)$, emitted spectral power distribution $E(\lambda)$ incident on an object with surface reflectivity $S(\lambda)$, as it is shown by the following equation:

$$R^{x,E} = \underline{a}^x \cdot \underline{n}^x I^x \int S^x(\lambda) E^x(\lambda) F^x(\lambda) d\lambda \tag{1}$$

where the unit vectors $\underline{a}^x$ and $\underline{n}^x$ represent the direction of light source and the surface normal, respectively. I^x represents the intensity of illumination on point x in the image. Assuming the $F(\lambda)$ can be modeled as a Dirac delta function centered on λ_i.We take the logarithm of both sides, simplify the geometric factor, and consider three sensor responses R1,R2,R3 related to peak sensitivities at wavelengths $\lambda_1 < \lambda_2 < \lambda_3$, the Eq.(1) can be write as:

$$I = \log(R_2) - \alpha \log(R_1) - (1-\alpha)\log(R_3) \tag{2}$$

After transformation, the one-dimensional color space *I* is independent of the other parameters if Eq.(3) is satisfied:

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{1\text{-}\alpha}{\lambda_3} \tag{3}$$

In Eq.(3), α is a parameter associated with the peak spectral responses of the camera's sensors. Finally the one-dimensional channel I is computed by

$$I = 0.5 + \log(R_2) - \alpha \log(R_1) - (1-\alpha)\log(R_3) \tag{4}$$

Considering that ADE20K is a collection of images taken by different cameras, where it is hard to decide the value of α directly. We choose several common α values to do the experiments and see the results. Finally we take α as 0.3 and complete the illumination invariance preprocessing. The preprocessed images are shown in Figure 2.

Figure 2. The illumination invariance preprocessing on an example image from Gardens Point Walking dataset.

Gardens Point Walking[34] is a dataset of a single route through the Gardens Point Campus, Queensland University of Technology, Australia. The route passes both indoors and outdoors, and images are collected with a hand-held iPhone 5, in day and night. This dataset is quite suitable for our research. The reasons are as the follows: the images are captured in different light conditions and from viewpoints close to those of pedestrians. This dataset contains the scenes not only in indoors and outdoors, but also in day and night, which covers most of the illumination changing situations. Some scenes even involve large-scale illumination variation in one single image, such as Figure 3. It's challenging for the network to accurately segment the things in the dark part of the image, which is exactly the problem we have to overcome in this paper. What's more, the view of Gardens Point dataset is forward facing from hand-held smart phone, which is similar to the view of visually impaired pedestrians. In addition, the routes of day and night are the same, which is convenient and fair for comparison between different light conditions.

We preprocess the ADE20K dataset by using the illumination invariance method, then employ the preprocessed dataset to train the network in Section 3.2. We test the network on the preprocessed Gardens Point dataset, and study if the robustness across light conditions in indoors and outdoors, day and night has been improved by using this preprocessing method.

Figure 3. Examples of the images from the Gardens Point Walking dataset under challenging light conditions.

3.2 Structure and backbone

The structure of our network is shown in Figure 1. In the encoder of Deeplabv3+ structure, the feature map of the input images is extracted by lightweight encoder, then goes through the ASPP to obtain multi-scale features. In the decoder, low-level feature is combined with output of encoder together, then restored to original size by an upsampling layer.

ASPP is used to obtain more levels of features during the pooling. It is consisted of several atrous convolutions[9] with different sampling rates. Atrous convolution with lager rate refers to larger receptive field, which is associated with the high-level information, while atrous convolution with smaller rate refers to smaller receptive field and low-level information. The combination of multi-scale features offers more information and improves the accuracy. About the trade-off between speed and accuracy, in [9], cooperated with Xception as encoder, four ASPP layers are applied to achieve a high accuracy on Cityscapes dataset. But when the task needs no so much precision but a higher speed or a smaller model, it's useful to reduce the number of ASPP layers, trading a part of precision for better real-time

performance and less computation. In addition, DeepLabv3+ also designed a parameter, output stride (OS) to adjust this speed-precision balance.

3.3 Training dataset

For the purpose of aiding the VIP, both indoor and outdoor scenes need to be semantically understood in a unified way. So ADE20K dataset[30] is an appropriate dataset, which contains 20210 training images and 2000 validation images obtained from indoors and outdoors.

ADE20K dataset includes 150 classes of objects such as sky and building. However, it's not necessary to segment all these classes for the VIP to provide feedback regarding the most important semantics. They only have to care about the classes which are useful or which are potential obstacles. In addition, it's more difficult to segment so many classes in a relatively high accuracy. Therefore, we reduce the category of the dataset by combining some of them together (e.g., vehicle and car can be combined together), and adjusting the dataset into 20 classes. Table 1 shows the reclassification of our dataset. The bold categories are combined by multiple categories.

Table 1. The reclassification of ADE20K dataset.

Number	Category	Number	Category
1	**wall**	11	chair
2	building	**12**	**bus**
3	**sky**	**13**	**person**
4	floor	14	vehicles
5	tree	**15**	**door**
6	**ceiling**	16	table
7	pavement	17	picture
8	stair	18	obstacles
9	**window**	19	furniture
10	water	**20**	**others**

3.4 Training details

We train the network on the adjusted ADE20K dataset. Before training, the weights of ShuffleNet V2 are initially loaded from the model pre-trained on ImageNet[31], making the encoder have the basic feature-extracting ability of many kinds of objects. Then we train the network on original ADE20K dataset to learn the feature of typical indoor and outdoor object classes. Finally, we fine-tune the network on the 20-class ADE20K dataset in order to learn more information under the VIP assistance circumstances. One layer of ASPP is added into the encoder to increase the precision.

DeepLab model zoo provides the pre-trained model with MobileNetV2 as encoder, pre-trained on ImageNet and ADE20K. The model's size is 24.8M and the mIoU (mean Intersection over Union) reaches 32.04% on the validation set of ADE20K. Obviously, it's far from enough for practical application in VIP assistance. Therefore, we train this model for more steps and improve its mIoU to make a comparison with ShuffleNet V2 encoder model.

The learning rate strategy is set up as 'poly', which is shown in the formula below:

$$lr = base\_lr \times (1 - global\_step / training\ number\ of\ step) \wedge learning\_power \tag{5}$$

4. EXPERIMENTS AND RESULTS

Experiments have been completed to prove the effectiveness of preprocessing method on our architecture. We train the DeepLab V3+ model with lightweight backbones, using datasets with preprocessing and without preprocessing.

4.1 Illumination invariance preprocessing

For the purpose of improving robustness, we preprocess images in 20-class ADE20K dataset with the preprocessing method in Section 3.1, and use this dataset to train the network with MobileNetV2 backbone and ShuffleNet V2 backbone. The model is pre-trained on ImageNet and 150-class ADE20K to learn the basic classification ability. The

training iteration is 100k, with a learning rate of 1e-4. For comparison, training on 20-class ADE20K without preprocessing is also carried out with the same training settings. Evaluations on the preprocessed ADE20K validation set are presented to investigate the effect of preprocessing. Results are shown in Figure 4.

On the whole, the mIoU of model trained on illumination invariance preprocessed dataset is not higher than that of model without preprocessing. However, observing the details in images of validation dataset, we still find that illumination invariance preprocessing has some positive impacts on the segmentation. As shown in Figure 4, the image in the first row has 2 glass doors in poor lighting scene, which is hard to be recognized by ordinary model, can be segmented using the model trained with preprocessing. The second row shows that details such as small trees are able to be segmented with the aid of preprocessing. Preprocessing can also convert unfamiliar categories into those similar to them, such as converting snowfield into ground, instead of being considered as background. These details count a lot in various different complicated scenes met in realistic task, challenging the robustness of semantic segmentation in terms of light, scale and unfamiliar classes. So illumination invariance preprocessing has obvious positive effects on semantic segmentation and improves the robustness.

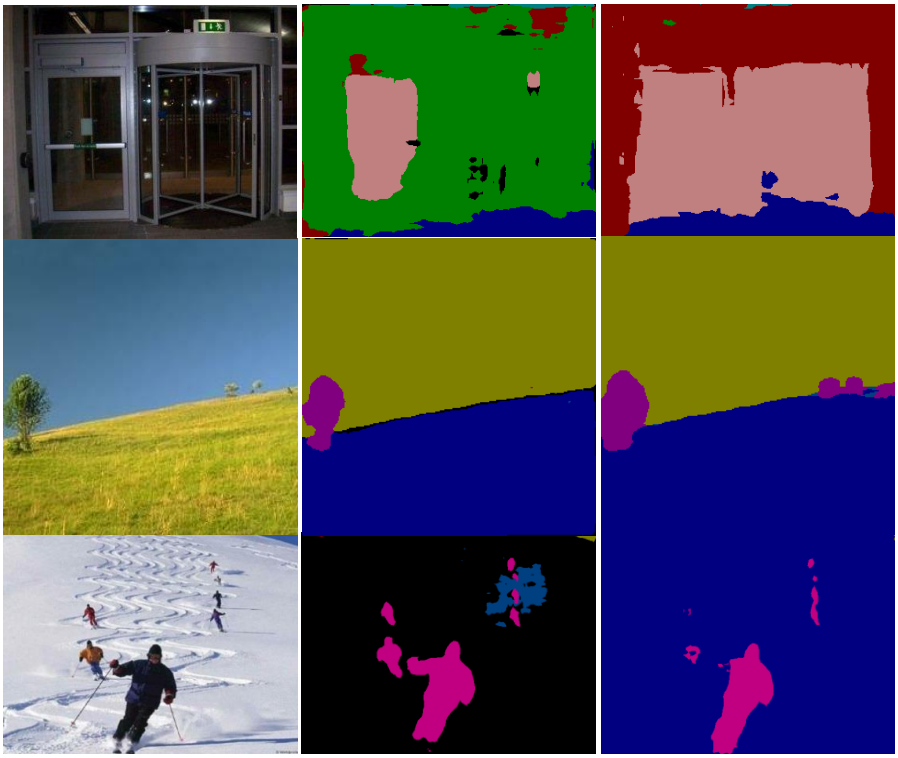

Figure 4. Semantic segmentation results of ADE20K images by model with MobileNetV2 as backbone.
The left column shows the input RGB images; the middle column shows the outputs segmented by model trained without preprocessing; the right column shows the outputs segmented by model trained with preprocessing.

To examine the performance on practicing situations, we test the preprocessing method on Gardens Point Walking dataset. Images in this dataset are segmented respectively by models trained on the preprocessed dataset and the un-preprocessed dataset. Both models with ShuffleNet V2 backbone and MobileNetV2 backbone are tested.

The segmentation results are shown in Figure 5 - Figure 7. Gardens Point Walking datasets contains both day and night images. It's clear to see that when the light condition in the different parts of the image differs a lot, it's hard for original models to recognize the objects in the dark parts due to the less illumination. Instead, the preprocessed models can segment objects in dark parts and provide more information and details of the image, because the preprocessing method reduces the influence of light and reflectance and shows the information about the material of objects. The glass doors in dark parts in Figure 5 are clearly segmented by the preprocessed model with MobileNetV2 as the backbone.

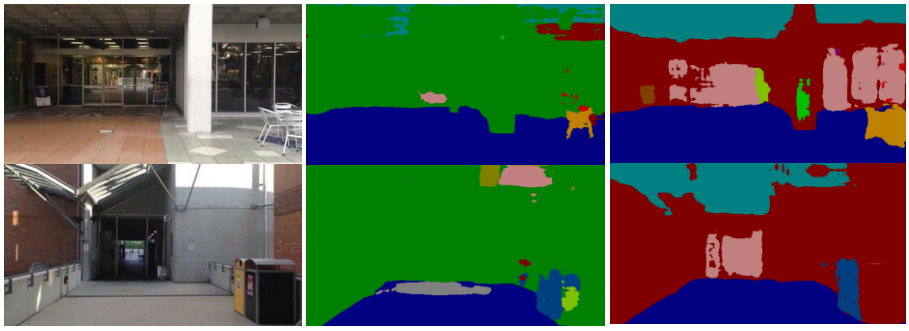

Figure 5. Semantic segmentation results of day images in Gardens Point dataset by model with MobileNetV2 backbone. The left column shows the input RGB images; the middle column shows the outputs segmented by model trained without preprocessing; the right column shows the outputs segmented by model trained with preprocessing.

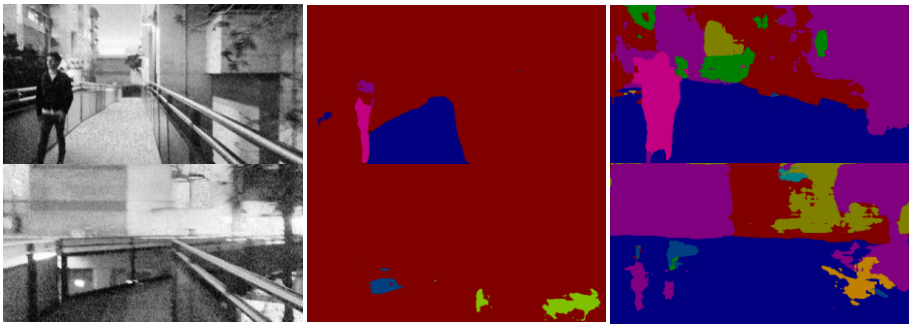

Figure 6. Semantic segmentation results of night images in Gardens Point dataset by model with MobileNetV2 backbone. The left column shows the input RGB images; the middle column shows the outputs segmented by model trained without preprocessing; the right column shows the outputs segmented by model trained with preprocessing.

For night situation, it's extremely hard for most of the original models to do semantic segmentation successfully because of the challenging light condition. Some of them are difficult to be recognized even for human eyes. Illumination invariance preprocessing can improve the result to a certain extent, providing as much information as possible from the night images. In Figure 6, models without preprocessing training almost segment nothing or only segment few object, while the model with preprocessing training segments more categories such as tree and sky, although not accurate enough. It shows the robustness improvement in night scenes brought by the illumination invariance preprocessing.

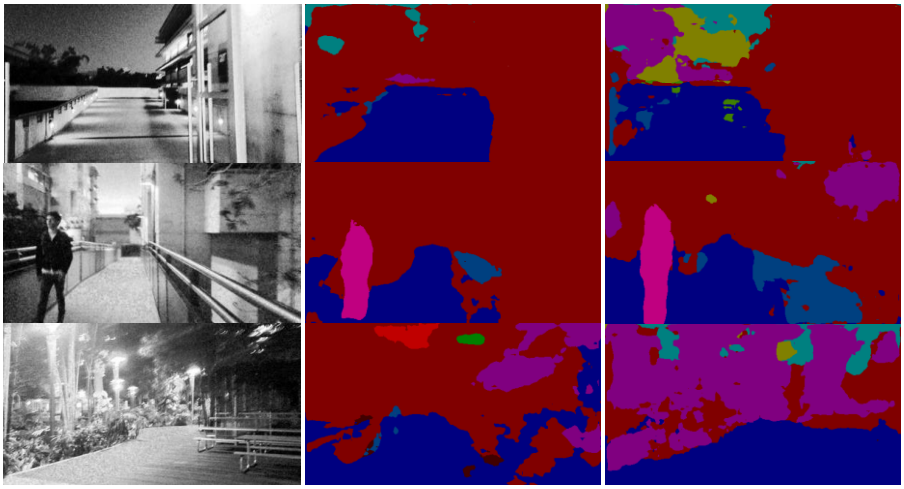

Figure 7. Semantic segmentation results of night images in Gardens Point dataset by model with ShuffleNet V2 backbone. The left column shows the input RGB images; the middle column shows the outputs segmented by model trained without preprocessing; the right column shows the outputs segmented by model trained with preprocessing.

Although illumination invariance preprocessing method can offer more information about objects in many scenes, there are also several cases where it cannot perform well and reduces the accuracy of the segmentation.

4.2 Backbone of model for mobile devices

During the training of DeepLabv3+ model with different backbone, we compare the two kinds of backbones, MobileNetV2 and ShuffleNet V2 to find the proper backbone for on lightweight semantic segmentation task on mobile devices.

The model with MobileNetV2 backbone is without ASPP layer, provided and pre-trained on ImageNet and 150-class ADE20K by [9]. The model with ShuffleNet V2 backbone is with one layer of ASPP [9] and atrous convolution mentioned in [32], pre-trained on ImageNet by [33]. Adding one layer of ASPP is aimed to improve the accuracy without increasing too much computation. We pre-train the model with ShuffleNet V2 on the 150-class ADE20K dataset. Then we train both models on the adapted 20-class ADE20K dataset for 200k iterations with an original learning rate of 1e-4.

Table 2. Test on 20-class ADE20K with the input size of 513×513.

	mIoU(%)	Size(MB)	FPS
MobileNetV2	50.9	9.4	64
ShuffleNet V2	45.8	9.0	83

Frame rate is tested on NVidia 1080Ti GPU. From Table 2, we find that both models are small and fast for running on mobile devices. But with the similar size and speed, model with MobileNetV2 has higher mIoU on the 20-class ADE20K. So it's more suitable to use MobileNetV2 as the backbone of DeepLabv3+ model for mobile devices.

By using Tensorflow Lite (TFLite), the trained model can be converted into TFLite format and put it on the mobile phone. The model is useful for both indoors and outdoors to help the VIPs' scene understanding.

5. CONCLUSION

To improve the robustness of semantic scene understanding system by CNN network, we utilize illumination invariance preprocessing method and perform the evaluation on Gardens Point Walking dataset. We utilize DeepLabv3+ structure with lightweight network as backbone to build the semantic segmentation model which can be developed on mobile device.

In the future, more complicated illumination invariance transformations can be considered to further improve the results. To improve accuracy of the model, the main task is on the training of the backbone network. For smaller model on mobile devices, model compression, network pruning and knowledge distillation can also be developed on the backbone network.

REFERENCES

[1] Bourne, Rupert RA, et al. "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis." The Lancet Global Health 5.9 (2017): e888-e897.
[2] Yang, Kailun, et al. "Expanding the detection of traversable area with RealSense for the visually impaired." Sensors 16.11 (2016): 1954.
[3] Yang, Kailun, et al. "Unifying terrain awareness through real-time semantic segmentation." 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018.
[4] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." arXiv preprint arXiv:1606.02147 (2016).
[5] Romera, E., et al. "Bridging the day and night domain gap for semantic segmentation." 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE. 2019.
[6] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[7] Yang, Kailun, et al. "Robustifying semantic cognition of traversability across wearable RGB-depth cameras." Applied optics 58.12 (2019): 3141-3155.
[8] Yang, Kailun, et al. "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception." 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE. 2019.
[9] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.
[10] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).
[11] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[12] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[13] Barrow, Harry, et al. "Recovering intrinsic scene characteristics." Comput. Vis. Syst 2.3-26 (1978): 2.
[14] Finlayson, Graham D., Mark S. Drew, and Cheng Lu. "Entropy minimization for shadow removal." International Journal of Computer Vision 85.1 (2009): 35-57.
[15] Maddern, Will, et al. "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles." Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China. Vol. 2. 2014.
[16] Alvarez, José M. Álvarez, and Antonio M. Lopez. "Road detection based on illuminant invariance." IEEE Transactions on Intelligent Transportation Systems 12.1 (2010): 184-193.
[17] Krajník, Tomáš, Jan Blažíček, and Joao M. Santos. "Visual road following using intrinsic images." 2015 European Conference on Mobile Robots (ECMR). IEEE, 2015.
[18] Alshammari, Naif, Samet Akcay, and Toby P. Breckon. "On the Impact of Illumination-Invariant Image Pre-transformation for Contemporary Automotive Semantic Scene Understanding." 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018.
[19] Wei, Chen, et al. "Deep retinex decomposition for low-light enhancement." arXiv preprint arXiv:1808.04560 (2018).
[20] Baslamisli, Anil S., et al. "Joint learning of intrinsic images and semantic segmentation." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[21] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[22] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
[23] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.
[24] Yang, Kailun, et al. "Unifying terrain awareness for the visually impaired through real-time semantic segmentation." Sensors18.5 (2018): 1506.
[25] Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." arXiv preprint arXiv:1412.7062 (2014).
[26] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).
[27] Zhang, Xiangyu, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[28] ai.googleblog.com/2018/03/mobile-real-time-video-segmentation.html
[29] Kim, Wonsuk, and Junhee Seok. "Indoor Semantic Segmentation for Robot Navigating on Mobile." 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, 2018.
[30] Zhou, Bolei, et al. "Scene parsing through ade20k dataset." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[31] Deng, Jia, et al. "ImageNet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
[32] Türkmen, Sercan, and Janne Heikkilä. "An efficient solution for semantic segmentation: ShuffleNet V2 with atrous separable convolutions." Scandinavian Conference on Image Analysis. Springer, Cham, 2019.
[33] TropComplique, 16 Sep 2018, https://github.com/TropComplique/shufflenet-v2-tensorflow

[34] https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets