Hybrid Feature based Pyramid Network for Nighttime Semantic Segmentation

Yuqi Li, Yinan Ma, Jing Wu and Chengnian Long*
Department of Automation, Shanghai Jiao Tong University, Shanghai, China

Keywords: Semantic Segmentation, Nighttime Dataset, Exposure-aware Network.

Abstract: In recent years, considerable progress has been made on semantic segmentation tasks. However, most existing works focus on only day-time images under favorable illumination conditions. In this work, we aim at nighttime semantic segmentation, which is remaining to be solved due to the problems of over- and under-exposures caused by complex lighting conditions and the lack of trainable nighttime dataset as pixel-level annotation requires extensive time and human effort. We (1) propose a hybrid network combining image pyramid network and Gray Level Co-occurrence Matrix (GLCM). GLCM is a significant descriptor of texture information, as statistical features to compensate the missing texture information in the over- and under-exposures problem at night. (2) design an exposure-awareness encoder network by fusing hybrid features hierarchically in GLCM fusion layers. (3) elaborately generate a trainable nighttime dataset, Carla-based Synthesis Nighttime dataset (CSN dataset), with 10027 synthesis images to resolve the problem of large-scale human annotations. To check whether the network trained on synthesized images is effective in the real world we also collect a real-world dataset called NightCampus with 500 nighttime images with annotations used as test dataset. We prove that our network trained on synthetic dataset yielding top performances on our real-world dataset.

1 INTRODUCTION

Semantic segmentation is to assign a specific class or category label to each pixel in the images according to its semantic meanings. Credit to deep Convolutional Neural Networks (CNNs) first introduced in this field by FCN (Long et al., 2015), many semantic segmentation methods have achieved real-time performance without sacrificing too much quality (Zhao et al., 2018; Yu et al., 2018; Nekrasov et al., 2018). However, in outdoor scenes we hardly see the applications of these computer vision works, which mainly focusing on only day-time images under favorable illumination conditions, exert their efficiency and reliability, as their accuracy declines significantly under challenging lighting conditions like nighttime.

Most of the current night semantic segmentation methods tend to use Far-Infrared (FIR) camera instead of visible light camera. However, FIR cameras are expensive and infrared images lack color information and texture information. Their contrast and signal-to-noise ratio are relatively low. So the methods based on infrared images is not suitable for semantic segmentation with high resolution requirements. In this work, we are focusing on nighttime semantic segmentation with RGB images.

Figure 1: Examples in NightCampus dataset. Nighttime images have both over-exposure problems and under-exposure problems. The texture and color information of vegetation almost disappear because there is no light source, while the street lights and headlight of cars area also loss texture information.

There are two main challenges to the problem of semantic segmentation of RGB images at night:

First, the reason why semantic segmentation using visible light cameras do not perform well at nighttime scenes is that extremely weak illuminance will degrade the structure, texture and color characteristics of input images. Nighttime images have both over-exposure problems and under-exposure prob-

Li, Y., Ma, Y., Wu, J. and Long, C.
Hybrid Feature based Pyramid Network for Nighttime Semantic Segmentation.
DOI: 10.5220/0010248503210328
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 4: VISAPP*, pages 321-328
ISBN: 978-989-758-488-6

lems. For example, in Figure. 1, the texture and color information of vegetation almost disappear because there is no light source, seems like an empty black region. The same situation happens in the street lights and headlight of cars area, but looks like an empty white region. Therefore, how to solve the problem of those information loss caused by over-exposure and under-exposure is the primary issue of semantic segmentation in nighttime scene.

Second, there are no large-scale labeled datasets available for nighttime scenes. Existing large data sets for semantic segmentation mainly contain daytime images, with little or no nighttime images. While the training of deep neural network is dependent on the input data. Therefore, the lack of trainable datasets in this field has impeded the development of nighttime semantic segmentation.

Aimed on the above challenges, we proposed a hybrid network combining image pyramid network and Gray Level Co-occurrence Matrix (GLCM). GLCM is a significant descriptor of texture information, as statistical features to compensate the missing information in the over- and under-exposures convolution and decline overfitting problems. Texture is an inherent global feature of an image and almost cannot be affected by noise. So its an important information used in the segmentation task, especially in nighttime tasks where color information is hardly exist. The Pyramid structure allow the multi-resolution input of original images and thus enable the network consider both global information and local detailed features, learning the weight between global and local information. We design an exposure-awareness encoder network by fusing hybrid features in GLCM fusion layers in a single hierarchy. The GLCM fusion layer is introduced to guide the network to learn where is the over-and under-exposure region in order to supplement GLCM features.

Besides, to tackle the lack of trainable nighttime dataset, and to solve the problem of extensive time and human effort wasted when annotating large trainable nighttime dataset, we elaborately generate 10072 synthesis campus-like nighttime images called CSN dataset based on Carla simulator instead. The CSN dataset is pixel-wise annotated by computer and has no annotation errors undoubtedly. To check whether the network trained on synthesized images is effective in the real world, we collect a real-world dataset called NightCampus with 500 real-world nighttime images with pixel-level label annotations used as test dataset.

We prove that our Hybrid Feature based Pyramid Network (HFPNet) trained on synthetic dataset yielding top performances on our real-world dataset. As shown in our experiments, our work slightly superior to other real-time algorithms in daytime as we got 73.9% mIoU at 32 FPS on CityScapes test dataset, and substantially improved performance on nighttime scenes as we obtained nearly 5% mIoU improvement comparing to the other dattime methods on CSN validation dataset and 88.3% mIoU on NightCampus test dataset.

2 RELATED WORKS

Infrared Information based Fusion Methods for Nighttime. Most of the current night semantic segmentation algorithms use thermal imaging images based on infrared cameras. (Zhiyi Liu, 2020) proposed a semantic segmentation algorithm for unmanned vehicle night infrared images based on the improved DeepLabv3+ network. A monocular vision system (Ge et al., 2009) is proposed for real-time pedestrian detection and tracking using near infrared (NIR) cameras while driving at night. Xu F. et al. (Xu and Fujimura, 2002) proposed a method for pedestrian detection and tracking using a single night vision camera mounted on a vehicle. (Wang,) collected infrared scene image data sets, combined with the ideas of InceptionNet and ResNet, proposed a 100-layer parallel residual network structure PresNet-100, and constructed a multi-scale semantic segmentation network Multi-PresNet based on PresNet.

However, infrared images contrast and signal-to-noise ratio are relatively low. Most of them it needs to be aligned and merged with RGB images, the operation of which is more complicated. Also data they used are usually in subdivided scenes. So the methods based on infrared images is not suitable for semantic segmentation with high resolution requirements.

GAN based Day-night Conversion Methods for Nighttime. Although most existing works focus on the standard daytime conditions under favorable illumination, there are also some works that address the challenging scenarios. Christos Sakaridis1 et al. (Sakaridis et al., 2019) designed a Curriculum framework by adapting daytime models to nighttime scenes without using nighttime scene annotations by GAN. They also collected the dark Zurich dataset, which included 151 pixel-level annotated night image as test dataset (dataset keeps private). Lei Sun et al.(Sun et al., 2019) trained a day-night conversion network by CycleGAN. CycleGANs are used to transfer nighttime images to the daytime and different ratio of daytime images in the dataset to the nighttime while keeping their label.

They require pixel-level matches, images corre-

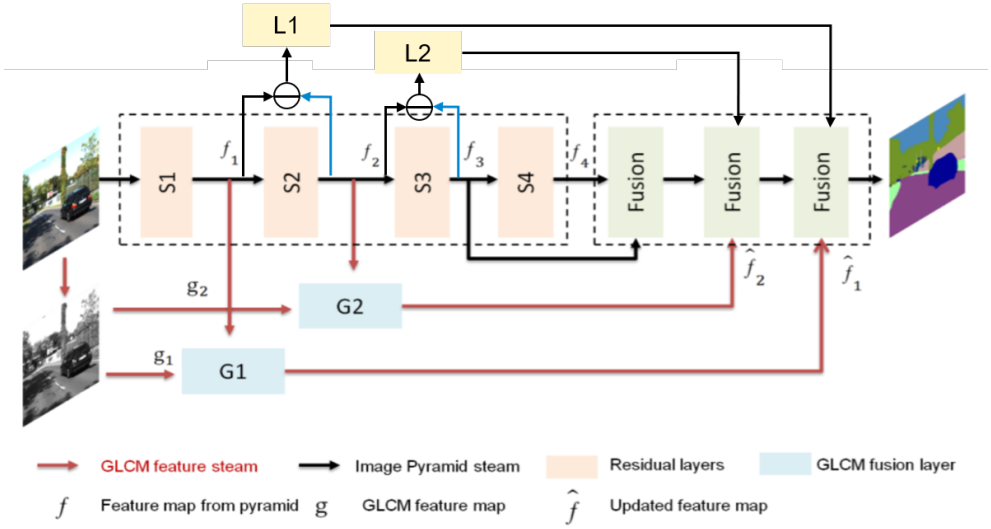

Figure 2: Our Hybrid Feature based Pyramid Network pipline. Given RGB images, ResNet extracts 4 resolution feature maps which form an image pyramid and corresponding Laplacian pyramid in the shallowest two layers. GLCM feature maps calculated by gray level input images are used as a mask of the fused feature maps in GLCM fusion layer. L1,L2 are Laplacian pyramid feature maps, blue line represent up-sampling.

sponding to the same position and the same angle during the daytime and other periods, which is harsh for dataset collecting. And their networks are mainly used to generate datasets, does not directly solve the problem of information loss caused by over- and under-exposure at nighttime.

Image Pyramid Structure. FPNNet (Lin et al., 2017) fuse the low-resolution feature maps with strong semantic information and the high-resolution feature maps with weak semantic information but rich spatial information under the premise of increasing less computational complexity. Oršić et al.(Oršić and Segvic, 2020) proposed a approach based on shared pyramidal representation and fusion of heterogeneous features along the upsampling path. The proposed pyramidal fusion approach is especially effective for dense inference in images with large scale variance due to strong regularization effects induced by feature sharing across the resolution pyramid.

3 HYBRID FEATURE BASED PYRAMID NETWORK

We propose a hybrid network (HFPNet), which combines an image pyramid network and GLCM. The pipeline of HFPNet is shown in Fig. 2. The well-designed exposure-awareness decoder network in HFPNet consists of GLCM fusion layers and Fusion Blocks hierarchically to force encoder network learn effective features in image pyramid.

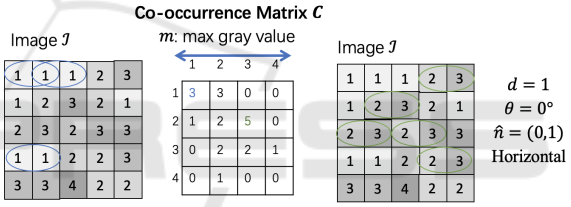

Figure 3: The calculation process of co-occurrence matrix.

3.1 Gray Level Co-occurrence Matrix Feature Map

The gray-level co-occurrence matrix is based on studies of the statistics of pixel intensity distributions. It is a matrix describing the gray-scale relationship between a pixel in a local or whole area of an image and neighboring pixels or pixels within a certain distance d. The co-occurrence matrices provide raw numerical data on the texture, however in practical applications, from the perspective of the calculation efficiency of texture features and the storage of the GLCM matrix, the gray level of the original image is usually first compressed. For example, from 8-bit images with gray levels of $0-255$ to 5-bit images with gray levels of 0-31, the dimension of the corresponding co-occurrence matrix is reduced from 256×256 to 32×32.

Calculation of GLCM. During the graying process, we use the formula

$$Gray = 0.114B + 0.587G + 0.299R, \qquad (1)$$

where R,G,B are the value of three channel of RGB

image. In order to construct a co-occurrence matrix, it is necessary to consider the fixed distance of all pixel pairs from each other, without considering the relative direction formed by the line connecting them and the image reference direction, which means that the only parameter is the distance d, the system can have any number of optional matrices:

$$\mathcal{C}(k,l;d) = \sum_i \sum_j \delta(k - g(i,j))\delta(l - (g((i,j) + d\hat{n})), \tag{2}$$

where $\hat{n}$ is the unit vector pointing in a chosen direction, $g(i,j)$ is the gray value of pixel (i,j), $(g((i,j) + d\hat{n}))$ is the gray value of another pixel that is at distance from pixel (i,j) and at the orientation defined by unit vector $\hat{n}$ (e.g. assume $\hat{n} = (0,1)$, representing the horizontal direction, thus $g((i,j) + d\hat{n}) = g(i,j+1)$, and $\mathcal{C}(k,l;d)$ is the total number of pairs of pixels at distance d from each other identified in the image, such that the first one has gray value k and the second has gray value l, as shown in Figure. 3. $\delta(x)$ represents the Dirac delta function, it equal to 1 if $x = 0$ and results in 0 when $x \neq 0$.

GLCM Feature. The 9 statistical attributes of texture features calculated form GLCM matrix used in this paper are Mean, Variance, Std, Homogeneity, Contrast, Dissimilarity, Entropy, Angular Second Moment, Correlation. Thus each pixel is associated to 9 statistical features obtained from GLCM, forming a $9 \times s \times s$ global feature map, where s is the size of original image. During the graying process, if the gray value of the target point is directly divided, it will cause the image sharpness to be reduced. Therefore, we first convert the picture to perform histogram equalization, so as to increase the dynamic range of the gray value, which increases the overall contrast effect of the image. Different textures shows distinguishing features, thus proving the GLCM features are applicable descriptor for textures.

3.2 Image Pyramid Structured Network

In order to enhance multi-resolution information and employ global and local features at the same time, we used a pyramid-structured CNN in our work. Image pyramid (Adelson et al., 1984) is a powerful but conceptually simple structure that can interpret images at multiple resolutions. Originally designed for machine vision and image compression applications. By multi-sampling the original image, images of different resolutions can be generated as a sequence.

Place the highest resolution images at the bottom and arrange them in a pyramid shape. Since the base level M is size $2^M \times 2^M$ or $N \times N$, where $J = log_2 N$, the intermediate level m is size $2^m \times 2^m$, where $0 \leq m \leq M$. Fully populated pyramids are composed of $M+1$ resolution levels from $2^M \times 2^M$ to $2^0 \times 2^0$.

That is, in general, the general limitation of P will reduce the resolution approximation of the original image; for example, the single-pixel approximation of the 1×1 or 512×512 image has little value. The total number of elements in a $P+1$ level pyramid for $P > 0$ is

$$N^2(1 + \frac{1}{(4)^1} + \frac{1}{(4)^2} + \cdots + \frac{1}{(4)^N}) \leq \frac{4}{3}N^2. \tag{3}$$

The Laplacian pyramid (Burt and Adelson, 1983) retains the blurred version of the difference image between each level. Only the minimum level is not a differential image, so that a higher-level differential image can be used to reconstruct a high-resolution image.

$$\mathcal{L}_j(x,y) = \mathcal{G}_j(x,y) - EXPAND(\mathcal{G}_{j+1}(x,y)), \tag{4}$$

where $\mathcal{L}_j(x,y)$ denotes the j-th level of Laplacian pyramids and $\mathcal{G}_j(x,y)$ and $\mathcal{G}_{j+1}(x,y)$ denotes the j-th level of Gaussian pyramid, the $EXPAND$ is the up-sampling operation.

3.3 Hybrid Feature based Network Structure

Backbone. We use the ResNet-101 (S1 to S4 in Figure. 2) pretrained over the ImageNet(Krizhevsky et al., 2012) dataset as the backbone to get 4 , 8 , 16 and 32 down-sampled feature maps.

GLCM Feature Map. As introduced in Section 3.1, GLCM features form a $9 \times s \times s$ feature map, and this global feature map is then sent into a 1×1 convolutional layer to expand channels as the same as CNN feature maps, thus we got final GLCM feature map g. In other words, GLCM Feature Maps g are extracted by compressing the input images into different-bits grayscale images.

GLCM Fusion Layer. To learn where are the texture-lost regions, we introduce the GLCM fusion layer G to augment GLCM features which contain compensated information from nearby pixels. This layer forces our model to learn the effective combination of GLCM features that help predict the correct label and guide the segmentation task towards an optimal performance. Semantic segmentation requires both large region context information and rich spatial information. The former is mostly obtained through deep hierarchies. For a single layer level, after we obtain fused feature from image pyramid feature maps and

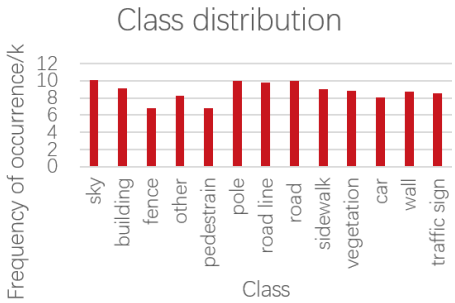

Figure 4: Class distribution in CSN dataset.

Laplacian pyramid feature maps mentioned above, this feature map and GLCM feature map g are sent into 1×1 convolutional layers separately followed by a sum operation.

Image Pyramid Structure in Encoder. The Pyramid structure enable a network consider both global information and local detailed features, which respectively corresponds to context information and spatial information for semantic segmentation.

Fusion Block. Fusion blocks are used in each hierarchy to fuse high-resolution feature map f_{i+1} obtained from image pyramid network and the feature map $\hat{f}_i$ obtained from GLCM Fusion Layer G. Inside the fusion block, each path is convolved with 3×3 convolution to determine the weight between two sets of feature maps and then upsampled to the same resolution as f_{i+1}. Two paths are summed up, and analagoulosly further propagated through next fusion blocks until the desired resolution is reached.

4 NIGHTCAMPUS AND CSN DATASETS

As annotating large trainable nighttime dataset requires extensive time and human effort, we elaborately generate 10072 synthesis campus-like nighttime images based on Carla instead. Carla is an open source simulator that can simulate real-world road scenes. The Carla dataset is pixel-wise annotated by

Table 1: Distribution on different maps.

Geographical environment	Maps	vehicles	pedestrian
Urban	town01	[0,50,100]	[0,100,200]
Urban	town02	[0,50,100]	[0,100,200]
Half-Urban Half-Suburbs	town03	[0,80,160]	[0,160,320]
Half-Urban Half-Suburbs	town05	[0,80,160]	[0,160,320]
Suburbs	town04	[0,100,200]	[0,200,400]

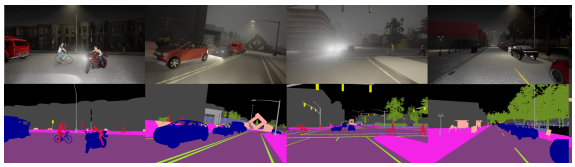

Figure 5: Sample images in CSN and corresponding ground-truth label maps.

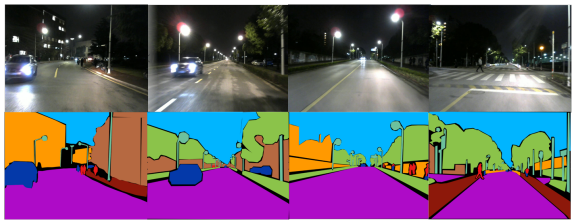

Figure 6: Sample images in NightCampus and corresponding ground-truth label maps.

computer and has no annotation errors undoubtedly. To check whether the network trained on synthesized images is effective in the real world, we collect a real-world dataset called NightCampus with 500 real-world nighttime images with pixel-level label annotations used as test dataset. We prove that the network trained on synthetic dataset is able to perform well in real-world scenes.

4.1 The CSN Dataset and Label Annotation

The Carla simulator is in a town environment, with roads, buildings, streets, traffic lights and so on. However, pedestrians or vehicles are added by running python script. We generate 10072 images which are similar to real scenes lighting effect, 6000 of which are used for training, 4072 of which are used for validation. Size of those images is 1280×720. We set different densities of pedestrians and vehicles, diversity of geographical environment, as listed in Table. 1. The class information is same as Cityscapes dataset, as shown in Figure. 4,the class distribution in each image is quite balanced. Figure. 5 shows sample images and corresponding ground-truth label maps.

4.2 The NightCampus Dataset

Images in our dataset are from real world nighttime road scene captured in a campus. These images are captured by Logitech c270 size of images is 1280×960. Compared to urban road scenes, NightCampus has more pedestrians, vegetation, and fewer vehicles, and darker scenes which means the contrast of images is greater and the problem of over-exposure and under-exposure is more serious. And in campus

scenes there are fewer types of vehicles, for example, there is no train in campus, truck and bus seldom appear, so we merge all types of cars into one 'vehicle' class. Similarly, terrain is in no need. Finally we get 9 classes in total, that is 'vehicle' 'road' 'sidewalk' 'sky' 'pole' 'vegetation' 'person' 'traffic sign'. Some regions that are too difficult to define even by humans are labeled as 'uncertain' so that they are ignored during training and evaluation. Figure. 6 shows sample images and corresponding ground-truth label maps.

5 EXPERIMENTS

5.1 Implementation Details

All experiments are on a workstation with Tesla V100 GPU under CUDA 9.0 and CuDNN 7.0. The experiments are conducted based on pyTorch. The network use mini-batch stochastic gradient descent (SGD) (Krizhevsky et al., 2012) with batch size 20, momentum 0.9 and weight decay $1e^{-5}$ in training. The initial learning rate is different with encoder and decoder, for encoder is $5e^{-4}$ and for decoder is $5e^{-3}$. Training is divided into three stages with different learning rate, at the beginning of each training stage, initial learning rate is reduced by half.

5.2 D Parameter Selection in GLCM

We tried different distance parameter $d = 1,2,3,4$ settings, to find the best performance for GLCM feature extraction on our CSN dataset. The experiment result is shown in Table 2. We use mean Intersection over Union (mIoU) to evaluate the semantic segmentation performance in our experiments.

When $d = 1,2$ pixel of interest (POI) cannot get enough texture information from nearby pixels, and when $d = 4$, POI may learn feature which is not from instrumental classes. The following experiments were all taken under $d = 3$.

5.3 Ablation Study

We have tried to add GLCM guidence layers to several different layers combinations. We number the layer level from the shallowest to deepest from 1 to 4. Experiments on CSN vilidation set are shown in Table 3. For single layer cases, the deeper layer did not achieve good results because for global features, context information is more important. The following experiments were all taken under $1+2$ situation.

Table 2: Experiments on CSN Validation Dataset with Different d.

GLCM para	$d=1$	$d=2$	$d=3$	$d=4$
mIoU(%)	92.1	93.4	94.2	92.8

And to verify the effectiveness of the pyramid structure, we conduct experiments on CSN validation dataset. when GLCM are excluded, performance drops significantly to 87.8% mIoU, which confirms the importance of GLCM focusing on the missing texture information. And pyramid structure also can help learn useful features at both w&w/o GLCM situation.

5.4 Experiment on Nighttime Datasets

We trained HFPNet on the training set of CSN, and compare it with other real-time networks on the CSN validation set and NightCampus dataset.

Experiment on CSN and NightCampus. As shown in Table 4, our work substantially improved performance on nighttime scenes as we obtained nearly 5% mIoU improvement comparing to the other dattime methods on CSN validation dataset and 88.3% mIoU on NightCampus test dataset. Due to the difference between synthetic and real pictures, the performance is reduced in NightCampus dataset, but not much severe. We prove that our HFPNet trained on synthetic dataset is able to perform well on real-world scenes.

The visualization examples are shown in Figure. 7. Compared to BiseNet, our HFPNet can segment over- and under-exposure regions more credibly. Particularly, our model can produce more accurate and clear boundaries in the segmentation of buildings and vegetation. In addition, it makes the sidewalk clearer and more complete. BiseNet ignores some poles, but our street lights have been successfully identified in those cases.

Experiment on WildDash 2. WildDash 2 has 325 nighttime images, we test with WildDash 2 using HFPNet pretrained on Carla and got 56.7% mIoU, performance declined much than on NightCampus because of severe motion blur and mirror reflection. Notice that classes of datasets are different so we set those not corresponded as ignore label thus we do not compare with others whose classes are inconsistent with ours.

5.5 Experiment on CityScapes Dataset

CityScapes (Cordts et al., 2016) is a urban street scene dataset from cars perspective, including 5000 high-resolution images as large as 1024×2048. CityScapes has 19 semantic classes not counting ignore label 255. We uploaded our semantic segmentation predictions to CityScapes organizers evaluation

Table 3: Experiments on Differenet Combination of Layers.

Layer Selection/single layer	*1*	*2*	*3*	*4*
mIoU(%) on CSN	89.2	91.5	88.2	83.8
Layer Selection/combination	*1+2*	*1+2+3*	*1+2+3+4*	
mIoU(%) on CSN	**94.7**	90.1	85.4	

Table 4: Results on CSN and NightCampus Dataset.

Method	**CSN validation set**	**NightCampus test set**
SegNet (Badrinarayanan et al., 2017)	51.3	45.6
BiSeNet (Yu et al., 2018)	89.6	84.3
ICNet (Zhao et al., 2018)	80.1	74.8
CCNet (Huang et al., 2018)	85.2	80.6
PSPNet (Zhao et al., 2017)	72.7	68.3
RefineNet-LW101 (Nekrasov et al., 2018)	85.9	82.8
HFPNet	**94.7**	**88.3**

Table 5: Results on Test Set of Cityscapes.

Method	**Backbone**	**Coarse**	**Mean IoU(%)**	**FPS**
Deeplab (Chen et al., 2017)	VGG16	×	63.1	0.25
FCN-8s (Long et al., 2015)	VGG16	×	65.3	2
Dilation10 (Yu and Koltun, 2015)	VGG16	×	67.1	0.25
LRR (Ghiasi and Fowlkes, 2016)	VGG16	×	69.7	-
LRR (Ghiasi and Fowlkes, 2016)	VGG16	√	71.8	-
ICNet* (Zhao et al., 2018)	Res101	×	69.5	30.3
GUN* (Mazzini, 2018)	DRN-D-22	×	70.4	33.3
PSPNet (Zhao et al., 2017)	Res101	√	81.2	0.78
RefineNet-LW101* (Nekrasov et al., 2018)	Res101	×	72.1	36.8
HFPNet*	Res101	×	73.9	32

*real-time semantic segmentation method.

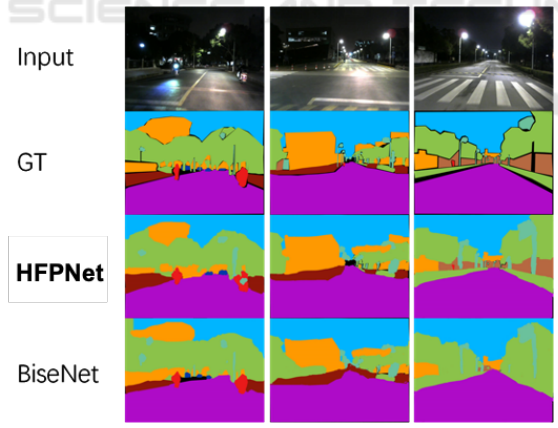

Figure 7: Visualization examples of our results on NightCampus.

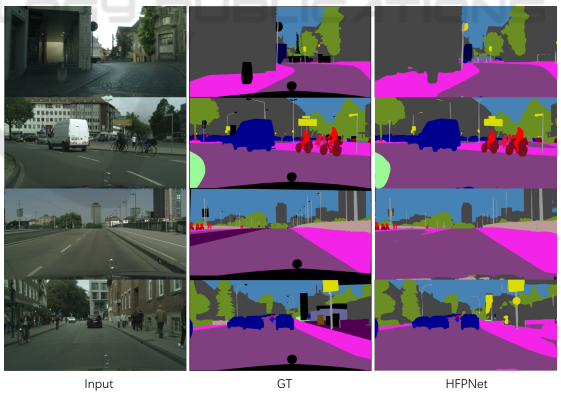

Figure 8: Visualization results on Cityscapes dataset.

server and obtained feedback scores In our experiments, we use fine annotated images only. In Table 5, Mean IoU and Frames Per Second(fps) are reported. HFPNet only slightly superior to other real-time algorithms in this daytime dataset as the visualization results shown in Figure. 8. That is because there are few overexposed and underexposed areas so the weights of GLCM features are approaching 0.

6 CONCLUSION

In this paper, we tackled the problem of nighttime semantic segmentation task, mainly focus on the missing texture information problem in over- and under-exposure region. We achieved that by introducing GLCM in our well-designed exposure-awareness de-

coder network to compensate the missing texture information. To tackled the lack of large trainable dataset in this field and to evaluate our HFPNet quantitatively, we presented a synthetic CSN dataset and a real-world NightCampus dataset. We demonstrated that HFPNet, which is trained on synthetic dataset, yielding top performances on real-world scenes.

ACKNOWLEDGEMENTS

This work was supported in part by the NSFC under Grants 62073215, 61873166, and 61673275.

REFERENCES

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., and Ogden, J. M. (1984). Pyramid methods in image processing. *RCA engineer*, 29(6):33–41.

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.

Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540.

Chen, L.-C., Papandreou, G., Kokkinos, I., 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):834–848.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Ge, J., Luo, Y., and Tei, G. (2009). Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):p.283–298.

Ghiasi, G. and Fowlkes, C. C. (2016). Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, pages 519–534. Springer.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2018). Ccnet: Criss-cross attention for semantic segmentation.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Mazzini, D. (2018). Guided upsampling network for real-time semantic segmentation. *arXiv preprint arXiv:1807.07466*.

Nekrasov, V., Shen, C., and Reid, I. (2018). Light-weight refinenet for real-time semantic segmentation. *arXiv preprint arXiv:1810.03272*.

Oršić, M. and Segvic, S. (2020). Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, page 107611.

Sakaridis, C., Dai, D., and Van Gool, L. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation.

Sun, L., Wang, K., Yang, K., and Xiang, K. (2019). See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion.

Wang, C. *Research on Infrared Image Semantic Segmentation Technology Based on Deep Learning*. PhD thesis, University of Chinese Academy of Sciences (Shanghai Institute of Technical Physics, Chinese Academy of Sciences).

Xu, F. and Fujimura, K. (2002). Pedestrian detection and tracking with night vision. In *Intelligent Vehicle Symposium, 2002. IEEE*.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhao, H., Qi, X., Shen, X., Shi, J., and Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhiyi Liu, Shaoyuan Sun, Z. R. (2020). Infrared image semantic segmentation of unmanned vehicles at night based on improved deeplabv3+. *Journal of Applied Optics*, 41(1).