See Clearer at Night: Towards Robust Nighttime Semantic Segmentation through Day-Night Image Conversion

Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, China

ABSTRACT

In recent years, intelligent driving navigation and security monitoring have made considerable progress with the help of deep Convolutional Neural Networks (CNNs). As one of the state-of-the-art perception approaches, semantic segmentation unifies distinct detection tasks widely desired by both autonomous driving and security monitoring. Currently, semantic segmentation shows remarkable efficiency and reliability in standard scenarios such as daytime scenes with favorable illumination conditions. However, in face of adverse conditions such as the nighttime, semantic segmentation loses its accuracy significantly. One of the main causes of the problem is the lack of sufficient annotated segmentation datasets of nighttime scenes. In this paper, we propose a framework to alleviate the accuracy decline when semantic segmentation is taken to adverse conditions by using Generative Adversarial Networks (GANs). To bridge the daytime and nighttime image domains, we made key observation that compared to datasets in adverse conditions, there are considerable amount of segmentation datasets in standard conditions such as BDD and our collected ZJU datasets. Our GAN-based nighttime semantic segmentation framework includes two methods. In the first method, GANs were used to translate nighttime images to the daytime, thus semantic segmentation can be performed using robust models already trained on daytime datasets. In another method, we use GANs to translate different ratio of daytime images in the dataset to the nighttime but still with their labels. In this sense, synthetic nighttime segmentation datasets can be generated to yield models prepared to operate at nighttime conditions robustly. In our experiment, the later method significantly boosts the performance at the nighttime evidenced by quantitative results using Intersection over Union (IoU) and Pixel Accuracy (Acc). We show that the performance varies with respect to the proportion of synthetic nighttime images in the dataset, where the sweet spot corresponds to most robust performance across the day and night. The proposed framework not only makes contribution to the optimization of visual perception in intelligent vehicles, but also can be applied to diverse navigational assistance systems.

Keywords: Convolutional Neural Networks, Semantic Segmentation, Generative Adversarial Networks

1. INTRODUCTION

Vision tasks like object detection and semantic segmentation (i.e. pixel-wise scene classification) are always the key points in security monitoring and autonomous driving. Convolutional Neural Networks(CNNs) have fueled the development of these methods thanks to the availability of larger datasets and computationally-powerful machines that have emerged in recent years. Semantic segmentation, which unifies distinct detection tasks with a single consumer camera[1–3] makes RADAR and LiDAR sensors become the second choices, freeing scene perception from complex multi-sensor fusion.[4–6] Some state-of-the-art CNN methods like PSPNet,[7] RefineNet,[8] DeepLab,[9] and ACNet[10] perform semantic segmentation with very high accuracies. In order to apply semantic segmentation to autonomous driving and security monitoring, we have proposed ERF-PSPNet,[1,11] a high-accuracy real-time semantic segmentation method in previous work, which is more computationally efficient than most of the state-of-the-art methods.

All these perception algorithms are designed to operate on images taken at daytime under good illumination conditions.[12–14] However, outdoor applications can hardly escape from challenging weather and illumination conditions. One of the reasons why computer vision system based on semantic segmentation have not been widely applied yet is because that it can not deal with adverse conditions. For example, semantic segmentation

Correspondence: wangkaiwei@zju.edu.cn

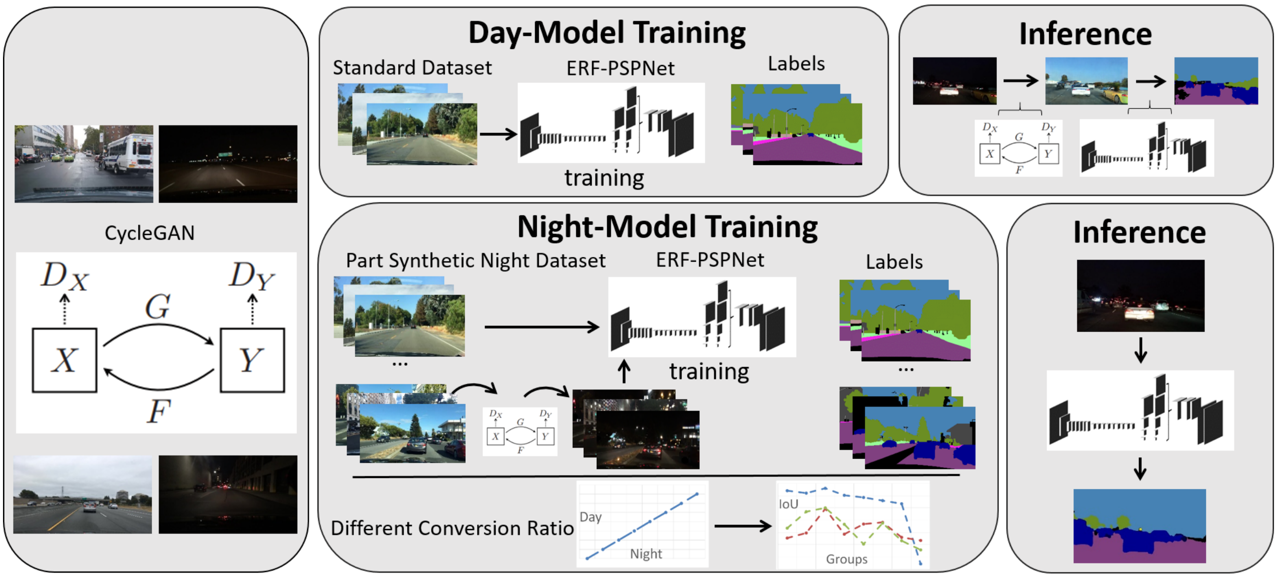


Figure 1. The figure shows the main frame of our work. On the left, a day-night converter is trained using unpaired datytime and nighttime images. The first row shows our first method: training a day-model and converting night domain images to the day domain before inference. Bottom row shows our second method: converting different ratios of images in the training set to nighttime images to train a night model. The ratio of synthetic nighttime images determines the model's accuracy in testing sets.

using visible light camera performs unsatisfactorily in the nighttime for the reason that when under extremely weak illuminance, the structure, texture and color features of objects change drastically. These features can either disappear because of the lack of the illuminance or being highly disturbed by artificial light. Thus, how to enhance the robustness of semantic segmentation has been an important issue in the computer vision domain.[15] In this work, we focus on improving nighttime semantic segmentation performance.

There are some researchers that have proposed to use Far-Infrared (FIR) camera instead of visible light camera.[16] FIR camera is a feasible measure, but they are expansive and only provide low-resolution images. In addition, there are rare FIR semantic segmentation datasets. In the other way, visible light camera is much cheaper, and there are a large number of datasets of daytime images. For these reasons, we choose visible light camera as the image acquisition device. On the other hand, high-accuracy semantic segmentation models are trained from large scale of annotated images. But annotating nighttime images requires extensive time and human effort, and it is impossible to annotate images at the pixel wise for all the other adverse conditions.

In this paper, we mainly propose a main frame (see Fig. 1) to overcome the problem of large accuracy downgrade from daytime to nighttime in semantic segmentation. Inspired by the idea of Generative Adversarial Networks (GANs),[17] nighttime images are converted on-the-fly during inference to the daytime domain as a pre-processing step of the first proposed method. In the other way, we augment an original large-scale semantic segmentation dataset such as the BDD database[18] by translating part of the daytime images to nighttime images. Among the experiments, the feasibility of improving robustness of semantic segmentation is validated. In addition, we record a dataset in the Yuquan Campus of Zhejiang university with both day and night images as well as GPS information by using our Multi-Modal Stereo Vision Sensor[19] (see Fig.2), which has been made publicly available.

2. RELATED WORKS

2.1 Semantic Understanding of the Road Scene

Semantic segmentation is important in understanding the content of images and finding target objects, and this technique is vital for the field of automatic driving.[20,21] Currently, most of state-of-the-art semantic segmentation models are based on fully convolutional end-to-end networks.[22] Inspired by SegNet,[23] semantic segmentation

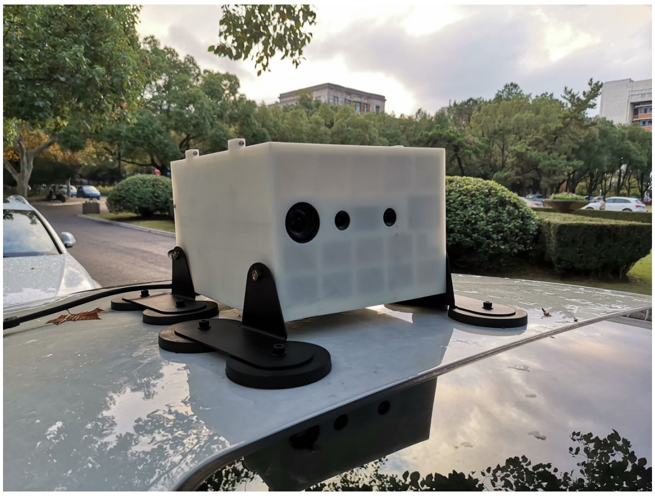


Figure 2. Our Multi-Modal Stereo Vision Sensor on the top of an instrumented vehicle used to capture the ZJU dataset.

models usually follow an encoder-decoder network architecture. The encoder is a vanilla CNN which is trained to classify the input, and the decoder is used to upsample the output of the encoder to the same size as the input image.[7,9,24–26] Further, more efficient networks were proposed to achieve the goal of real-time semantic segmentation.[15,27,28] Our works are based on ERF-PSPNet,[1,11] a state-of-the-art semantic segmentation network designed for navigation assistance systems.

2.2 Model Adaption

Generally, CNNs only learn features from the domain of training datasets, and may perform much worse in a different domain. This is also the reason why semantic segmentation model trained in daytime domain drops accuracy in the nighttime domain. To improve the generalization of convolutional neural network, many methods were proposed. Most commonly, data augmentation techniques like random cropping, random rotation and flipping are used to make networks perform stably in unfamiliar domains.[12] Effective use of synthetic data has been preliminarily studied in.[29,30] There is another domain adaptation-based approach that was proposed to adapt semantic segmentation models from synthetic images to real environments.[31] The other attempts that are most close to our work, which also improve the model robustness at the nighttime, make use of twilight images to transfer knowledge from standard daytime conditions to nighttime images.[32,33] Similar efforts were also made to address robust foggy scene parsing[34,35] and rainy scene semantic segmentation.[36,37] Unsupervised learning has also been frequently leveraged to pre-process input images, in order to prevent performance from degrading catastrophically when the input domain differs significantly from previously seen domains.[15,38] Specifically, this research line is also highly related to topological localization,[38,39] where modern visual localizers like[40,41] can also benefit from the input adaptation to perform more reliably against variation challenges. More recently, model distillation/imitation were applied to make model behave stable in unseen domains.[42,43]

2.3 Image Stylization

Since I. Goodfellow et al. proposed Generative Adversarial Networks (GANs),[17] GANs have become the most promising method for image stylization. Formally, GANs simultaneously contain two models: a generative model G that captures the critical distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than generated by G. Although state-of-the-art GANs like Pix2Pix[44] perform impressively for style transfer, the training data in both domains have to be pre-formatted into a single X/Y image pair that holds tight pixel-wise correlation. A recently proposed CycleGAN[45] is designed to perform a full translation cycle, and make it possible to make use of images in two different domains without paring images, which is suitable for our work to translate image across daytime and nighttime domains.

3. METHOD

In our work, two methods are proposed to narrow the gap between daytime and nighttime images in semantic segmentation. These methods respectively correspond to converting nighttime images to daytime images and the vice versa. Fig. 1 shows our framework. In both methods, we train a CycleGAN to perform domain converting. In the first method, we convert nighttime images to daytime, in order to shift the images to the suitable domain. Next, the ERF-PSPNet[1,11] trained on daytime images predict semantic maps in the inference. In the second method, CycleGAN converts parts of daytime images in the training set to nighttime images to extend the domain coverage of the datasets. After that, the adapted training dataset with a certain percent of nighttime images is used to train ERF-PSPNet, in order to improve its performance at the nighttime.

3.1 Training CycleGAN for Night-day Domain Converting

In this subsection, our work is training a GAN to translate nighttime images to daytime or reverse. Image-to-image translation is a class of vision problems in which the goal is to learn the mapping between input images and output images by using a training set of aligned image pairs. But for our task, large scale of paired training data is not available. Although collecting paired datasets by ourselves is theoretically possible, it is impractical to collect datasets for every different styles of scenes. What we need is an universal night-day domain converter that can be utilized at the dataset level.

CycleGAN[45] is an approach for learning to translate an image from a source domain to a target domain in the absence of paired examples, which suits our needs. CycleGAN contains two sets of GANs. Each GAN contains a generator and a discriminator. Generator and discriminator make translator, to translate image from domain X to domain Y or vice versa. Two GANs represent two generators: F and G, and they are inverses of each other. we have trained both the mapping G and F simultaneously and adding a cycle consistency loss that encourages

$$F(G(x)) \approx x \tag{1}$$

and

$$G(F(y)) \approx y \tag{2}$$

This loss makes the unpaired image-to-image translation possible.

In our work, we select 6000 daytime images and 6000 nighttime images from the BDD100K dataset,[18] as two image domains to train CycleGAN. Limited to GPU memory, we resize images to 480×270 to train our CycleGAN. In this way, we obtain our day-night and night-day converters.

3.2 Converting Images to the Daytime During Inference

The first option is to convert nighttime images to daytime images. More specifically, nighttime images acquired by the camera are converted to synthetic daytime images, which is the suitable domain for semantic segmentation.

This method does not need to train the semantic segmentation model again. In other words, the advantage of the method is that we can make use of the original weights in the trained ERF-PSPNet, which has already been demonstrated to be stable in most datasets[20] and actual scenarios.[11,12] Additionally, the night-day conversion and semantic segmentation inference are separated, which makes it easier to adjust.

But the computational cost of the inference process is increased, which is a disadvantage for real-time semantic segmentation. For each inference process, the converted images from CycleGAN are fed into ERF-PSPNet. While efficient segmentation network is readily available for predicting accurate semantic maps, the forward pass of CycleGAN costs nearly 1 second for each 480×270 image. This is too slow for a real-time semantic segmentation model like ERF-PSPNet and the system losses its real-time performance. Another disadvantage is that compared to original images, the synthetic image produced by GAN may be biased. For example, GAN may convert faraway buildings to trees etc.

3.3 Generating Nighttime Images to Expand the Training Set

The second option is to convert daytime images in the BDD10K training set with segmentation labels to nighttime images. Then, the training set with part of synthetic nighttime images is fed into ERF-PSPNet, with focal loss as the loss function.[46,47] The idea comes from the lack of the nighttime datasets with precise segmentation labels.

The advantage of the method is that for a trained model, no extra calculation is introduced in the inference process. For this reason, ERF-PSPNet can keep its real-time property. In our experiment, we conduct experiment to explore how the ratio of synthetic nighttime images influences the accuracy of the semantic segmentation model. Based on the experiment, we perform the discussion on the necessity of adding nighttime images with precise labels in the training dataset.

However, the disadvantage of the method is the time-consuming process of re-training model, and the model may not always be robust for all kinds of environments. In addition, because of the large scale of parameters in CycleGAN, we have to resize images in BDD100K to 480×270 to train the GAN. In this way, GAN can only produce images with the size of 480×270, which is below 1280×720, the resolution of images in BDD100K. So we have to upsample the synthetic images to 1280×720 before feeding into the segmentation model. Such operation makes unavoidable influence on the accuracy of the final prediction result.

4. EXPERIMENTS

4.1 Datasets and Training

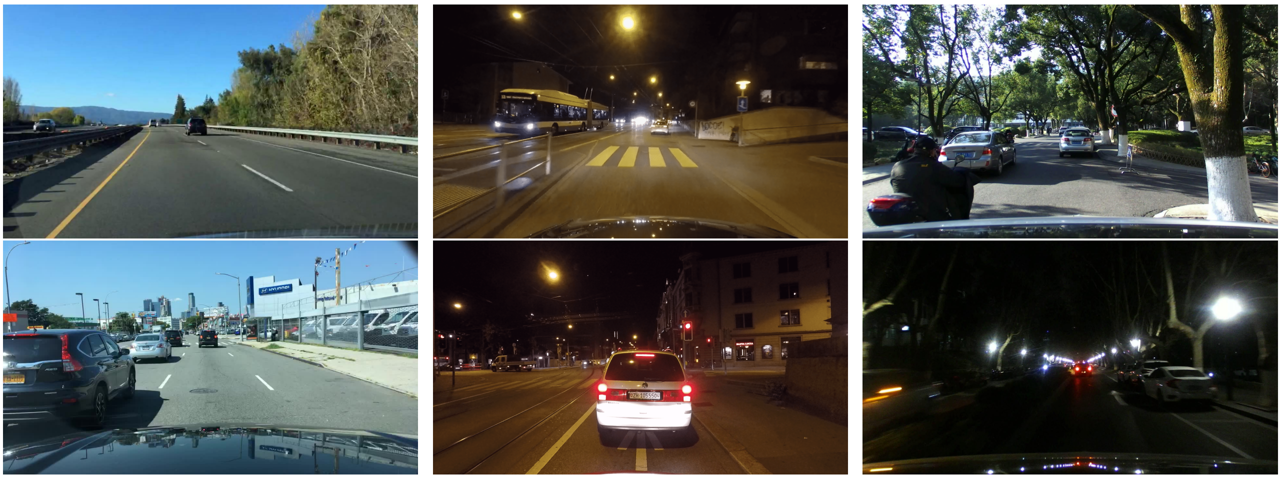
The first step of our experiment is to train a GAN converting nighttime images to daytime images and vice versa. As describe above, paired images are not necessary for CycleGAN,[45] but the datasets must contain diverse driving images at both daytime and nighttime. As nighttime images are absent in most mainstream datasets for semantic segmentation like Cityscapes[13] and Mapillary Vistas,[14] finally we utilize the lately released BDD datasets.[18]

BDD datasets contain two parts: BDD100K and BDD10K. The former contains 10,0000 driving images captured in diverse conditions and different time with GPS information, object detection annotation, lane and drivable area annotation. The latter one contains 9,000 pixel-wise semantically annotated images and 1,000 test images with 19 labeled classes.

To qualitatively verify our methods for real-world applications, we also collect daytime and nighttime images in the Yuquan Campus of Zhejiang University by using our Multi-Modal Stereo Vision Sensor[19] fixed on the top of an instrumented car. More than 1,500 images were acquired for our research. Furthermore, we utilize Nighttime Driving Test dataset provided by D. Dai et al.,[32] which contains 50 nighttime images with precise segmentation annotation. Table. 1 and Fig. 3 show details about these datasets and all three datasets are again itemized below:

BDD Dataset.[18] In total, BDD dataset have 100,000 driving images collected from more than 50,000 rides, covering New York, San Francisco Bay Area, and Berkeley. The dataset contains diverse scene scenarios such as city streets, residential areas, and highways. BDD100K contains plenty of nighttime images, making it possible to train a day-night converter. However BDD10K contains only 32 nighttime images with pixel-wise labels.

ZJU Dataset. The dataset was captured in Zhejiang University, Yuquan Campus (Hangzhou, China) with our Multi-Modal Stereo Vision Sensor.[19] During the nighttime collection, we capture the images under two settings: with headlight-illumination and without headlight-illumination. The most significant feature of this dataset is that it has more trees and pedestrians than the others, and the road is very narrow. ZJU Dataset has also been used in our preliminary work[15] that only uses the headlight-illuminated nighttime image sequence. In this work, we use the image sequence without headlight-illumination which is more challenging. Both image sequences including daytime, nighttime with/without headlight-illumination of the ZJU dataset have been made publicly available at `https://github.com/elnino9ykl/ZJU-Dataset`.



(a) BDD dataset (b) Nighttime Driving dataset (c) ZJU dataset

Figure 3. Examples from three datasets. The streetscape of images in ZJU dataset varies from that in BDD dataset and Nighttime Driving dataset.

Nighttime Driving Dataset.[32] The dataset was collected during 5 rides with a car inside multiple Swiss cities and their suburbs using a GoPro Hero 5 camera, consisting of images of real driving scenes at nighttime and twilight time, with 35,000 unlabeled and 50 densely annotated images. In this paper, we only utilize the 50 annotated nighttime images for evaluation as well as comparison by taking the method proposed by D. Dai et al.[32] as a baseline. In general, the streetscape in this dataset is very similar to BDD dataset.

Table 1. Main information of the three datasets.

Dataset	**Resolution**	**Number of Images**		**Comment**
		Day	**Night**	
BDD100K	1280×720	52511	39986	No semantic segmentation labels
BDD10K	1280×720	7691	309	Precise semantic segmentation labels
ZJU	1920×1080	1700	1700	Different streetscape from other datasets
Nighttime Driving test	1920×1080	0	50	Good illumination by street lamp

Because of the huge style differences between the BDD datasets and ZJU datasets, we trained the GANs respectively for two datsets. In BDD100K, we select 6,000 daytime and 6,000 nighttime images, both in clear weather. These images are fed into CycleGAN to train a day-night converter named BDD-GAN. Because of the massive computation cost of CycleGAN, images are resized to 480×270. Similarly, 850 day-time image pairs in ZJU datasets are used to train a ZJU-GAN.

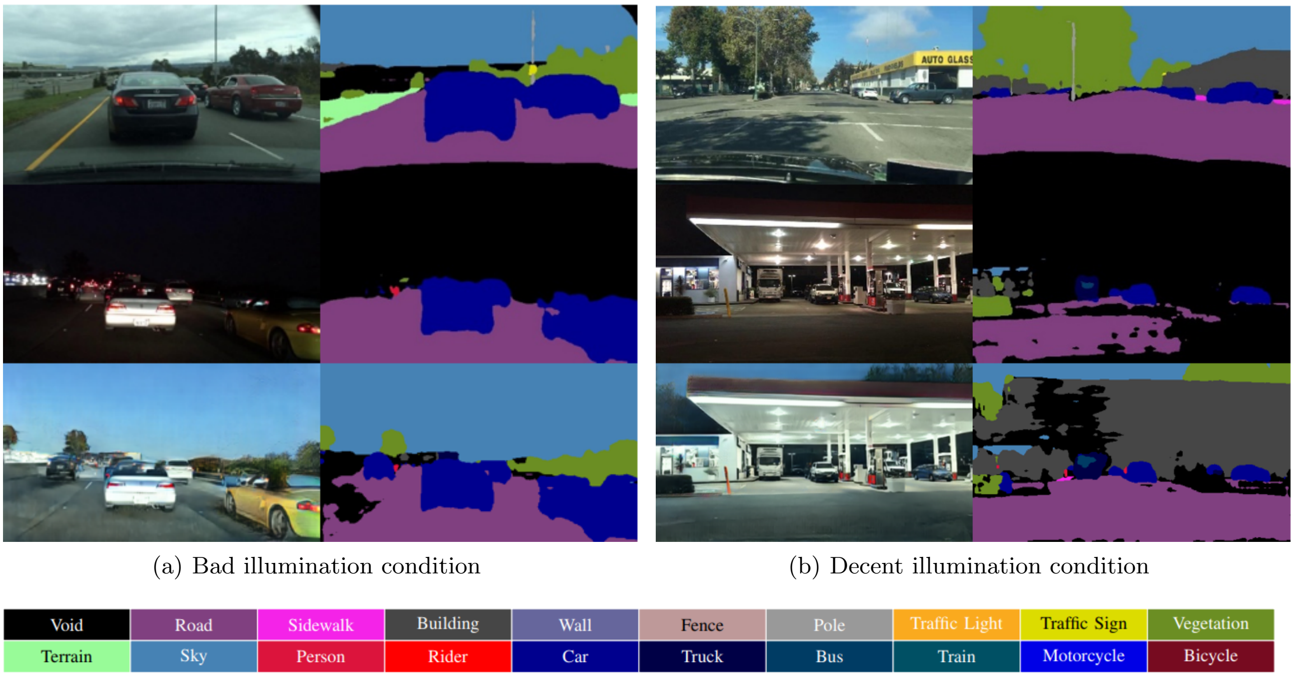
Like most semantic segmentation models, ERF-PSPNet[1,11] is composed of two parts: encoder and decoder. The encoder part has been trained on ImageNet[48] already, and all the training tasks for ERF-PSPNet lie in the training of the decoder part of the model. In the first method, ERF-PSPNet is trained on BDD10K. Nighttime images during inference are converted on-the-fly to daytime domain by CycleGAN. In the second method, different ratios of images in training set of BDD10K are used to train ERF-PSPNet. To quantitatively validate our method, we use the 32 nighttime images with segmentation annotation in the validation set of BDD10K and 50 nighttime images with precise segmentation annotations in the Nighttime Driving Test datasets. The style of images in Nighttime Driving Test dataset is similar to BDD10K, which makes it reasonable to apply BDD-trained semantic segmentation models on it.

4.2 Qualitative Results

In the first method, we use night-to-day converter to generate synthetic daytime images, which is the comfortable domain for ERF-PSPnet. Fig. 4 and Fig. 5 shows some representative results of our experiment. The first row and second row show the daytime images and nighttime images respectively, and the bottom row shows the synthetic daytime image converted form the images in second row.

Fig. 4 shows our results in BDD Dataset. First rows of both subfigures show the stably-behaving performance of ERF-PSPNet in daytime driving images. Nearly all the classes are labeled precisely. In the second row, subfigure (a) is worse illuminated than subfigure (b). Our daytime-trained model fails to detect the sky in both images. Cars in the left are ignored by the model in subfigure (a) and the whole gas station is missed in subfigure (b). In the bottom row, the model recognizes the sky, more cars and the whole gas station in the synthetic images successfully. In this way, GANs help the model improve performance at night. However, GANs also bring some problems: textures of objects like roads and cars in synthetic images are different from the ones in the real world, causing confusion to the model. As we can see in the bottom row of subfigure (a), the left part of the road is not labeled completely.

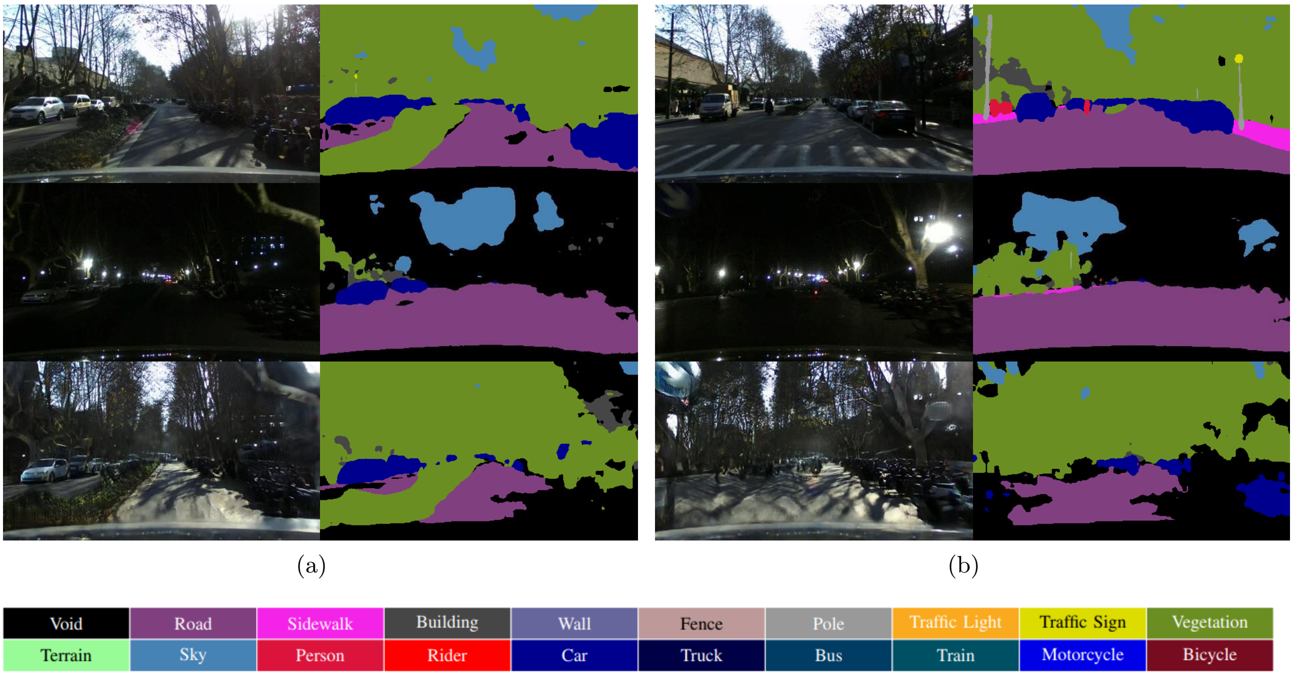
Fig. 5 shows our results in ZJU dataset. Because our semantic segmentation model is trained on BDD dataset, in which the streetscape significantly varies from ZJU datset, semantic segmentation outputs in the first row are not as ideal as ones in BDD dataset due to the geographical location-related domain gap between BDD (Berkeley) and ZJU (Hangzhou). In the second row we can find that all of the bushes and motorcycles beside the road are recognized as road, which is extremely dangerous for autonomous driving. In the bottom row, synthetic daytime images perform much better than nighttime images in the second row in general (night-to-day converter helps the model to correctly label the trees and bushes, which are very dark at night), but the lost details of the synthetic images make part of the road and the traffic sign missed in the labeled images.


(a) Bad illumination condition

(b) Decent illumination condition

Figure 4. Examples from BDD dataset (Top: day input, Mid: night input, Bottom: night-to-day converted input). Right image is better illuminated than left image. The model labeled the whole sky more precisely for synthetic daytime images than nighttime images. The night-to-day converter helps ERF-PSPNet perform better at night.

In the second method, different ratio of images with annotated labels in training set are converted to nighttime images to improve the robustness of semantic segmentation model in face of nighttime images. Fig. 6 shows performance of the model on the validation set of the BDD dataset and the testing set of the Nighttime Driving


(a)

(b)

Figure 5. Examples from ZJU dataset (Top: day input, Mid: night input, Bottom: night-to-day converted input). In the second row, all the most of the trees and some cars are missed in labeled images due to the darkness. In the bottom row, the segmentation of cars and vegetation has improved. But model performs not so well around the corner of the images.

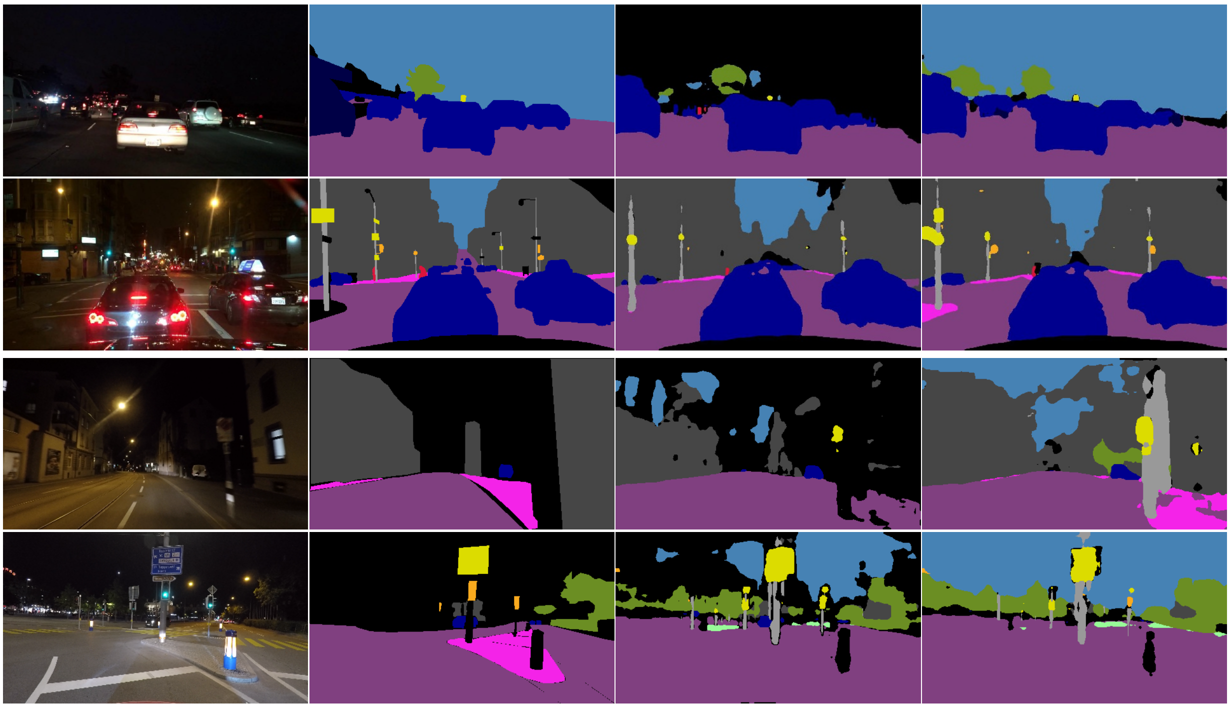
Dataset. In this example, our method relies on training ERF-PSPNet with 2,000 synthetic nighttime images and 5,000 original images in the training set. Different ratios of synthetic nighttime images in training set will be discussed in the next subsection. As shown in Fig. 6, the traffic signs and sky are significantly better labeled with our method. ERF-PSPNet yielded with the original training set performs decently for nighttime images with good illuminance (bottom row), but much worse under bad illuminance conditions. However, our method shows enhanced robustness in all conditions.

4.3 Quantitative Results and Discussion

Table. 2 shows the quantitative results of our two methods and contrast method. In the table, the first three rows are results from contrast method: training ERF-PSPNet with BDD10K dataset and testing with 968 daytime images and 32 nighttime images in the BDD10K test set. The contrast method represents the baseline for our proposed two methods.

The results of first method (converting nighttime images on-the-fly to synthetic daytime images during inference), are shown in the forth and fifth rows. As we can see, the results are below the baseline. In general, this method performs worse than no methods applied, but in some classes such as sky, car and trucks, accuracies improve remarkably. The main cause is that the textures of synthetic images from GANs are not as detailed as those of natural images. Semantic segmentation model is trained on natural daytime images, causing low accuracy in labeling synthetic daytime images from the test sets. Another possible reason is that nighttime images in BDD dataset and Nighttime Driving dataset are illuminated decently, so our method may not be much better than the baseline.

The results of second method: converting parts of the daytime images in training set to nighttime images on the stage of training, are shown in last four rows, together with a baseline method DarkModelAdaptation proposed by D. Dai et al.[32] validated on their dataset. Mean IoU increases remarkably for the nighttime images in the BDD testing set and Nighttime Driving test set, and keeps the same level of accuracy as the baseline for daytime images. Compared to the method proposed by D. Dai et al.,[32] our method rises nearly 4% on the same



(a) Input (b) Ground Truth (c) Ordinary ERF-PSPNet (d) Our Method

Figure 6. Examples from BDD Dataset and Nighttime Driving Dataset (Top two rows: BDD Dataset, Bottom two rows: Nighttime Driving Dataset). In general, our method (Converting 2,000 images to synthtic nighttime images in training) performs better than ordinary ERF-PSPNet(trained in original BDD10K training set). All the classes especially sky are recognized better

Table 2. Results of our two methods in BDD val set and Nighttime Driving test set.

Train/Method	**Test**	**Mean IoU**	**Mean Acc**
BDD dataset	Day in BDD test set	52.10%	75.52%
BDD dataset	Night in BDD test set	32.72%	75.67%
BDD dataset	Nighttime Driving test set	36.73%	72.38%
BDD dataset	Night2day in BDD test set	29.94%	56.87%
BDD dataset	Night2day in Nighttime Driving test set	32.74%	66.46%
Day2night BDD dataset	Day in BDD test set	**53.03%**	75.96%
Day2night BDD dataset	Night in BDD test set	**43.14%**	68.93%
DarkModelAdaptation[32]	Nighttime Driving test set	41.60%	NA
Day2night BDD dataset	Nighttime Driving test set	**45.09%**	72.82%

Nighttime Driving test set, even though our ERF-PSPNet is much smaller than RefineNet[8] adopted by them. In general, our method dramatically improves the performance of ERF-PSPNet in face of nighttime images.

We perform an important experiment to explore how the ratio of synthetic nighttime images in the training set influences the result. Fig.7 shows the mean IoU of our day-to-night method with respect to different ratios of synthetic nighttime images in the training dataset. When varying the number of synthetic nighttime images in the training set, we find that in the range of 0 - 2,000, as the ratio of synthetic nighttime images gets higher, the model performs increasingly better for nighttime images. The model learns well the illumination of scene elements in synthetic nighttime images and textures of real objects in daytime images. But as the synthetic nighttime images gets more, IoU gets down on the contrary. Additionally, at 5,000, the curve reaches another peak. The reason may be that 5,000 is a symmetrical number to 2,000 (7,000 in total), and the model learns the texture from daytime images and the illumination from synthetic nighttime images in a complementary way, but the daytime performance has already degraded to a low level. When all images in train set are converted to nighttime images, the IoU gets even lower than 30% for the same reason as the low IoU of the first method: the textures in synthetic images are different from that in real images. In the end, it turns out that the sweet spot is to use 2,000 synthetic nighttime images and 5,000 real daytime images as training set, as this ratio reaches the best accuracy for nighttime semantic segmentation, while daytime semantic segmentation also remains robust.

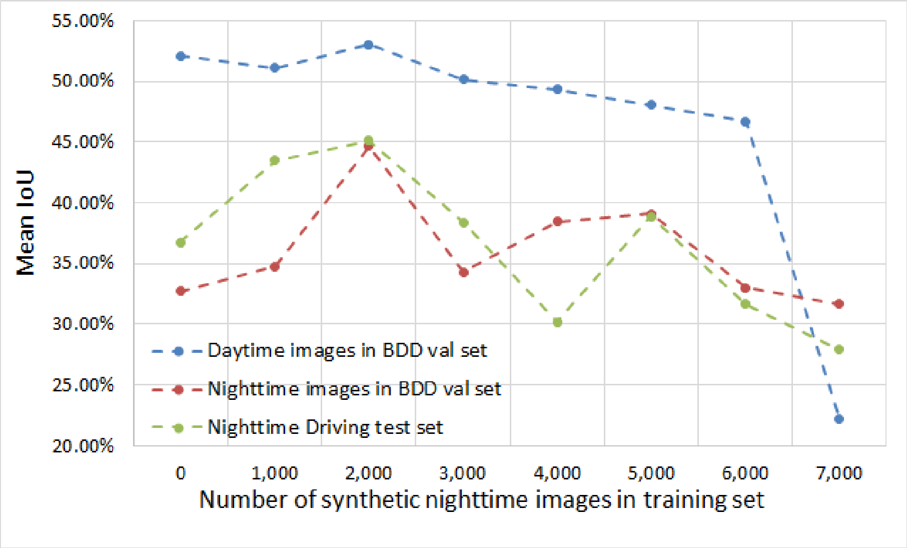


Figure 7. Number of synthetic nighttime images in training set - Mean IoU curve. IoU value peak appears at 2,000 synthetic nighttime images in the train set.

5. CONCLUSIONS

In this paper, we investigate the problem of semantic image segmentation of nighttime scenes. To improve the performance, two methods are proposed by training CycleGAN as a two-way day-night converter. In the first method, nighttime images are converted to daytime domain on-the-fly during inference as a pre-processing step. In the second method, a critical part of images of BDD training set are converted to synthetic nighttime images, improving the robustness of the segmentation model in the training process.

To validate our methods, three datasets are leveraged to obtain qualitative and quantitative results. Our comprehensive set of experiments indicates the path to follow, and the sweet spot to determine the training strategy, in order to attain the best robustness across the day and night. Overall, these results demonstrate that our methods improve the model performance observably, making state-of-the-art efficient networks like ERF-PSPNet work robustly at night.

REFERENCES

[1] Yang, K., Bergasa, L. M., Romera, E., Cheng, R., Chen, T., and Wang, K., "Unifying terrain awareness through real-time semantic segmentation," in [*2018 IEEE Intelligent Vehicles Symposium (IV)*], 1033–1038, IEEE (June 2018).

[2] Yang, K., Wang, K., Bergasa, L. M., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T., and López, E., "Unifying terrain awareness for the visually impaired through real-time semantic segmentation," *Sensors* **18**(5), 1506 (2018).

[3] Yang, K., Cheng, R., Bergasa, L. M., Romera, E., Wang, K., and Long, N., "Intersection perception through real-time semantic segmentation to assist navigation of visually impaired pedestrians," in [*2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*], 1034–1039, IEEE (December 2018).

[4] Long, N., Wang, K., Cheng, R., Yang, K., and Bai, J., "Fusion of millimeter wave radar and rgb-depth sensors for assisted navigation of the visually impaired," in [*Millimetre Wave and Terahertz Sensors and Technology XI*], **10800**, 1080006, International Society for Optics and Photonics (2018).

[5] Pfeuffer, A. and Dietmayer, K., "Robust semantic segmentation in adverse weather conditions by means of sensor data fusion," *arXiv preprint arXiv:1905.10117* (2019).

[6] Yang, K., Wang, K., Hu, W., and Bai, J., "Expanding the detection of traversable area with realsense for the visually impaired," *Sensors* **16**(11), 1954 (2016).

[7] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J., "Pyramid scene parsing network," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 6230–6239, IEEE (2017).

[8] Lin, G., Milan, A., Shen, C., and Reid, I., "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 5168–5177, IEEE (2017).

[9] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "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 (2017).

[10] Hu, X., Yang, K., Fei, L., and Wang, K., "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," *arXiv preprint arXiv:1905.10089* (2019).

[11] Yang, K., Hu, X., Bergasa, L. M., Romera, E., Huang, X., Sun, D., and Wang, K., "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in [*2019 IEEE Intelligent Vehicles Symposium (IV). IEEE*], 374–381 (June 2019).

[12] Yang, K., Bergasa, L. M., Romera, E., and Wang, K., "Robustifying semantic cognition of traversability across wearable rgb-depth cameras," *Applied optics* **58**(12), 3141–3155 (2019).

[13] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B., "The cityscapes dataset for semantic urban scene understanding," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 3213–3223, IEEE (2016).

[14] Neuhold, G., Ollmann, T., Bulò, S. R., and Kontschieder, P., "The mapillary vistas dataset for semantic understanding of street scenes," in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 5000–5009, IEEE (2017).

[15] Romera, E., Bergasa, L. M., Yang, K., Alvarez, J. M., and Barea, R., "Bridging the day and night domain gap for semantic segmentation," in [*2019 IEEE Intelligent Vehicles Symposium (IV)*], 1184–1190, IEEE (June 2019).

[16] Li, C., Xia, W., Yan, Y., Luo, B., and Tang, J., "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *arXiv preprint arXiv:1907.10303* (2019).

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in neural information processing systems*], 2672–2680 (2014).

[18] Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., and Darrell, T., "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687* (2018).

[19] Sun, D., Huang, X., and Yang, K., "A multimodal vision sensor for autonomous driving," in [*Artificial Intelligence and Machine Learning in Defense Applications*], International Society for Optics and Photonics (2019).

[20] Xiang, K., Wang, K., and Yang, K., "Importance-aware semantic segmentation with efficient pyramidal context network for navigational assistant systems," *arXiv preprint arXiv:1907.11066* (2019).

[21] Xiang, K., Wang, K., and Yang, K., "A comparative study of high-recall real-time semantic segmentation based on swift factorized network," *arXiv preprint arXiv:1907.11394* (2019).

[22] Shelhamer, E., Long, J., and Darrell, T., "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 640–651 (2016).

[23] Badrinarayanan, V., Kendall, A., and Cipolla, R., "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495 (2017).

[24] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062* (2014).

[25] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H., "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587* (2017).

[26] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in [*Proceedings of the European conference on computer vision (ECCV)*], 801–818 (2018).

[27] Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E., "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147* (2016).

[28] Romera, E., Alvarez, J. M., Bergasa, L. M., and Arroyo, R., "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems* **19**(1), 263–272 (2017).

[29] Sadat Saleh, F., Sadegh Aliakbarian, M., Salzmann, M., Petersson, L., and Alvarez, J. M., "Effective use of synthetic data for urban scene semantic segmentation," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], 84–100 (2018).

[30] Xu, Y., Wang, K., Yang, K., Sun, D., and Fu, J., "Semantic segmentation of panoramic images using a synthetic dataset," in [*Artificial Intelligence and Machine Learning in Defense Applications*], International Society for Optics and Photonics (2019).

[31] Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S. N., and Chellappa, R., "Learning from synthetic data: Addressing domain shift for semantic segmentation," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 3752–3761, IEEE (2018).

[32] Dai, D. and Van Gool, L., "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in [*2018 21st International Conference on Intelligent Transportation Systems (ITSC)*], 3819–3824, IEEE (2018).

[33] Sakaridis, C., Dai, D., and Van Gool, L., "Semantic nighttime image segmentation with synthetic stylized data, gradual adaptation and uncertainty-aware evaluation," *arXiv preprint arXiv:1901.05946* (2019).

[34] Sakaridis, C., Dai, D., Hecker, S., and Van Gool, L., "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], 687–704 (2018).

[35] Dai, D., Sakaridis, C., Hecker, S., and Van Gool, L., "Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding," *International Journal of Computer Vision* , 1–23 (2019).

[36] Porav, H., Bruls, T., and Newman, P., "I can see clearly now: Image restoration via de-raining," *arXiv preprint arXiv:1901.00893* (2019).

[37] Hu, X., Fu, C.-W., Zhu, L., and Heng, P.-A., "Depth-attentional features for single-image rain removal," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 8022–8031 (2019).

[38] Porav, H., Bruls, T., and Newman, P., "Don't worry about the weather: Unsupervised condition-dependent domain adaptation," *arXiv preprint arXiv:1907.11004* (2019).

[39] Larsson, M., Stenborg, E., Hammarstrand, L., Pollefeys, M., Sattler, T., and Kahl, F., "A cross-season correspondence dataset for robust semantic segmentation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 9532–9542 (2019).

[40] Lin, S., Cheng, R., Wang, K., and Yang, K., "Visual localizer: Outdoor localization based on convnet descriptor and global optimization for visually impaired pedestrians," *Sensors* **18**(8), 2476 (2018).

[41] Cheng, R., Wang, K., Lin, S., Hu, W., Yang, K., Huang, X., Li, H., Sun, D., and Bai, J., "Panoramic annular localizer: Tackling the variation challenges of outdoor localization using panoramic annular images and active deep descriptors," *arXiv preprint arXiv:1905.05425* (2019).

[42] Hinton, G., Vinyals, O., and Dean, J., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531* (2015).

[43] Gupta, S., Hoffman, J., and Malik, J., "Cross modal distillation for supervision transfer," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2827–2836 (2016).

[44] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 5967–5976, IEEE (2017).

[45] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 2242–2251, IEEE (2017).

[46] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 2999–3007, IEEE (2017).

[47] Yang, K., Bergasa, L. M., Romera, E., Huang, X., and Wang, K., "Predicting polarization beyond semantics for wearable robotics," in [*2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*], 96–103, IEEE (2018).

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision* **115**(3), 211–252 (2015).