SUPPLEMENTARY: WHAT'S THERE IN THE DARK

Sauradip Nag[*1] *Saptakatha Adak*[2] *Sukhendu Das*[3]

Visualization and Perception Lab, Department of CS&E, IIT Madras, Chennai-600 036, India
Email: [1]sauradipnag95@gmail.com, [2]sapta@cse.iitm.ac.in, [3]sdas@iitm.ac.in

A. NIGHT IMAGE SEGMENTATION NETWORK (NISENET) TRAINING

Most of the modern Deep based architectures used in computer vision requires large amount of training data. However, the dearth in semantic annotations of night time scenes has enforced us to generate night time images from day time images having semantic annotations. Since, we are creating night images from day images we term it as semi-real night image. We use the same semantic maps of day time images as ground truth for the semi-real night images where we can get good semantic annotation which is better than human annotation for real night images. Hence, the model can capture semantics of object which are occluded in dark and that makes the system reliable. We trained the CycleGAN [1] with $\sim 26.5K$ images of BDD [2] unlabelled day as source and same number of night images as target images. Since Mapillary-Vistas [3] has less number of night images, we did data augmentation to make the number of day and night images to $8K$. The day and semi-real night images are collected from validation and training set as it has annotations. We trained the CycleGAN [1] with these images to convert daytime test images of mapillary to convert them to night images. For training the Real channel, the optimizer used is Stochastic Gradient Descent (SGD) having Nesterov acceleration with momentum as 0.8 and learning rate as 0.05. The Real channel was finetuned for UNDD Dataset since it has less number of annotations. A total of 35 densely annotated labels and its augmentation which adds up to 700 images has been used to fine-tune the model trained for BDD Dataset since it shares same structural similarity with the scenes. For validation, 10 labels and its augmentation which sums up to 200 images have been used. Remaining 30 labels and its augmentation which adds up to 600 images has been used for testing. For training of Adaptive channel, we used the same number of classes for both source and target domain to prevent false positives and degrading semantic quality. The output and the input image dimensions of this channel is same as Real Channel which is 256×256. The generator network in the Adaptive channel is trained for 500 epochs with early stopping having SGD as optimizer with Nesterov acceleration where the momentum is 0.9 and the weight decay is 10^{-4}. The initial learning rate was 0.0002 and is decreased using the polynomial decay with power of 0.9 as mentioned in [4] For training the discriminator, we followed the same protocols as mentioned in Adaptsegnet [5]. Adam optimizer [6] is used as optimizer with the learning rate as 10^{-4} and the same polynomial decay as the generator network. The momentum is set as 0.9. For fusing the multi-channel outputs, we trained a multi-scale based CNN to determine the similarity between the patch of the predicted segmentation map with its ground truth patch. It is a binary classification problem where 2 classes are "good patch", a patch which is semantically similar to ground truth and "bad patch", which is semantically not similar to ground truth. For training the Multi-Scale Fusion Module the following assumptions are made :

1. A good patch is taken from the ground truth.
2. A bad patch is taken from a segmentation map which is obtained after passing a real image using a model pre-trained for synthetic day images.
3. Both the good and bad patches are overlapping and represents part of the same scene.
4. The CNN will be trained with test images of target domain used in adaptive channel.

Since, this is a two class classification problem, we label each good patch as $[1\ 0]^T$ and each bad patch as $[0\ 1]^T$. The patches of size 20×20 from both real and adaptive channel output representing the same scene and spatial location is passed to the multi-scale network. They are convolved to same number of spatial dimensions using 1×1 convolution and merged together. The resulting output is passed into a series of fully connected layers to classify whether the patch is good or bad. The *argmax*

*Corresponding Author

of the output of softmax layer is taken and the index returned by the $argmax$ operation indicates which patch among Real and Adaptive channel to consider for patch centered at location i. The optimizer used for training is adadelta [7].

B. URBAN NIGHT DRIVING DATASET (UNDD)

The *Urban Night Driving Dataset* (UNDD) is compiled using 12 videos, where 8 videos are taken at different times of the day and 4 videos are taken at night. The apparatus used for capturing the video is a high end smart phone having 20MP camera with Optical Image Stabilization mounted as dashboard camera inside a car which was driven in and around 3 cities and its suburbs. The videos are typically 20 seconds in length on average and captured at 30 fps. Each frame is extracted at a rate of one frame per second, leading to 7200 images day and night combined. Out of these 4800 are daytime images and 2400 are night time images. The images captured are mostly of 1080p resolution, but there are images with 720p and its lower resolutions. There are ~ 80 key-frames among all night time images. Out of them, 75 key-frames have been densely annotated with pixel level annotations. The class definitions are same as that of Cityscapes [8] dataset. The summary of the dataset is given in table B.1. The comparison of our proposed dataset with other existing datasets have been given below in table B.2. Examples of day and night scenes are exhibited in Fig. B.1.

Table B.1. Details of our proposed Urban Night Driving Dataset (UNDD).

Type	Images	With Augmentation
Full	7200	-
Day	4800	-
Night	2400	-
Labelled	75	1500
Train	35	700
Validation	10	200
Testing	30	600

Table B.2. Comparative study of UNDD dataset with existing scene segmentation datasets.

Specifications	KITTI [9]	Cityscapes [8]	Mapillary [3]	BDD [2]	Nighttime Driving [10]	**UNDD(Ours)**
# Sequences	22	∼50	N/A	100,000	5	12
# Images	14,999	5000	25,000	120,000,000	35,000	7200
# Night Images	0	0	<250	∼27,000	9500	2400
# Labelled Night Images	0	0	<250	<400	50	75
Multiple Cities	No	Yes	Yes	Yes	Yes	Yes
Multiple Weather	No	No	Yes	Yes	No	No
Multiple Time of the Day	No	No	Yes	Yes	Yes	Yes
Multiple Scene Types	Yes	No	Yes	Yes	Yes	Yes

Fig. B.1. Examples of day and night scenes from our proposed UNDD dataset. The last row contains the Ground truth segmentation corresponding to the night images in the previous row (best viewed in color).

C. COMPARATIVE STUDY OF QUALITATIVE RESULTS

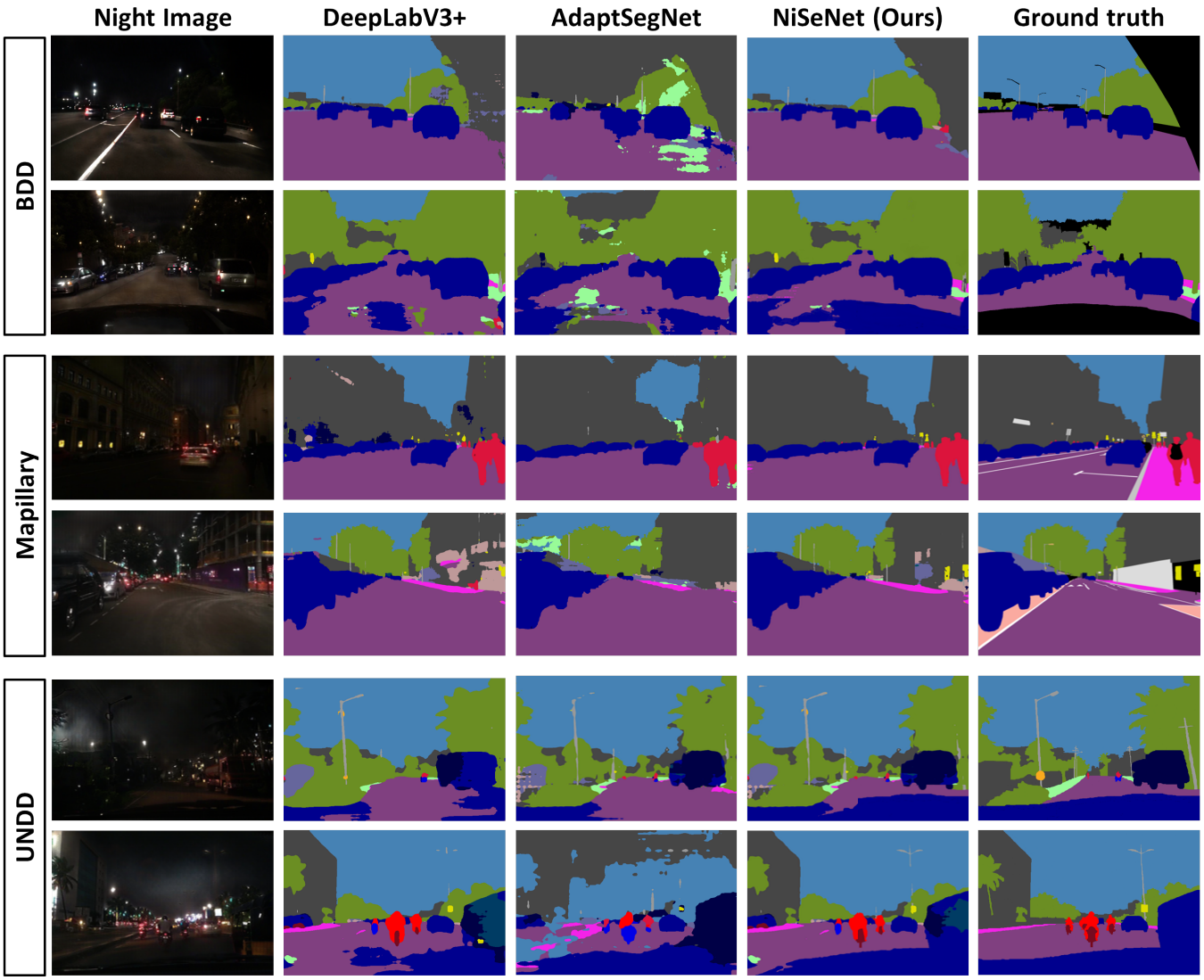

Fig. C.1. Comparitive study of visual results of our proposed NiSeNet with existing state-of-the-art methods on night images of three real-world datasets (best viewed in color).

D. REFERENCES

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[2] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[3] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[5] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, "Learning to adapt structured output space for semantic segmentation," in *CVPR*, 2018.

[6] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[7] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[10] Dengxin Dai and Luc Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *ITSC*, 2018.