WHAT'S THERE IN THE DARK

SAURADIP NAG*[1], SAPTAKATHA ADAK*[2], SUKHENDU DAS[3]
VISUALIZATION AND PERCEPTION LABORATORY, DEPARTMENT OF CS & E
INDIAN INSTITUTE OF TECHNOLOGY MADRAS, CHENNAI, INDIA

SCOPE AND HIGHLIGHTS

- Parsing complex real-world night scenes are often hard due to unfavourable weather and illumination conditions, as well as less avaibility of annotated real-world datasets.
- We propose a novel multi-scale fusion based architecture which takes real and synthetically generated segmentation maps and selects "best-of-both worlds" to produce an accurate final segmentation map.
- We additionally compiled a new dataset Urban Night Driving Dataset (UNDD) having 75 annotated maps of night scenes.
- Extensive performance evaluation on BDD and Mapilllary dataset, untouched yet in the domain of night image segmentation.

NETWORK ARCHITECTURE

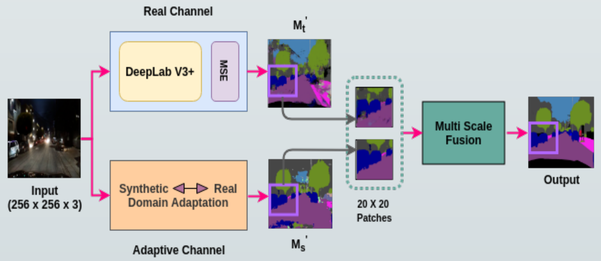

Fig. 1: The proposed NiSeNet architecture

- Given an input night image of dimension 256 x 256 , the network predicts semantic segmentation of the input.
- The model is composed of 2 streams namely Real Channel (uses real training data) and Adaptive Channel (uses synthetic training data).
- Finally, a novel multi-scale comparator based fusing network is used to fuse the output of two streams.
- Each components are trained individually and then the overall model is trained end-to-end.

TWO STREAM NETWORK

- **Real Channel:**

The real channel uses DeepLabv3+ coupled with MSE Loss to compensate for the less number of annotated real night-time images for training.

$$\mathcal{L}_{real}(M', M_t) = \lambda_c \mathcal{L}_{ce}(M', M_t) + \lambda_m \|M' - M_t\|_2$$

where, M' represents predicted map and M_t represents ground truth(GT) map of night input image. $\mathcal{L}_{ce}$ represents softmax cross-entropy criterion between predicted and GT maps. $[\lambda_c , \lambda_m]$=[0.4 , 0.6].

- **Adaptive Channel:**

The adaptive channel uses a GAN based Domain Adaptation technique which is trained using SYNTHIA Night Image dataset.

$$\mathcal{L}_{adapt}(M'_s, M_s, I_t) = \lambda_s \mathcal{L}_{seg}(M', M_t) + \lambda_a \mathcal{L}_{adv}(I_t)$$

Segmentation loss between generated source map (M'_s) and its ground truth (M_s), is:

$$\mathcal{L}_{seg}(M'_s, M_s) = \mathcal{L}_{ce}(M'_s, M_s) - \lambda_j \mathcal{L}_{jacc}(M'_s, M_s)$$

This is a combination of cross entropy loss ($\mathcal{L}_{ce}$) and a Jaccardian loss ($\mathcal{L}_{jacc}$), as $(\lambda_c , \lambda_m , \lambda_j)$ = (0.55, 0.45, 0.6).

$$\mathcal{L}_{jacc}(M'_s, M_s) = \frac{1}{N}\sum_{i=0}^{N} \frac{M_s \odot M'_s}{M_s + M'_s - M_s \odot M'_s}$$

where, M'_s is predicted softmax , M_s is ground truth label. $\odot$ represents the Hadamard product between 2 vectors. $\mathcal{L}_{jacc}$ helps in stabilizing unbalanced training data and slow convergence.

MULTI-SCALE FUSION

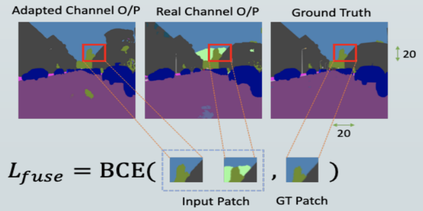

Fig. 2: Binary Cross Entropy Loss is calculated between a predicted output map (calculated using 20 x 20 image patches over multiple scales and concatenating features maps) and the GT map at the same location.

- The 20 x 20 patches are passed through same convolution layers at multi-scales (5x5, 7x7) .
- **Assumption :** While training, a patch having semantic similarity with GT is considered good (labelled $[1\ 0]^T$) and bad (labelled $[0\ 1]^T$) otherwise.

COMBINED OBJECTIVE FUNCTION

- The Real and Adaptive Channels are trained individually and then used to train the multi-scale fusion module.
- The overall loss function for training NiSeNet (end-to-end):

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{real}(M', M_t) + \lambda_2 \mathcal{L}_{adapt}(M'_s, M_s, I_t) + \lambda_3 \mathcal{L}_{fuse}(M''_t, M_t)$$

where, $\lambda_1, \lambda_2, \lambda_3$ are weighing parameters, having weights 0.4 , 0.3 and 0.3 respectively.

URBAN NIGHT DRIVING DATASET

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. It amounts to 4800 daytime and 2400 night-time images. Out of the available night images, 75 keyframes are densely annotated.

Fig 3: Samples from our proposed Urban Night Driving Dataset (UNDD)

RESULTS (QUANTITATIVE)

Methods	Mean IOU (%)		
	BDD	Mapillary	UNDD
DenseASPP [1]	33.56	35.21	34.55
CBST [2]	41.87	40.70	37.97
DeeplabV3+ [3]	38.41	37.64	39.23
AdaptSegnet [4]	42.17	39.55	38.42
NiSeNet (ours)	**53.52**	**48.32**	**45.56**

Table 1: Comparison of performance of the proposed frame prediction method with recent and state-of-the-art approaches. Best results are in bold.

RESULTS (QUALITATIVE)

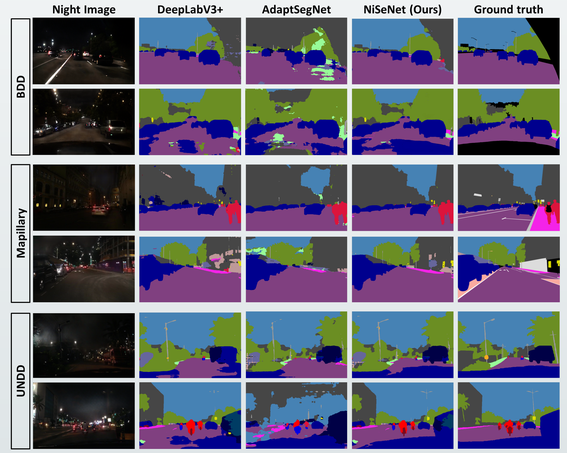

Fig. 4: A comparative study of the qualitative results on 3 real world datasets

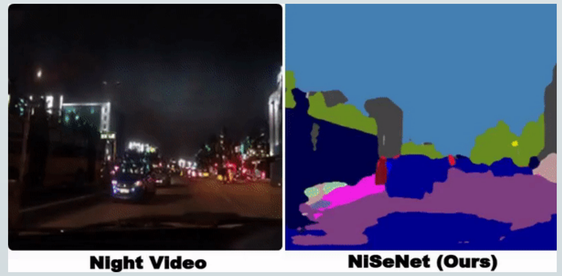

Fig. 5: Result of segmentation of NiSeNet on a real-world videoframe.

CONCLUSION

- The proposed NiSeNet is one of the early works in the domain of semantic segmentation on night images.
- Admirable performance in terms of mIOU(%) score in night-dataset evaluation for 3 real world dataset BDD, Mapillary, UNDD.
- A standard dataset UNDD has been compiled using 7500 night images where 75 keyframes are densely annotated.

References

[1] Yang, Maoke, et al. "Denseaspp for semantic segmentation in street scenes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
[2] Zou, Yang, et al. "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training." Proceedings of the European Conference on Computer Vision (ECCV). 2018. .
[3] 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.
[4] Tsai, Yi-Hsuan, et al. "Learning to adapt structured output space for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018

Contact

[1]sauradipnag95@gmail.com [2]sapta@cse.iitm.ac.in
[3]sdas@iitm.ac.in www.cse.iitm.ac.in/~vplab
* Both authors contributed equally