Appendix for DepthSegNet24: A Label-Free Model for Robust Day-Night Depth and Semantics

Phan Thi Huyen Thanh[1⋆[0009−0007−1781−4827]], The Hiep Nguyen[2,3∗[0009−0007−5765−0468]], Minh Huy Vu Nguyen[2,3[0009−0003−4599−7182]], Trung Thai Tran[2,3[0009−0002−1422−9685]], Tran Vu Pham[2,3], Duc Dung Nguyen[2,3[0000−0001−7321−7401]], Truong Vinh Truong Duy[1[0000−0001−8000−2214]], and Natori Naotake[1[0000−0002−6128−4145]]

[1] Tokyo Research Center, Aisin Corporation, Japan
[2] AITech Lab., Ho Chi Minh City University of Technology (HCMUT)
[3] Vietnam National University Ho Chi Minh City (VNUHCM)
{thanh.phan,duy.truong,naotake.natori}@aisin.co.jp
{hiep.nguyena113872,huy.vu.cse.9,thai.tran241002,ptvu,nddung}@hcmut.edu.vn

Section 1 provides details information on the datasets, and Section 2 presents more quantitative and qualitative results on Oxford Robotcar.

1 Datasets

1.1 Oxford RobotCar

The Oxford RobotCar dataset is a large-scale driving dataset captured in various driving scenarios at different times throughout the day [8]. Following NightDepthADFA [9], we utilize the left images collected by the front-view trinocular stereo camera (Bumblebee XB3) of two daytime and nighttime sequences "2014-12-09-13-21-02" and "2014-12-16-18-44-24" for training, respectively. We follow ADDS-DepthNet's [6] splits and data setup with 21,932 daytime and 22,811 nighttime training images. The testing data includes 451 daytime images and 411 nighttime images from the other splits of the dataset. All images are center-cropped to 640 × 1280 resolution and then downscaled to 256 × 512. It should be noted that the dataset does not provide SS ground-truth labels.

1.2 nuScenes

The nuScenes dataset offers over 1,000 challenging urban driving scenarios with detailed object annotations, including semantic segmentation [1]. We extend the STEPS split [10] for nighttime training to 13,261 images. In line with previous practices [3, 4], we make use of the official split with 15,129 daytime training images. The testing split comprises 4,449 daytime images and 602 nighttime images at a resolution of 320 × 576 pixels.

⋆ Equal contributors

Table 1: Depth on Oxford RobotCar (depth = 60 m). Best: **bold**, second best: underlined. Our multi-task model achieves the best or second-best performance compared to depth-specifically-optimized dedicated SOTA methods.

Night - Oxford RobotCar Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Monodepth2 [5] (day)	0.432	5.366	11.267	0.463	0.361	0.653	0.839
Monodepth2 [5] (night)	0.580	21.446	12.771	0.521	0.552	0.840	0.920
HR-Depth [7]	0.462	5.660	11.009	0.477	0.374	0.670	0.842
ADDS-DepthNet [6]	0.231	2.674	8.800	0.286	0.620	0.892	0.956
Ours	**0.196**	**2.056**	**7.921**	**0.259**	**0.709**	**0.902**	**0.965**
Day - Oxford RobotCar Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Monodepth2 [5] (day)	0.124	0.931	5.208	0.178	0.844	0.963	**0.989**
Monodepth2 [5] (night)	0.294	2.533	7.278	0.338	0.541	0.831	0.934
HR-Depth [7]	0.129	1.013	5.468	0.184	0.825	0.958	**0.989**
ADDS-DepthNet [6]	**0.115**	**0.794**	**4.855**	**0.168**	**0.863**	**0.967**	**0.989**
Ours	0.117	0.810	4.921	0.178	0.846	0.963	0.988

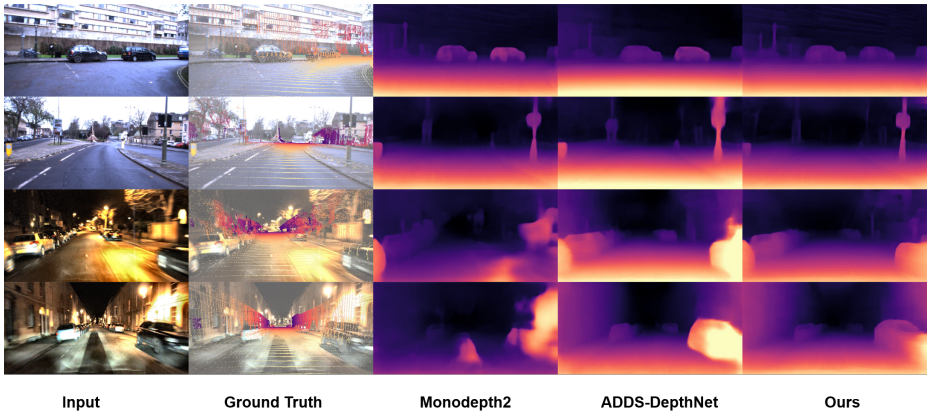
2 Additional Results

2.1 Depth on Oxford RobotCar

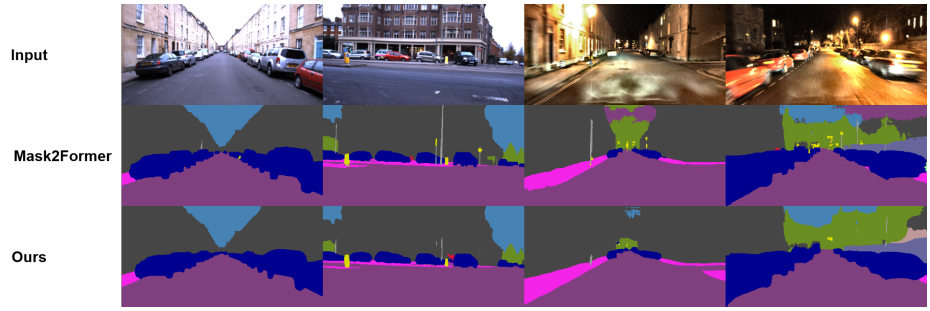
Table 1 illustrates the performance comparisons of different depth estimation models on nighttime and daytime scenes up to 60 m of this Oxford RobotCar dataset. For nighttime scenes, Monodepth2 [5], trained on daytime and nighttime data, achieved absolute relative scores of 0.432 and 0.580, respectively. HR-Depth [7] recorded a score of 0.462, ADDS-DepthNet [6] achieved 0.231, and our model attained a score of 0.196. In daytime scenes up to 60 m, Monodepth2 achieved scores of 0.124 and 0.294 for the trained daytime and nighttime splits, respectively. HR-Depth scored 0.129, ADDS-DepthNet achieved 0.115, and our model recorded a score of 0.117. These results highlight the increased difficulty of nighttime scenes compared to daytime scenes. Our model demonstrated relatively good performance in nighttime and daytime scenarios, achieving the lowest score in nighttime scenes and performing comparably to Monodepth2 (day) in daytime scenes. Notably, our model is not a single-task model specifically tailored for this task, indicating its robust performance across various domains. It is important to note that Monodepth2, HR-Depth, and ADDS-DepthNet are single-task models designed and optimized for the MDE task.

2.2 Qualitative Results

Qualitative comparisons in Figure 1 illustrate the depth and semantic segmentation predictions for some representative day and night scenes on Oxford RobotCar. Our model provides precise depth estimates for various objects, including small traffic sign, and generates clear depth maps for both day and night scenes. Additionally, the segmentation maps accurately recognize and classify objects and boundaries such as cars and roads, assigning appropriate semantic labels. In the nighttime scenes, it even visualizes better results than the teacher model



(a) Depth on Oxford RobotCar.



(b) Semantic segmentation on Oxford RobotCar.

Fig. 1: Qualitative results on Oxford RobotCar. Our results display clear depth and correct classification of the objects and their sharp boundaries in both day and night scenes.

Mask2Former [2] especially in the sky and road classes. These results demonstrate our multi-task model's ability to seamlessly integrate MDE and SS tasks, resulting in visually consistent and coherent perception.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11618–11628. Computer Vision Foundation / IEEE (2020)
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)

3. Gasperini, S., Koch, P.N., Dallabetta, V., Navab, N., Busam, B., Tombari, F.: R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. 2021 International Conference on 3D Vision (3DV) pp. 751–760 (2021), https://api.semanticscholar.org/CorpusID:236965838
4. Gasperini, S., Morbitzer, N., Jung, H., Navab, N., Tombari, F.: Robust monocular depth estimation under challenging conditions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8177–8186 (2023)
5. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 3827–3837. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00393
6. Liu, L., Song, X., Wang, M., Liu, Y., Zhang, L.: Self-supervised monocular depth estimation for all day images using domain separation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12717–12726 (2021), https://api.semanticscholar.org/CorpusID:237142261
7. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: High resolution self-supervised monocular depth estimation. ArXiv **abs/2012.07356** (2020), https://api.semanticscholar.org/CorpusID:229152988
8. Maddern, W.P., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**, 15 – 3 (2017), https://api.semanticscholar.org/CorpusID:22556995
9. Vankadari, M.B., Garg, S., Majumder, A., Kumar, S., Behera, A.: Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:222133092
10. Zheng, Y., Zhong, C., Li, P., Gao, H., Zheng, Y., Jin, B., Wang, L., Zhao, H., Zhou, G., Zhang, Q., Zhao, D.: Steps: Joint self-supervised nighttime image enhancement and depth estimation. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 4916–4923 (2023), https://api.semanticscholar.org/CorpusID:256503497