A. Dataset Construction

Dataset for self-supervised monocular depth training in nighttime is under-explored. To make up this lack, we build two nighttime datasets, named RobotCar-Night (RC-N) and nuScenes-Night (NS-N). The two datasets consist of many video clips from Oxford RobotCar [5] and nuScenes [2], along with carefully generated ground truth using the official toolbox[1].

In RobotCar, the number of LiDAR points in one frame is relatively small, so multiple frames are combined to generate the ground truth depth using official scripts. This process is based on Structure-from-Motion (SFM), therefore moving objects lead to wrong outputs. For example, a generated depth map is visualized in Fig. 1. It shows an obvious mistake on the moving car framed by a red box. To tackle this problem, we manually select scenes without moving object and carefully pick up many high-quality outputs among them. This approach is different from the previous work [7], in which random sampling is used to choose test samples.

Figure 1. Sample from Oxford RobotCar containing moving objects. The red box indicates a wrong construction of depth.

Remark. In main text, there are several samples containing moving objects in Fig. 5. These samples are from the same video sequences as RC-N but neither included in test nor training split. This is also the case for the last sample in Fig. 1.

By contrast, the LiDAR data in nuScenes contain more than 3,000 valid points in one frame and covers a wide range of depth values. Thus, data from single frame is used to prepare the ground truth depth maps and random sampling is applied to form the final test split.

Furthermore, some video clips containing daytime scenarios are selected from Oxford RobotCar and nuScenes to separately build RobotCar-Day (RC-D) and nuScenes-Day (NS-D), which are used to generate referenced depth maps.

B. Parameter Setting

Here, we discuss the parameter setting about σ in MCIE and ϵ in SBM. Images captured in low light environments are usually noisy, thus a smaller σ should be set to avoid an excessive enhancement on noise. In darker scenarios, more textureless pixels need to be masked out. Therefore, ϵ should be set to a bigger value. In our experiment, (σ, ϵ) is set to $(0.008, 10)$ and $(0.004, 20)$ on RC-N and NS-N dataset, respectively. This can be a empirical reference to set these two parameters.

[1]Oxford RobotCar: https://github.com/ori-mrg/robotcar-dataset-sdk, nuScenes: https://github.com/nutonomy/nuscenes-devkit

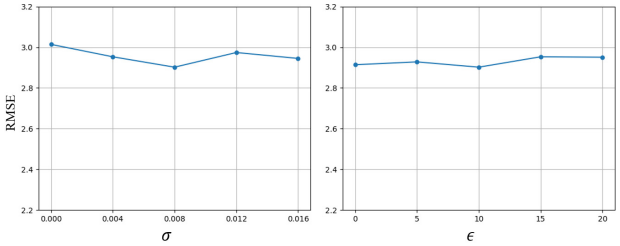

Figure 2. The left and right chart separately show the effect of σ and ϵ. The y axis is RMSE error and the x axis denotes different values of these two parameters.

To explore the effect of these two parameters, we conduct series comparison tests on RC-N and report the RMSE error in Fig. 2. The variables in the left and right chart are σ and ϵ, respectively. Zero indicates the corresponding module is not enabled. Overall, these two parameters impact little to the framework which performs the best when $\sigma = 0.008, \epsilon = 10$.

Generally speaking, $[0.002, 0.01]$ and $[10, 20]$ are proper ranges for σ and ϵ, respectively. Besides, comparison tests can help to choose the best parameter setting.

C. Selection of Referenced Scene

In our framework, the reference depth maps are generated by a depth estimation network Φ'_d trained on RC-D and NS-D in a self-supervised manner. They provide prior knowledge about depth distributions and are unpaired with nighttime scenarios. Generally speaking, depth maps in various driving scenes can be used as references, since they share similar depth distributions. To explore the effect of different reference scenarios, we train the framework in two referenced scenarios and report quantitative results in Tab. 1. The method *Our (RobotCar-Day)* achieves a slightly worse but similar performance compared to *Our (nuScenes-Day)* and significantly outperforms other SOTA methods. This illustrates that depth maps from other scenarios are also able to regularize training.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
MonoDepth2 [3]	1.1848	42.3059	21.6129	1.5699	0.1842	0.3598	0.5044
SfMLearner [8]	0.6004	8.6346	15.4351	0.7522	0.2145	0.4166	0.5961
SC-SfMLearner [1]	1.0508	30.5865	19.6004	0.8854	0.1823	0.3673	0.5422
PackNet [4]	1.5675	61.5101	25.8318	1.3717	0.1387	0.2980	0.4313
FM [6]	1.1383	41.6166	20.8481	1.1483	0.2376	0.4252	0.5650
Our (nuScenes-Day)	**0.3150**	**3.7926**	**9.6408**	**0.4026**	0.5081	**0.7776**	**0.8959**
Our (RobotCar-Day)	0.3285	4.3069	10.2651	0.4197	**0.5142**	0.7642	0.8813

Table 1. Quantitative results on nuScenes-Night, using depth maps from nuScenes-Day and RobotCar-Day as references, respectively.

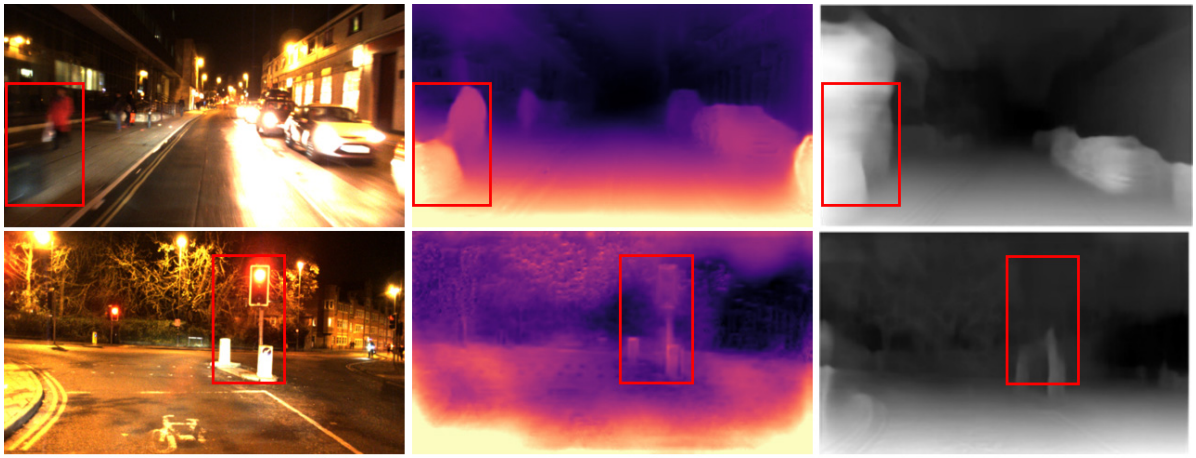

Figure 3. Qualitative comparison between our method and ADFA [7] on two samples containing saturated and blurred regions. These two images come from the Fig. 4 of ADFA.

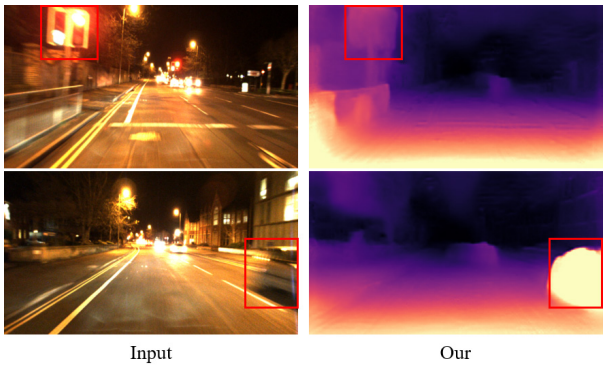

Figure 4. Qualitative results of our method on two samples containing saturated and blurred regions. These two images are from RobotCar-Night.

Figure 5. Qualitative results on very low-light images, where the top one and last two depth maps are generated by ADFA [7] and our method, respectively. The first image comes from the Fig. 4 of ADFA [7] and the last two images are from nuScenes-Night.

D. Comparison with ADFA in Challenging Cases

ADFA [7] claims three challenging cases that lead to its failure, including nighttime images with very low-illumination conditions, blurred image regions and saturated regions (bright light spots). Here, we further compare our method with ADFA in these three cases.

Blurred and Saturated Image Regions. Fig. 3 shows a comparison on two image samples containing saturated and blurred regions. Compared with ADFA, our method achieves better performance on presenting the shape of objects. Furthermore, two similar samples are shown in Fig. 4 to further illustrate the advantages of our method.

Images with very Low-Illumination Conditions. Very low-light images are a huge challenge for self-supervised depth estimation. Fig. 5 shows three samples in very low illuminated environments, where the top one and last two are generated by ADFA and our method, respectively. On the first sample, ADFA produces a blurry and inaccurate depth map. In contrast, our method is still able to make a plausible prediction on the second sample. The last sample is captured in a very dark environment. Our method makes a coarse estimation on some objectives but fails to depict the depth of entire scene.

E. More Qualitative Result

Here, we show more qualitative results on RobotCar-Night and nuScenes-Night datasets in Fig. 6 and Fig. 7, respectively. Five SOTA methods are evaluated for comparison, including SfMLearner [8], SC-SfMLearner [1], PackNet [4], MonoDepth2 [3] and FM [6].

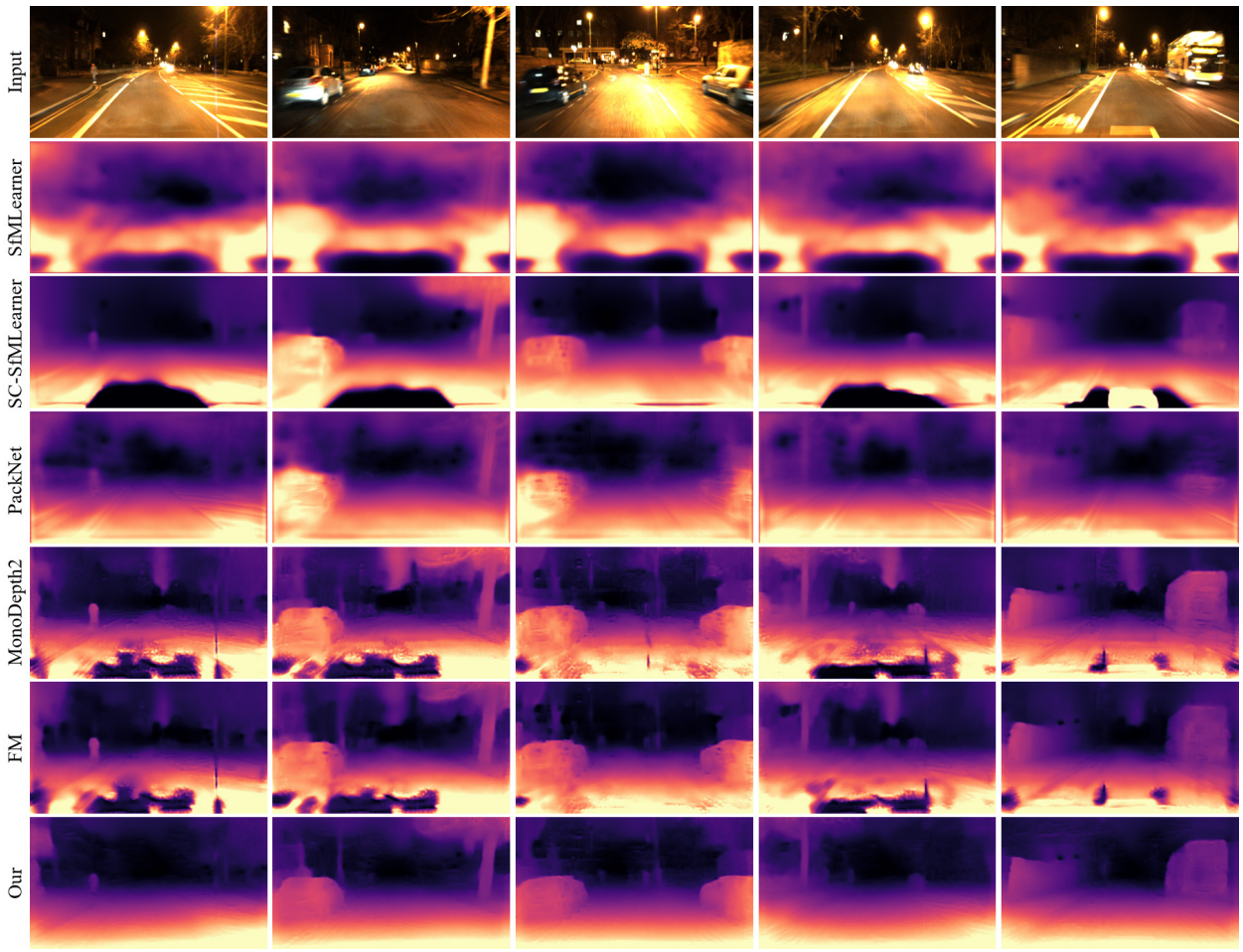

Figure 6. Qualitative comparison on RobotCar-Night.

F. Evaluation Metrics

There are seven standard metrics are used for evaluation, including Abs Rel, Sq Rel, RMSE, RMSE log, δ_1, δ_2 and δ_3, which are presented by

$$\begin{aligned}
\text{Abs Rel} &= \frac{1}{|D|}\sum_{d^*\in D}|d^*-d|/d^*, \\
\text{Sq Rel} &= \frac{1}{|D|}\sum_{d^*\in D}\|d^*-d\|^2/d^*, \\
\text{RMSE} &= \sqrt{\frac{1}{|D|}\sum_{d^*\in D}\|d^*-d\|^2}, \\
\text{RMSE log} &= \sqrt{\frac{1}{|D|}\sum_{d^*\in D}\|logd^*-logd\|^2}, \\
\delta_i &= \frac{1}{|D|}|\{d^*\in D \max(\frac{d^*}{d},\frac{d}{d^*})<1.25^i\}|,
\end{aligned} \tag{1}$$

where d and $d*$ separately denotes predicted and ground truth depth maps, D indicates a set of valid ground truth depth values in one image, $|.|$ returns the number of elements in the input set.

G. More discussion on experiment results

More Analysis on Experiments. In Table. 1 of main text, we show the evaluation results on RC-N. One may notice that, *MonoDepth2 (Day)* and *FM (Day)* achieve better results on the first four error metrics yet worse on the last three accuracy ones than their counterparts. Here, we present a possible explanation on this phenomenon. Fig. 8 shows two depth maps generated by MonoDepth2 (abbr. MD2) and MonoDepth2 (Day), respectively. The former produces more detailed results but with big holes while the later generates blurry outputs without holes. The big holes indicate a very large depth value and differ greatly from the Ground Truth, thus MD2 gets higher average errors on Abs_Rel, Sq_Rel, RMSE and RMSE_log. Conversely, $\sigma_{1,2,3}$ denote the percentage of pixels below a certain threshold, therefore MD2 outperforms MD2 (Day) by more accurate predictions within non-hole areas.

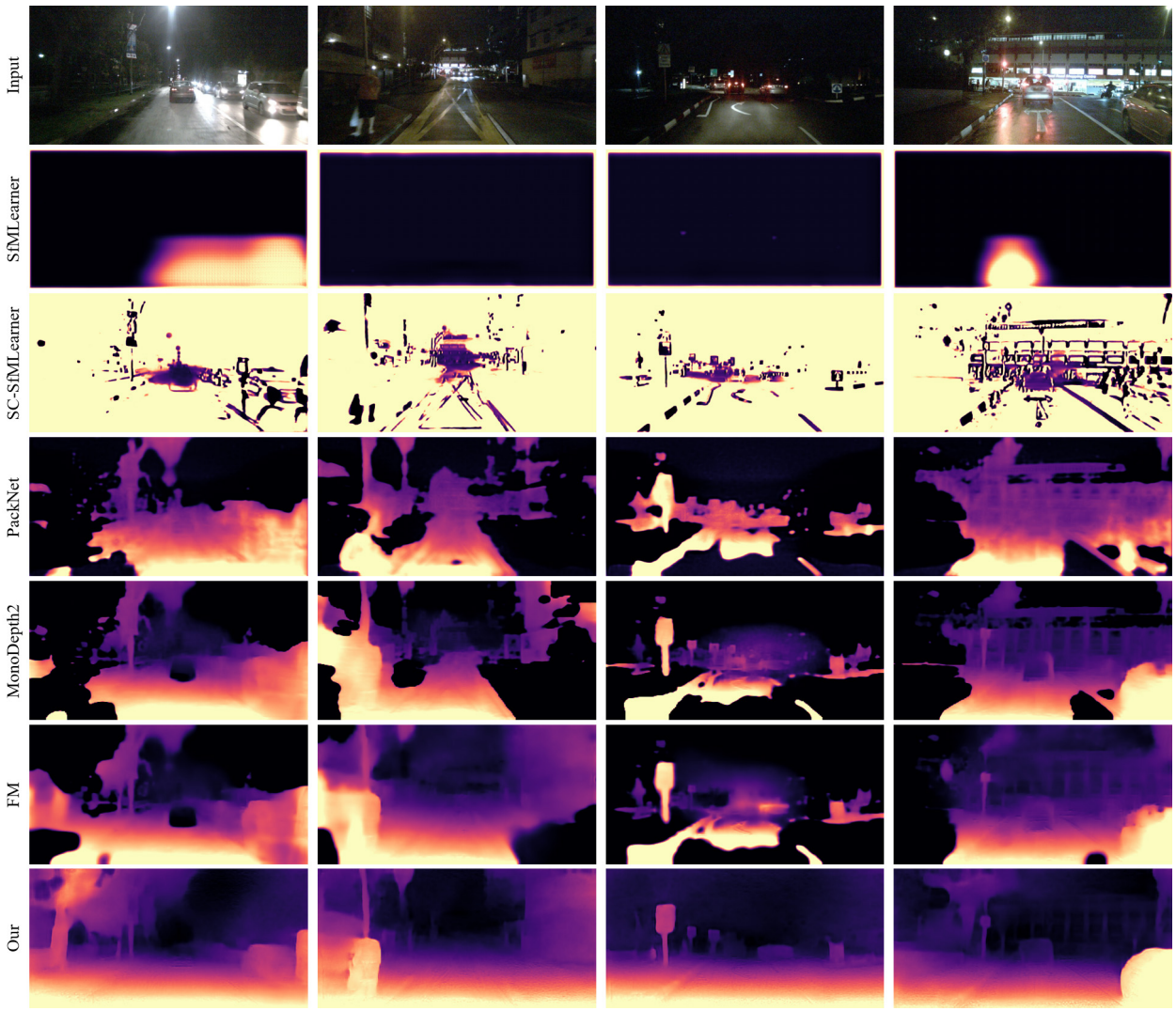

Figure 7. Qualitative comparison on nuScenes-Night.

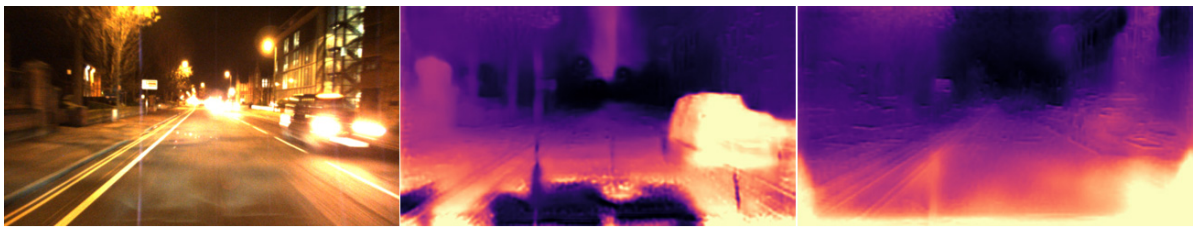

Figure 8. Qualitative comparison between MonoDepth2 (middle) and MonoDepth2 (Day) (Right).

Mixed Data Training. We train MD2 and FM with mixed daytime and nighttime data from nuScenes and report the results (*MonoDepth2 (Mix)* and *FM (Mix)*) in Table. 2. These two methods achieve better performance than their baselines but still keep a large gap to *Our*.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
MonoDepth2	1.185	42.306	21.613	1.570	0.184	0.360	0.504
FM	1.138	41.617	20.848	1.148	0.238	0.425	0.5650
MonoDepth2 (Mix)	1.070	38.336	20.117	1.191	0.269	0.451	0.586
FM (Mix)	0.956	34.052	18.794	0.798	0.305	0.507	0.652
Our	**0.315**	**3.793**	**9.641**	**0.403**	**0.508**	**0.778**	**0.896**

Table 2. Quantitative results on mixed daytime and nighttime data of nuScenes. Baseline methods are underlined.

References

[1] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NIPS*, volume 32, 2019.

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[4] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020.

[5] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*, 36(1):3–15, 2017.

[6] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, pages 572–588. Springer, 2020.

[7] Madhu Vankadari, Sourav Garg, Anima Majumder, Swagat Kumar, and Ardhendu Behera. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *ECCV*, pages 443–459. Springer, 2020.

[8] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.