Supplementary Material
Multi-Frame GAN: Image Enhancement for Stereo Visual Odometry in Low Light

Eunah Jung[1,2*] **Nan Yang**[1,2*] **Daniel Cremers**[1,2]
[1]Technical University of Munich [2]Artisense

1 Introduction

In this supplementary material, we firstly show the details of the network architecture of MFGAN. Then, we introduce more information regarding how Oxford RobotCar dataset [1] is used to evaluate our method and present the qualitative results of flow estimation on different light environment. Next, we show the experiments of New Tsukuba dataset [2] and additional results on Oxford RobotCar dataset. Additionally, we also provide a supplementary video to demonstrate the performance of MFGAN for the frame consistency as well as the improvement for stereo VO methods.

2 Network Details

The detailed architecture for the encoder and decoder of the generators is shown in Figure 1 and Figure 2, respectively. We extend the architecture of [3] to take into account temporal consistency and stereo image pairs. During training, the networks share the weights in Siamese networks fashion [4] to push consistency over frames and optimize the parameters all together after forwarding two temporally neighboring stereo pairs. Note that we do not separately train the networks for optical flow estimation but use the supervision of the estimated optical flow only in the training phase.

3 Details of using Oxford RobotCar Dataset

The Oxford RobotCar dataset [1] contains a large amount of data collected while traversing approximately 10km in central Oxford, UK for a year in different time slots and weathers. We divide the entire sequence into 10 sub-sequences where each sequence is around 700m distance and has certain characteristics of the trajectories. These sub-sequences are shown on Oxford map in Figure 3. Most of sequences contain one or more curves, Seq. 3 has difficult U-turn route, and Seq. 6 is straight-shaped road. For training, we split the sequences such that they are equally distributed on the entire map. We use Seq. 0, 2, 3, 5, 8 as training set and Seq. 1, 4, 6, 7, 9 as testing sequences. Note that there are no overlaps between the training set and the test.

We take the same sequences from different conditions using GPS/INS data from the Oxford RobotCar dataset. First, we fix the start point of each sequence according to GPS location from Day set and take the corresponding sequence from Night set using the fixed GPS position. Then, we set the end of each sequence by measuring the same distance. To generate ground-truth poses for the evaluation of VO methods, since the timestamps of frames are not synchronized with GPS/INS data, we interpolated the pose of timestamps based on GPS/INS data. In addition, we cut the car head part fixed at the bottom in the frames because the fixed objects unrelated with the scene cause difficulties to run VO methods.

*These two authors contributed equally. Correspondence to: {jungeu,yangn}@in.tum.de

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.

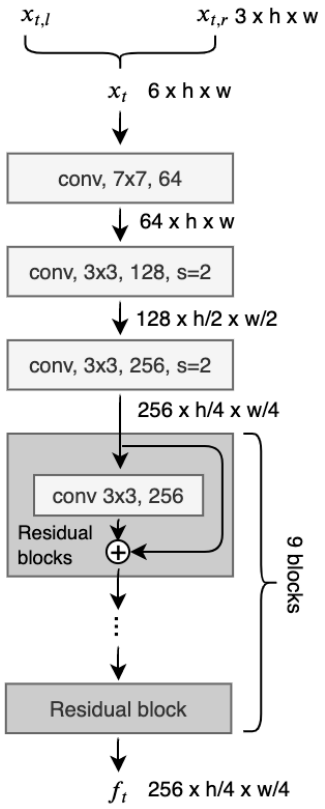


Figure 1: Encoder architecture of the generator network. The encoder takes the concatenated stereo image pair at the same timestamp and generates the corresponding feature. The convolutional layers are shown with the kernel size and the number of output channels followed by 9 residual blocks.

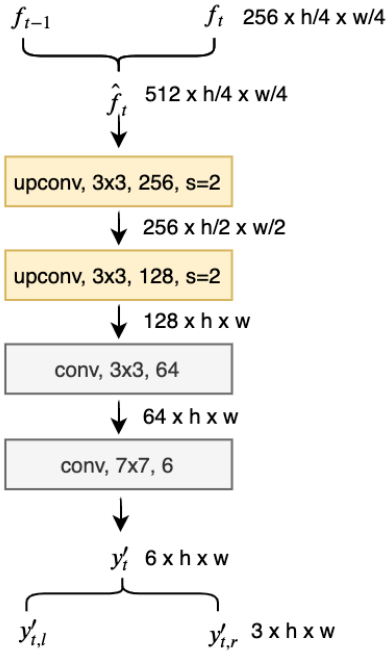


Figure 2: Decoder architecture of the generator network. We generate the features from two consecutive temporal stereo pairs and concatenate the features to feed into the decoder part. Transposed convolutional layers for upsampling are presented in yellow color.

Figure 3: The entire route of the Oxford RobotCar dataset (left image) is represented in green route and 10 sub-sequences taken from the whole trajectory (right image) are marked in yellow color respectively.

4 Flow Estimation on Different Lighting Images

We check that FlowNet2 in general shows good performance for not only bright lighting but also dark illumination images. The predicted optical flow on Day and Night scene of the Oxford RobotCar dataset is shown in Figure 4 with a warped image based on the flow. Additionally, the evaluation results on the main paper show that, with the estimation quality delivered by FlowNet2, MFGAN is able to generate temporal as well as stereo consistent sequences.

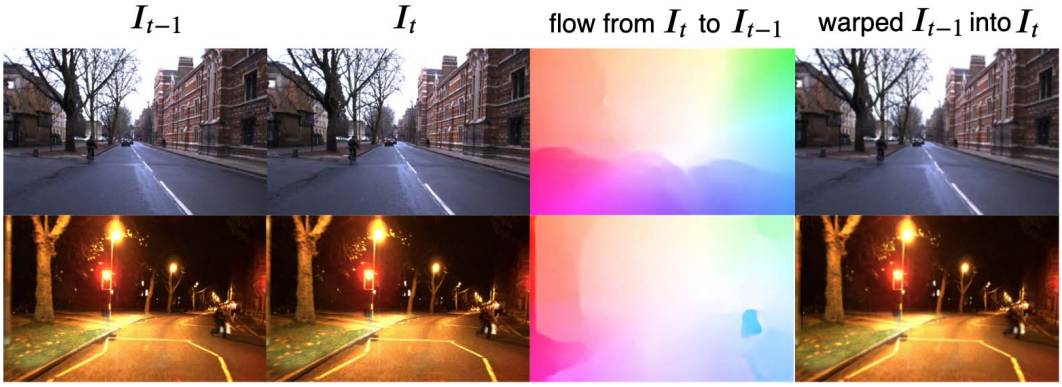


Figure 4: Flow estimation and warped images of Day and Night scene in the Oxford RobotCar dataset. The first row is made of images of Day set and the second row consists of images from Night set. The estimated flows show the fair quality of warped images leading to reasonable supervision for consistency.

5 More Experiments

5.1 New Tsukuba Dataset

This synthetic dataset [2] captures static office scene. The dataset contains 1800 stereo image pairs with ground-truth camera pose, disparity maps, occlusion maps and discontinuity maps. The stereo camera travels the fixed trajectory under different lighting such as Daylight, Fluorescent, Lamps and Flashlight, and the camera poses demonstrate strong rotation change. We test our method on one pair of unpaired sets, *Fluorescent* and *Flashlight*, and used 1000 images as a training set and 800 images for the evaluation. Although the images in *Fluorescent* and *Flashlight* set are paired, we utilize them in unpaired setting when training.

The experiments about frame consistency and VO performance on test sequences are shown in Table 1 and 2, respectively. The enhanced sequence by the model cy, tmp trained with temporal consistency as well as cycle consistency shows better frame consistency and improves the VO performance of both direct and indirect methods compared to when using original *Flashlight* and the cy outputs. Further adding the stereo consistency did not improve the results on this dataset.

	cy	cy, tmp
median	0.75	**0.54**
mean	2.20	**1.68**

Table 1: EPE E_{tmp} of the New Tsukuba dataset for temporal consistency.

	Flashlight		cy		cy, tmp	
	t_{abs}	r_{abs}	t_{abs}	r_{abs}	t_{abs}	r_{abs}
DSO	2.055	39.35	0.194	2.34	**0.144**	**2.06**
ORB	X	X	0.284	8.63	**0.239**	**7.86**

Table 2: Evaluation on the New Tsukuba dataset. $t_{abs}(m)$ and $t_{abs}(^\circ)$ are absolute translation and rotational RMSE. The image translation model cy allows stereo VO methods outperform on dark scene, and adding temporal consistency cy, tmp leads to even better performance.

Note that we measure the absolute trajectory error because the test sequence is short, i.e., less than 50m.

5.2 Oxford RobotCar Dataset

The estimated trajectories of Seq. 1, 6, 7 by Stereo DSO are shown in the left side of Figure 5, 6 and 7, and those by stereo ORB-SLAM are shown in the right side of the figures.

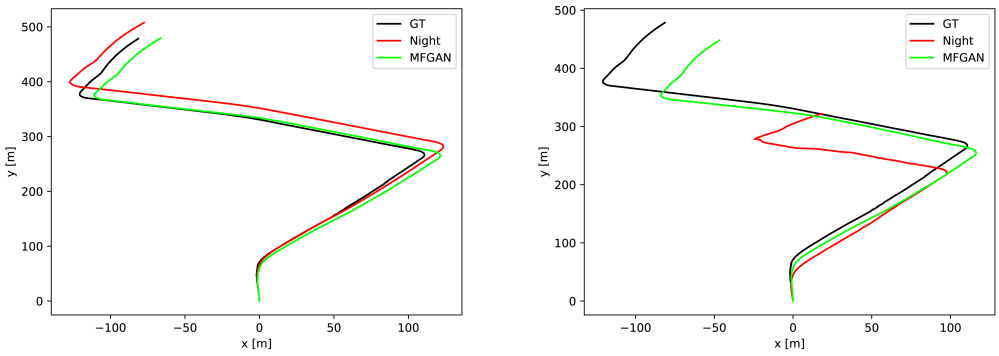


Figure 5: Seq. 1 of Oxford RobotCar dataset from Stereo DSO (left) and stereo ORB-SLAM (right).

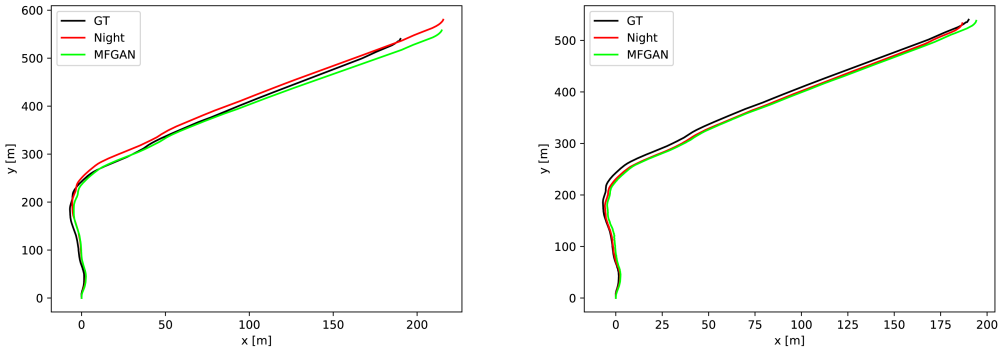


Figure 6: Seq. 6 of Oxford RobotCar dataset from Stereo DSO (left) and stereo ORB-SLAM (right).

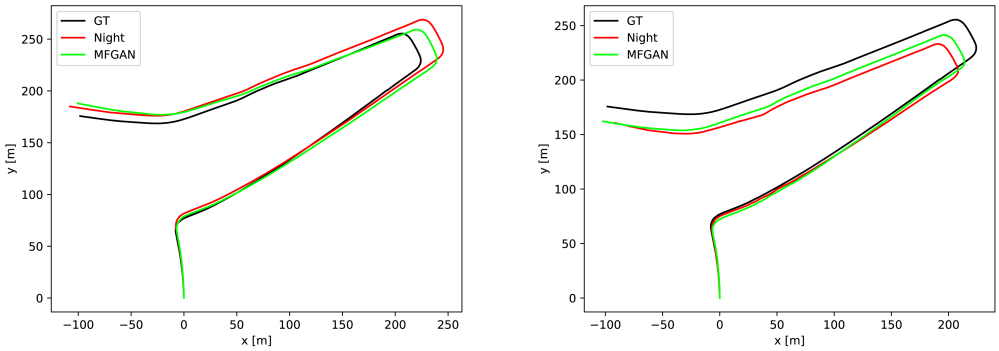


Figure 7: Seq. 7 of Oxford RobotCar dataset from Stereo DSO (left) and stereo ORB-SLAM (right).

5.3 Stereo Visual Odometry with Different Lighting

We evaluate the VO performance using both direct and indirect methods on different lighting conditions, i.e., a bright optimal scene and a dark challenging scene without any image enhancement.

With New Tsukuba dataset, the results on Fluorescent and Flashlight are shown in Table 3. It is clearly observed that that Stereo DSO and stereo ORB-SLAM on Fluorescent scene deliver accurate as well as robust performance compared to on Flashlight scene e.g., losing tracking.

	Fluorescent		Flashlight	
	t_{abs}	r_{abs}	t_{abs}	r_{abs}
DSO	**0.092**	**1.30**	2.055	39.35
ORB	**0.177**	**5.61**	X	X

Table 3: Evaluation on the New Tsukuba dataset. X indicates the tracking is lost.

With Oxford RobotCat dataset, we evaluate the test sequences of a Day set, and the results are shown in Table 4 with the results on Night scene. As proposed in the KITTI Odometry Benchmark [**?**], we evaluate the relative translational error (t_{rel}) and relative rotational error (r_{rel}) as a function of trajectory length. Specifically, the metrics are defined as

$$t_{rel}(\mathcal{F}) = \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} ||(\hat{\mathbf{p}}_j \ominus \hat{\mathbf{p}}_i) \ominus (\mathbf{p}_j \ominus \mathbf{p}_i)||_2 \tag{1}$$

$$r_{rel}(\mathcal{F}) = \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} \angle[(\hat{\mathbf{p}}_j \ominus \hat{\mathbf{p}}_i) \ominus (\mathbf{p}_j \ominus \mathbf{p}_i)] \tag{2}$$

where $\mathcal{F}$ is a set of frames (i, j), $\hat{\mathbf{p}} \in SE(3)$ and $\mathbf{p} \in SE(3)$ are estimated and true camera poses, respectively, $\ominus$ denotes the inverse compositional operator and $\angle[\cdot]$ is the rotation angle.

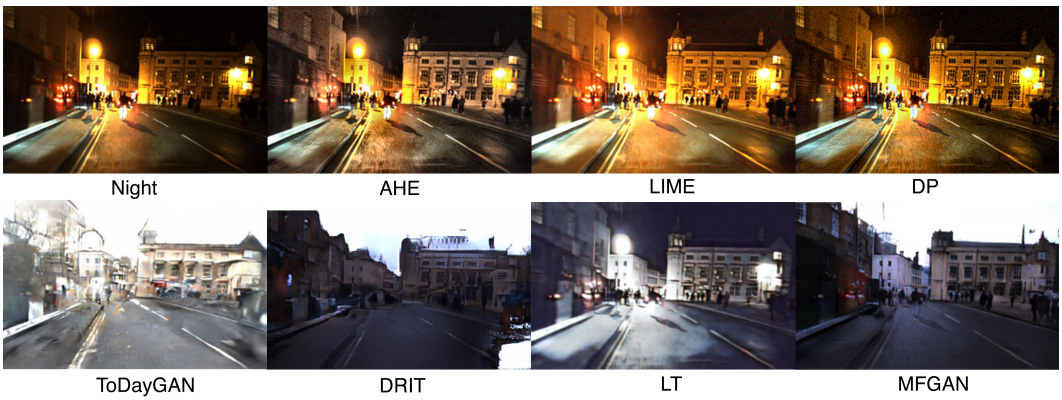
The results show that a Day scene with bright illumination gives more accurate and reliable performance of stereo VO methods than a Night scene.

Seq.		Day		Night	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	DSO	7.86	**1.71**	**7.16**	2.91
	ORB	**7.16**	**2.10**	X	4.80
04	DSO	**7.13**	**2.01**	24.78	5.28
	ORB	**3.78**	**1.90**	X	11.00
06	DSO	**9.07**	**0.71**	9.86	0.87
	ORB	5.73	**0.61**	**5.52**	0.86
07	DSO	**6.07**	**1.79**	6.38	2.38
	ORB	**3.52**	**2.26**	6.35	2.58
09	DSO	**5.32**	**2.07**	7.87	4.96
	ORB	**4.27**	**2.20**	14.16	9.21
mean	DSO	**7.09**	**1.80**	11.21	3.28
	ORB	**4.89**	**1.81**	16.94	5.69

Table 4: Comparison of Day and Night sequences of Oxford RobotCar dataset. $t_{rel}(\%)$ and $r_{rel}(^\circ)$ are the relative translational and rotational errors [5]. Note that although the corresponding Day and Night sequences share similar trajectories, there are no 1-to-1 paired images. Overall, both VO systems deliver more accurate results on Day sequences. For the sequences where Night is better than Day, our observation is that these Night sequences are with fairly good lighting conditions and contain less dynamic objects than the corresponding Day sequences.

5.4 Qualitative Results of Other Methods

In the paper, we compare MFGAN with other methods to evaluate VO performance. The example frames of Oxford RobotCar dataset generated by other methods are shown in Figure 8. We tested the enhanced sequence by photo enhancement methods: adaptive histogram equalization(AHE) [6], low-light image enhancement(LIME) [7], and deep photo enhancer(DP) [8]. The frames translated by style transfer methods such as ToDayGAN [9], DRIT [10], and LinearTransfer(LT) [11] are presented as well.



(a)

(b)

Figure 8: Enhanced frame examples. The top left image is the original Night image and the other images in the first row are generated from photo enhancement methods. On the second row, the first three images are translated by other style transfer methods, and the last image is by MFGAN.

References

[1] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[2] S. Martull, M. Peris, and K. Fukui. Realistic CG stereo image dataset with ground truth disparity maps. In *ICPR workshop TrakMark2012*, volume 111, pages 117–118, 2012.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[4] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[6] K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.

[7] X. Guo. LIME: A method for low-light image enhancement. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 87–91. ACM, 2016.

[8] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6306–6314, 2018.

[9] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. *arXiv preprint arXiv:1809.09767*, 2018.

[10] H. Y. Lee, H. Y. Tseng, J. B. Huang, M. K. Singh, and M. H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[11] X. Li, S. Liu, J. Kautz, and M.-H. Yang. Learning linear transformations for fast arbitrary style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.