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

Abstract: We propose the concept of a multi-frame GAN (MFGAN) and demonstrate its potential as an image sequence enhancement for stereo visual odometry in low light conditions. We base our method on an invertible adversarial network to transfer the beneficial features of brightly illuminated scenes to the sequence in poor illumination without costly paired datasets. In order to preserve the coherent geometric cues for the translated sequence, we present a novel network architecture as well as a novel loss term combining temporal and stereo consistencies based on optical flow estimation. We demonstrate that the enhanced sequences improve the performance of state-of-the-art feature-based and direct stereo visual odometry methods on both synthetic and real datasets in challenging illumination. We also show that MFGAN outperforms other state-of-the-art image enhancement and style transfer methods by a large margin in terms of visual odometry.

Keywords: Visual Odometry, Style Transfer, Generative Adversarial Network

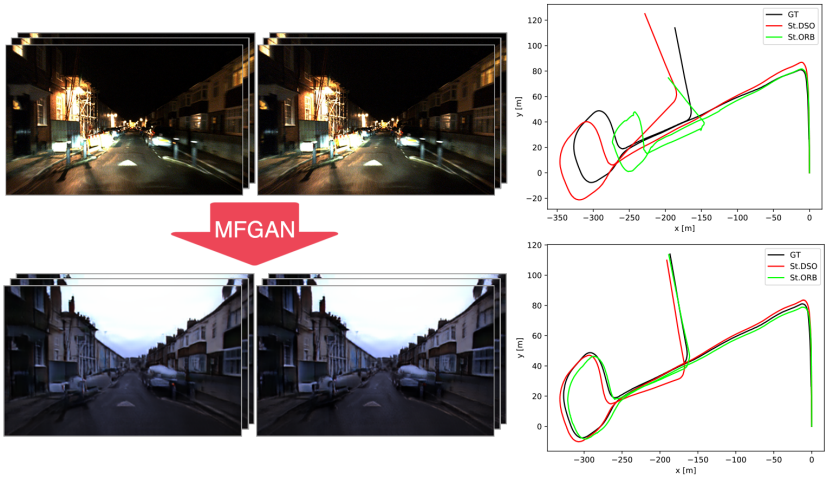


Figure 1: We propose Multi-Frame GAN (MFGAN) for stereo VO in challenging low light environment. The MFGAN takes two consecutive stereo image pairs and outputs the enhanced stereo images while preserving temporal and stereo consistency. On the right side, the estimated trajectories by the state-of-the-art stereo feature-based VO method Stereo ORB-SLAM and the state-of-the-art direct VO method Stereo DSO are presented. Due to the low image gradient, dynamic lighting and halo, Stereo DSO and Stereo ORB-SLAM cannot achieve good tracking accuracy in the night scene. With the translated images from MFGAN, the performance of both methods is notably improved.

1 Introduction

Visual odometry (VO) and simultaneous localization and mapping (SLAM) have been actively studied due to their wide usage in robotics, AR/VR and autonomous driving. Particularly, stereo

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

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

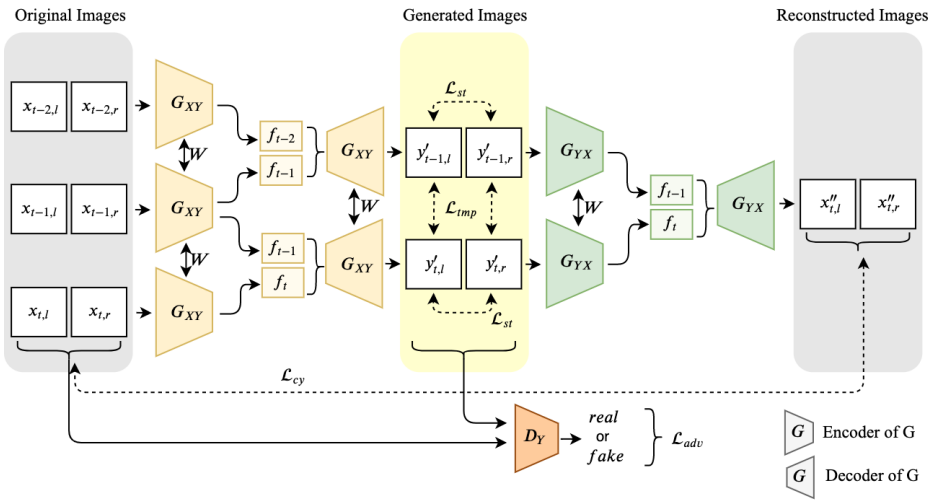


Figure 2: Overview of MFGAN. Only the forward cycle with G_{XY} and D_Y is shown for the simplicity. The encoder of the generator takes a stereo image pair at each timestamp and the decoder part takes the concatenated feature maps from previous and current frames to output the current enhanced images. During training phase, the adversarial loss $\mathcal{L}_{adv}$ is computed by using discriminator network and the cycle consistency loss $\mathcal{L}_{cy}$ compares the original and reconstructed images. The network generates two consecutive fake frames and computes the temporal consistency loss $\mathcal{L}_{tmp}$ between different times and the stereo consistency loss $\mathcal{L}_{st}$ between stereo image pairs.

VO[1, 2] delivers more reliable and accurate results than monocular systems[3, 4, 1] by eliminating the scale ambiguity[5, 6, 7].

However, tracking the camera pose in poorly-lit conditions, such as night driving scenes, which is crucial for autonomous driving, is still a challenge for current stereo systems. Both, feature-based methods [8, 1] and direct methods [4, 3] rely on image gradient-based key point extraction, which provides fewer high-quality points in low-light scenes. Moreover, with dynamic lighting and halo that are abundant in dark scenes, it is difficult to track the same key points as the illumination is changing. To overcome these limitations, several approaches have recently been presented, e.g., camera exposure time control [9, 10], camera model optimization [11, 10], and robust feature descriptors [12, 13]. Yet, these approaches either need paired datasets for training or address only one of many aspects for the challenging night scene.

In this paper, we propose a learning-based sequence enhancement method for stereo VO methods, named Multi-Frame GAN (MFGAN). MFGAN makes use of Generative Adversarial Networks (GANs) to perform a domain transfer from bad illumination to good illumination. In this way, manual engineering for different aspects causing VO failure at night is avoided. Based on CycleGAN [14], we make use of *unpaired* data for training, thus avoiding substantial costs for pairing or labeling in real-world applications. Yet, CycleGAN transforms images independently whereas for stereo VO we need to preserve spatial (i.e., inter-camera) and temporal consistency of the domain transfer. To this end, we carefully design the temporal and stereo consistency loss terms leveraging optical flow in order to ensure consistency of the transformed brightness across cameras and in time.

We validate our approach with state-of-the-art VO methods on the synthetic indoor New Tsukuba dataset [15] and the challenging outdoor Oxford RobotCar datatset [16] which contains various illumination situations. Specifically, we use a direct method, Stereo DSO [2], and a feature-based method, Stereo ORB-SLAM [1]. To the best of our knowledge, this is the first work exploring the potential of GAN-based image translation for VO. And the experiments show that our method leads to significant improvement in accuracy and robustness for both direct as well as indirect methods. We also compare MFGAN with other photo enhancement [17, 18, 19] and image/video translation [20, 21, 22] methods. The results show the superiority of MFGAN for improving VO in the challenging lighting condition.

2 Related Work

Robust Visual Odometry (VO). Feature-based methods [1, 8] rely on feature matching and estimate the camera poses by minimizing the re-projection error. Direct methods [4, 3] do not rely on feature descriptors and directly optimize the photometric error. To improve the performance of VO in challenging lighting conditions, Pascoe et al. [13] proposed a direct monocular SLAM algorithm using a newly designed metric considering entropies in the frame instead of intensities. Alismail et al. [12] introduced a binary feature descriptor for direct VO methods for poor light environment. While the classical vision methods have been actively researched, there are few learning-based methods. Gomez et al. [23] trained a neural network with LSTM units using synthetic paired datasets to produce enhanced images and evaluated their method on real-world static scene. Compared to their method, we make use of unpaired datasets and explicitly address temporal and spatial coherence using optical flow. We validate our method on challenging synthetic as well as real-world datasets.

Image and Video Translation. Gatys et al. [24] used the pretrained networks to capture the content and style respectively and optimize texture transfer. Image translation using Generative Adversarial Networks (GANs) [25, 14, 26] has become very popular due its superior performance. Isola [25] proposed a conditional adversarial networks with paired samples in different styles and this work is extended to CycleGAN by Zhu et al. [14] who suggested two pairs of unconditionally trained adversarial networks. One of the applications of image translation is retrieval-based visual localization [20, 27]. They used image translation to close the domain gap for matching images from different conditions. TodayGAN [28] adapts the architecture of CycleGAN and improves the performance of image retrieval. Beyond a single image style transfer, video synthesis into other styles becomes an active topic [29, 30, 31]. Especially, synthesizing video requires temporal consistency over the contiguous frames, since estimating individual frames leads to flickering effect. Therefore, many works use an approach comparing the target image and the warped image based on the optical flow to push consistency [30, 29]. While we also utilize image-warping based on flow field, we design our consistency loss function taking into account both stereo-spatial as well as temporal context.

3 Multi-Frame GAN for Space-time Consistent Domain Transfer

Our method, Multi-Frame GAN (MFGAN), is based on the cycle-consistent network architecture proposed by Zhu et al. [14]. We extend it into a multi-frame scheme, such that MFGAN translates a given sequence of stereo images in one domain into another domain of sequence. With the proposed temporal and stereo consistency terms, MFGAN is able to generate the translated sequence preserving the coherence of the input sequence. We implement this coherence with differentiable image warping using optical flow.

Inspired by [14], MFGAN consists of two generator-discriminator-sets, $\{G_{XY}, D_Y\}$ and $\{G_{YX}, D_X\}$ where X and Y denote different image domains, e.g., X for the poor lighting and Y for good lighting condition. The generator G_{XY} translates the input images from the source domain X to the target domain Y, and the discriminator D_Y aims to distinguish between the original images from the domain Y and the translated fake images. Likewise, the other set, $\{G_{YX}, D_X\}$, functions the same but with opposite domains. We denote the part with G_{XY} and D_X as the forward cycle and the part involving G_{YX} and D_Y as the backward cycle.

The overview of our networks in the training phase is shown in Figure 2. For simplicity, we present the method regarding the forward cycle and skip the backward cycle. In the training phase, the generator G_{XY} takes *three* pairs of stereo images and generates *two* pairs of stereo images for the target domain,

$$y'_{t-1} = G_{XY}(x_{t-2}, x_{t-1}) \quad \text{and} \quad y'_t = G_{XY}(x_{t-1}, x_t). \tag{1}$$

where $x_{t-2}, x_{t-1}, x_t \in X$ are the original image pairs and $y'_{t-1}, y'_t \in Y$ are the fake image pairs. Note that a stereo pair x_t means $(x_{t,l}, x_{t,r})$ for the left and right images. The generator G_{YX} takes (y'_{t-1}, y'_t) as the input and reconstructs the images back in the source domain X,

$$x''_t = G_{YX}(y'_{t-1}, y'_t). \tag{2}$$

where $x''_t \in X$ is the reconstructed stereo image.

Adversarial Loss. The adversarial loss $\mathcal{L}_{adv}$ [32, 14, 27, 20] is defined as:

$$\begin{aligned} \mathcal{L}_{adv} &= \mathcal{L}_{gen} + \mathcal{L}_{disc}, \\ \mathcal{L}_{gen} &= (D_Y(y'_t) - 1)^2, \\ \mathcal{L}_{disc} &= (D_Y(y_t) - 1)^2 + (D_Y(y'_t))^2 \end{aligned} \tag{3}$$

The discriminator is trained to distinguish the given real and fake inputs correctly through $\mathcal{L}_{disc}$, while the generator aims to synthesize as realistic image in domain Y as possible by minimizing $\mathcal{L}_{gen}$.

Image Consistency Loss. The image consistency loss term computes how similar two images are. We use a linear combination of L1 loss and single scale SSIM [33] as the measurement:

$$\mathcal{F}(a, b) = \alpha \frac{1 - \text{SSIM}(a, b)}{2} + (1 - \alpha)|a - b|, \tag{4}$$

where α is set to 0.8. Then, the cycle consistency loss is formed using the above image similarity metric,

$$\mathcal{L}_{cy} = \mathcal{F}(x_t, x''_t) \tag{5}$$

where x''_t is from Equation 2. This loss term resolves unconstrained difficulties due to unpaired datasets by reconstructing back the generated images into the original image domain.

Additionally, we introduce the temporal consistency loss $\mathcal{L}_{tmp}$ and the stereo consistency loss $\mathcal{L}_{st}$ to jointly optimize the image coherence over the multiple temporal and stereo-spatial frames. For temporally neighboring frames, we warp the images y'_{t-1} into y'_t using the estimated optical flow W^t_{t-1}, and for stereo frames warp the right image $y'_{t,r}$ into $y'_{t,l}$ similarly making use of the estimated optical flow W^l_r. The temporal consistency loss $\mathcal{L}_{tmp}$ and the stereo consistency loss $\mathcal{L}_{st}$ are then formed as below,

$$\mathcal{L}_{tmp} = \mathcal{F}(\omega^t_{t-1}(y'_{t-1}), y'_t) \tag{6}$$

$$\mathcal{L}_{st} = \mathcal{F}(\omega^l_r(y'_{t,r}), y'_{t,l}) \tag{7}$$

where $\omega^t_{t-1}(\cdot)$ is the backward warping function using the optical flow W^t_{t-1}, and $\omega^t_{t-1}(x_{t-1})$ gives the warped images of x_{t-1} into x_t.

Finally, the total loss $\mathcal{L}$ integrates the adversarial loss, cycle consistency loss, temporal consistency as well as stereo consistency loss,

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{cy}\mathcal{L}_{cy} + \lambda_{tmp}\mathcal{L}_{tmp} + \lambda_{st}\mathcal{L}_{st} \tag{8}$$

where λ is the weight for each corresponding loss term. In summary, this loss function extends the cycle loss of [14] by a temporal and a stereo consistency loss thereby assuring that the resulting domain transfer preserves a spatio-temporal regularity. Note that we compute $\mathcal{L}_{adv}$, $\mathcal{L}_{cy}$ and $\mathcal{L}_{tmp}$ for the left and right image in a stereo pair correspondingly and each loss term includes both translation direction between generators, e.g. from the domain X into the domain Y and vice versa.

The generator networks contain down-sampling and up-sampling convolutional layers in an encoder-decoder scheme. The feature maps of the temporal consecutive two neighbor frames are channel-wise concatenated and then fed into the decoder. For the discriminators, we adopt the PatchGAN architecture [14, 25] and exploit its usage through the loss terms. Please refer to the supplemental material for the detailed architecture.

4 Experiments

We evaluate MFGAN on the synthetic indoor New Tsukuba dataset [15] as well as the real outdoor Oxford RobotCar dataset [16]. To evaluate the frame consistency, we propose a new metric using the optical flow between frames and discuss the consistency quantitatively and qualitatively. For the evaluation of VO, we validate the performance of two state-of-the-art feature-based and direct stereo VO methods, namely Stereo ORB-SLAM [1] and Stereo DSO [2], respectively. We run the two VO methods on both original sequences with bad lighting conditions (Flashlight for New Tsukuba dataset and Night for Oxford RobotCar dataset) and the corresponding translated good lighting conditions (*Fluorescent* for New Tsukuba dataset and *Day* for Oxford RobotCar dataset)

	cy		cy, tmp		MFGAN	
	E_{tmp}	E_{st}	E_{tmp}	E_{st}	E_{tmp}	E_{st}
01	1.64	2.19	1.03	1.21	1.14	0.73
04	2.05	2.68	1.23	1.38	1.3	0.76
06	1.37	1.81	0.95	1.15	1.01	0.74
07	1.76	2.14	1.18	1.29	1.26	0.82
09	1.92	2.46	1.16	1.35	1.2	0.83

Table 1: EPE E_{tmp} and E_{st} of the Oxford RobotCar dataset for frame consistency.

with MFGAN. Both, frame consistency as well as VO performance are significantly improved by MFGAN on both datasets. In this section, we introduce the extensive evaluation results on Oxford RobotCar dataset. Please refer to our supplementary materials for the evaluation on New Tsukuba dataset.

Oxford RobotCar dataset [16] provides a massive amount of data collected while driving an approximately 10km route over 1 year in different time slots. The dataset recorded almost 20 million images from 6 cameras mounted on the car, with LIDAR, GPS, INS ground truth, and includes the data in different weather conditions, seasons and daytimes, e.g. summer, winter, rain, night, of the same trajectory. We select Day and Night scene where Day is the overcast dataset 2015/02/10 and Night is the night-tagged dataset 2014/12/16 to train MFGAN and evaluate the VO performance. Specifically, we choose 10 sub-sequences around 700m long from the entire route, such that each sub-sequence includes several characteristics like multiple corners, straight-shaped route. Please refer to our supplementary materials for the locations of each sub-sequences. We use the Seq. 00, 02, 03, 05, 08 as the training set and Seq. 01, 04, 06, 07, 09 as the testing set. Note that there is no overlap segments between training sequences and the testing sequences. Approximately 9000 frames are used for training MFGAN. The train and test set are split to be geographically equally distributed.

In order to compute the consistency loss functions based on the warped images, we use the predicted optical flow by the state-of-the-art flow estimation network FlowNet2 [34]. We measure temporal optical flow and stereo spatial optical flow to compute the total consistency loss function. To be specific, we use the optical flows predicted in one domain, for example, domain X, to warp the generated images in another domain e.g. Y, and check the consistency and vice versa for the other direction. We implement MFGAN with PyTorch and train with batch size 1, 15 epochs and Adam optimizer. The learning rate remains 0.0002 for the first 10 epochs and decays linearly to zero over the next 5 epochs. The weights for each loss term in Equation 8 are set as λ_{adv} to 1.0, λ_{cy} to 10.0, λ_{tmp} to 3.0 and λ_{st} to 3.0. With the image resolution 320x192, MFGAN shows 111 FPS inference performance with NVIDIA GeForce GTX 1080.

In the following, we show the evaluation results on the testing set. For the simplicity, we use cy for the model trained with cycle consistency loss, cy, tmp for the model cy with additional temporal consistency loss, cy, st for the model cy with additional stereo consistency loss, and MFGAN means the proposed model trained with cycle, temporal and stereo consistency loss.

4.1 Frame Consistency

In this section, we evaluate the visual consistency over multiple contiguous frames. We investigate this in terms of both temporal and stereo consistency.

Quantitative results. We introduce a metric using optical flow between frames as the optical flow is measured by matching the corresponding points of two frames by their appearance. Under the assumption that the optical flow between two frames has no flaw, the optical flow W_t^{t+2} from t-th frame to $(t+2)$-th frame equals the addition of the optical flow W_t^{t+1} and W_{t+1}^{t+2} if three contiguous frames at $t, t+1, t+2$ timestamps are temporally consistent in image appearance.

Therefore, we measure the temporal consistency E_{tmp} by the endpoint error [34] as below:

$$EPE(W_t^{t+2}, W_t^{t+1} \oplus W_{t+1}^{t+2}) \tag{9}$$

where $\oplus$ means the addition of two optical flows. EPE indicates the end point error which is used to measure the error of two optical flows as $\|W_1 - W_2\|_2$ where W_1 and W_2 are flows. The optical flow addition is done by adding the first operand to the sampled second operand based on the first optical flow.

Seq.		Night		cy		cy, tmp		cy, st		MFGAN	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	DSO	7.16	2.91	X	X	X	X	10.00	2.26	**5.41**	**2.18**
	ORB	X	4.80	21.14	2.97	78.94	26.10	81.73	26.00	**11.06**	**2.75**
04	DSO	24.78	5.28	X	X	X	X	9.76	**2.89**	**4.75**	3.20
	ORB	X	11.00	82.73	26.22	74.70	30.74	73.03	33.43	**4.59**	**3.89**
06	DSO	9.86	**0.87**	X	X	X	X	19.11	1.58	**8.08**	0.88
	ORB	**5.52**	**0.86**	74.50	38.13	X	X	15.07	2.70	5.63	1.34
07	DSO	6.38	2.38	6.46	**2.29**	6.34	2.38	6.21	2.34	**4.55**	2.36
	ORB	6.35	**2.58**	94.40	9.48	67.10	30.52	65.98	41.08	**4.92**	2.82
09	DSO	7.87	4.96	X	X	X	X	12.78	**2.83**	**5.57**	3.16
	ORB	14.16	9.21	67.10	42.44	X	X	33.35	24.50	**5.39**	**3.78**
mean	DSO	11.21	3.28	X	X	X	X	12.35	2.44	**5.67**	**2.36**
	ORB	16.94	5.69	76.32	32.15	73.58	29.12	53.83	25.54	**6.83**	**2.92**

Table 2: Evaluation on the test sequences from the Oxford RobotCar dataset. $t_{rel}(\%)$ and $r_{rel}(^\circ)$ are the relative translational and rotational errors [35], respectively. X means lost tracking and the sequences which lose tracking are not used for calculating the mean. Overall, MFGAN improves both Stereo DSO and Stereo ORB-SLAM in terms of average t_{rel}. MFGAN also shows the superior results to other variants of the models.

(a) (b)

Figure 3: Qualitative results on temporal consistency (a) and stereo consistency (b) for the Oxford RobotCar dataset. Each row shows from top to bottom original Night images, the outputs from cy, cy, tmp, and MFGAN. (a) Left: t-th frame. Center: $(t+1)$-th frame. Right: $(t+2)$-th frame. (b) Left: left image of stereo. Right: right image of stereo. MFGAN presents more coherent appearance regarding temporal as well as stereo consistency. Please refer to our supplementary video for clearer demonstration.

Likewise, we compute the stereo and temporal consistency E_{st} as below:

$$EPE(W_{t,r}^{t+1,r} \oplus W_{t+1,r}^{t+1,l}, W_{t,r}^{t,l} \oplus W_{t,l}^{t+1,l}) \tag{10}$$

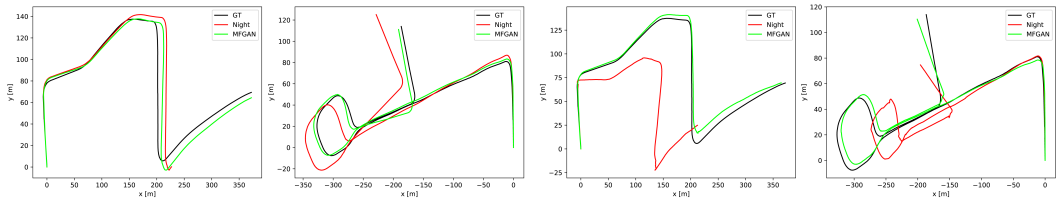
where l, r means right and left side of stereo image pair respectively.

The results of the temporal and stereo-spatial consistency is shown in Table 1. While the model cy, tmp gives comparably low EPE for temporal consistency E_{tmp}, it gives high error values for temporal and stereo consistency E_{st}. On the other hand, MFGAN, which is trained with both temporal and stereo consistency, shows lower EPE than cy, tmp. Overall, this shows that cy, tmp is more consistent in terms of temporally contiguous frames but considering stereo sequences, MFGAN is more consistent for the entire stereo sequences.

Qualitative results. The qualitative results of temporal consistency converting the Night scene to Day scene of the Oxford RobotCar dataset are shown in Figure 3a. While the generated outputs from the model cy give fluctuating artifacts, cy, tmp and MFGAN presents consistent appearance over the contiguous frames, especially on the areas marked in red rectangle. The comparison for

Seq.		AHE [17]		LIME [18]		DP [19]		MFGAN	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	DSO	7.82	1.84	7.80	1.87	7.66	**1.79**	**5.41**	2.18
	ORB	46.85	10.43	38.07	3.96	53.69	12.97	**11.06**	**2.75**
04	DSO	7.22	4.40	6.59	**2.75**	6.49	2.92	**4.75**	3.20
	ORB	35.00	12.40	27.43	10.31	40.77	17.41	**4.59**	**3.89**
06	DSO	10.08	**0.73**	9.77	0.80	9.63	0.79	**8.08**	0.88
	ORB	6.56	1.40	**5.69**	**0.89**	34.87	0.93	5.67	1.34
07	DSO	6.46	**2.29**	6.34	2.38	6.21	2.34	**4.55**	2.36
	ORB	11.91	3.06	5.50	2.57	24.48	**2.28**	**4.92**	2.82
09	DSO	6.85	3.22	6.51	2.82	6.49	**2.73**	**5.57**	3.16
	ORB	24.97	13.81	17.32	13.20	24.39	4.11	**5.39**	**3.78**
mean	DSO	7.69	2.50	7.40	**2.12**	7.30	**2.12**	**5.67**	2.36
	ORB	25.06	8.22	18.80	6.19	35.64	7.54	**6.83**	**2.92**

Table 3: Comparison with other photo enhancing methods.



(a) Stereo DSO

(b) Stereo ORB-SLAM

Figure 4: Results of Stereo VO methods on Seq. 4 and 9 of the Oxford RobotCar dataset. Clearly MFGAN improves the performance of both Stereo DSO and Stereo ORB-SLAM. Please refer to our supplementary materials the trajectories of other sequences.

stereo consistency for the same trained model is presented in Figure 3b. The stereo image pair from MFGAN delivers consistent image appearance. Please refer to our supplementary video for clearer demonstration.

4.2 Stereo Visual Odometry

We run Stereo DSO and stereo ORB-SLAM 5 times for each method and each original day sequences (Day), the original night sequences (Night) and the generated Day sequences (MFGAN). We use the *median* relative translational error $t_{rel}(\%)$, and relative rotational error $r_{rel}(^\circ)$ as proposed in the KITTI Odometry benchmark [35] as the evaluation metric. Please refer to our supplementary materials for the formulas of calculating t_{rel} and r_{rel}. We run Stereo DSO with the default settings on all sequences and Stereo ORB-SLAM with lower FAST corner thresholds on Night due to low gradient magnitude of dark scenes. Both methods show in general good performance on Day but declined performance on Night. With the translated images from MFGAN, both VO methods are significantly improved. Please refer to our supplementary materials for the evaluation results on Day.

As shown in Table 2, both Stereo DSO and Stereo ORB-SLAM are improved on MFGAN on all the sequences except for Seq. 06 for which Stereo ORB-SLAM can already deliver very accurate tracking on Night. The results from cy, cy, tmp, cy, st and MFGAN show the effectiveness of the proposed consistency loss terms and only with our full approach, MFGAN, the performance is improved consistently. Overall Stereo DSO delivers better results than Stereo ORB-SLAM on Night, and the improvement from MFGAN is more significant for Stereo ORB-SLAM. The results of Seq. 01 and 04 on which Stereo ORB-SLAM loses tracking on Night show that MFGAN improves the robustness of Stereo ORB-SLAM. With the day-ification from MFGAN, the *quantity* of tracked points for each frame increase from 106 to 289, which improves the robustness of ORB-SLAM. The tracking accuracy of Stereo DSO is improved on MFGAN, since the brightness is more consistent by removing the active lighting / halo while preserving the coherence of consecutive frames, which reduces the amount of outliers for the photometric error minimization. For ORB-SLAM, on Seq. 07 and 09, the tracking of Stereo ORB-SLAM does not fail on Night and the amounts of tracked points are similar for Night and MFGAN – in average, 289 and 293 points are tracked for each frame, respectively. Therefore, we also measure the *quality* of the features by taking the average re-projection residuals of the local map points for each frame when the pose bundle adjustment is finished. Please refer to [1] for the details of the optimization. The histograms of the residuals for

Seq.		ToDayGAN [20]		DRIT [21]		LT [22]		MFGAN	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	DSO	77.6	9.76	82.57	10.18	7.22	**2.08**	**5.41**	2.18
	ORB	82.97	26.22	X	X	11.78	2.92	**11.06**	**2.75**
04	DSO	74.69	15.19	74.71	11.40	6.05	3.23	**4.75**	**3.20**
	ORB	75.13	38.19	X	X	12.62	7.41	**4.59**	**3.89**
06	DSO	13.90	2.61	7.73	1.18	9.46	0.97	**8.08**	**0.88**
	ORB	97.76	9.55	98.69	10.07	5.85	**0.80**	**5.67**	1.34
07	DSO	21.47	3.55	67.55	5.99	X	X	**4.55**	**2.36**
	ORB	67.80	42.44	68.30	42.97	X	X	**4.92**	**2.82**
09	DSO	36.20	10.60	29.79	12.83	6.23	3.20	**5.57**	**3.16**
	ORB	62.84	44.51	62.60	45.61	10.77	7.78	**5.39**	**3.78**
mean	DSO	44.77	8.34	52.47	8.32	7.85	**2.08**	**5.67**	2.36
	ORB	77.30	32.18	76.53	32.88	8.31	4.29	**6.83**	**2.92**

Table 4: Comparison with other style transfer methods.

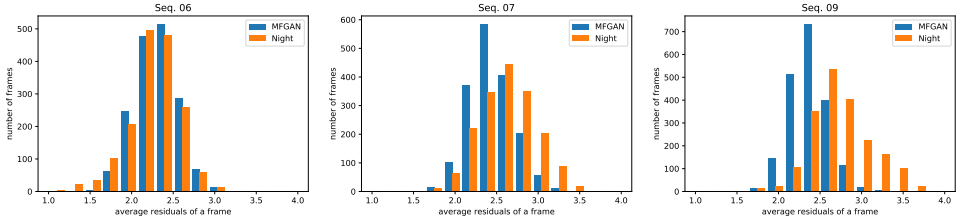


Figure 5: Histogram of the average frame residuals.

Seq. 07 and 09 are shown in Figure 5. We can see that on MFGAN Stereo ORB-SLAM can deliver more points with lower re-projection residuals and less points with higher residuals, which improves the tracking accuracy. The higher average residuals from Night indicates that there are more wrong matches due to the active lighting or inaccurate matches caused by less reliable descriptors due to the dark scene with less image gradients. As a comparison, we also show the histogram of Seq. 06 in Figure 5.

In Table 3, we compare MFGAN with other photo enhancing methods including adaptive histogram equalization(AHE) [17], low-light image enhancement(LIME) [18], deep photo enhancer(DP) [19]. We use the Matlab implementation for AHE and LIME, and the pre-trained model for DP. From the table we can see that MFGAN is able to deliver consistent improvement for both Stereo DSO and Stereo ORB-SLAM on all the sequences.

To show the advantage of MFGAN for VO compared with other recent style transfer methods, in Table 3, we show the results obtained by using the translated sequences from ToDayGAN [20], DRIT [21], and LinearTransfer(LT) [22]. For ToDayGAN we use the pre-trained model, since it also trained for Day-Night style transfer. We train DRIT with our split and do the inference with a constant noise vector. For LT, we use a Day image as the style, and run the video inference on the test sequences. From the table we can see that MFGAN outperforms other methods in terms of both accuracy and robustness for VO.

5 Conclusion

In this paper, we present a learning-based approach to improve stereo VO methods in low lighting conditions. To this end, we introduced the concept of Multi-Frame GAN (MFGAN) which performs a spatio-temporally consistent domain transfer. MFGAN takes advantage of unpaired datasets by leveraging a novel cycle adversarial network and learns to generate frames with temporal and stereo coherence. With the proposed metric for frame consistency, we quantitatively validate that our method successfully generates images with temporal as well as stereo consistency. Experiments regarding VO on both a synthetic indoor dataset and a real outdoor dataset show that our method improves the performance of both indirect and direct VO methods in low light environments. We also show that MFGAN outperforms other photo enhancement and image/video translation methods by a notable margin. In future work, we will explore the generalization capability of the proposed temporal and stereo consistency loss on other style transfer methods.

References

[1] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[2] R. Wang, M. Schworer, and D. Cremers. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3903–3911, 2017.

[3] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018.

[4] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 834–849, 2014.

[5] H. Strasdat, J. Montiel, and A. J. Davison. Scale drift-aware large scale monocular SLAM. *Robotics: Science and Systems VI*, 2(3):7, 2010.

[6] N. Yang, R. Wang, J. Stuckler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[7] N. Yang, R. Wang, X. Gao, and D. Cremers. Challenges in monocular visual odometry: Photometric calibration, motion bias, and rolling shutter effect. *IEEE Robotics and Automation Letters*, 3(4):2878–2885, 2018.

[8] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1–10, 2007.

[9] I. Shim, J.-Y. Lee, and I. S. Kweon. Auto-adjusting camera exposure for outdoor robotics using gradient information. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1011–1017, 2014.

[10] Z. Zhang, C. Forster, and D. Scaramuzza. Active exposure control for robust visual odometry in HDR environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3894–3901, 2017.

[11] P. Bergmann, R. Wang, and D. Cremers. Online photometric calibration of auto exposure video for realtime visual odometry and slam. *IEEE Robotics and Automation Letters*, 3(2):627–634, 2017.

[12] H. Alismail, M. Kaess, B. Browning, and S. Lucey. Direct visual odometry in low light using binary descriptors. *IEEE Robotics and Automation Letters*, 2(2):444–451, 2017.

[13] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman. NID-SLAM: Robust monocular SLAM using normalised information distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1435–1444, 2017.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] R. Gomez-Ojeda, Z. Zhang, J. Gonzalez-Jimenez, and D. Scaramuzza. Learning-based image enhancement for visual odometry in challenging HDR environments. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 805–811, 2018.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[26] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018.

[27] H. Porav, W. Maddern, and P. Newman. Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1011–1018, 2018.

[28] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019.

[29] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1105–1114, 2017.

[30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[31] C. Gao, D. Gu, F. Zhang, and Y. Yu. Reconet: Real-time coherent video style transfer network. *arXiv preprint arXiv:1807.01197*, 2018.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017.

[35] 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.