Self-supervised Monocular Depth Estimation for All Day Images using Domain Separation

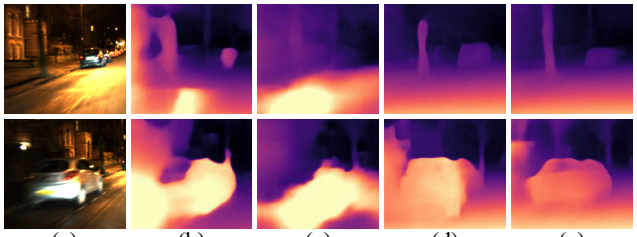
Lina Liu[1,2], Xibin Song[2,4]*, Mengmeng Wang[1], Yong Liu[1,3]* and Liangjun Zhang[2,4]
[1]Institute of Cyber-Systems and Control, Zhejiang University, China
[2]Baidu Research, China [3]Huzhou Institue of Zhejiang University, China
[4]National Engineering Laboratory of Deep Learning Technology and Application, China
{linaliu,mengmengwang}@zju.edu.cn,song.sducg@gmail.com,liangjunzhang@baidu.com,yongliu@iipc.zju.edu.cn

Abstract

Remarkable results have been achieved by DCNN based self-supervised depth estimation approaches. However, most of these approaches can only handle either day-time or night-time images, while their performance degrades for all-day images due to large domain shift and the variation of illumination between day and night images. To relieve these limitations, we propose a domain-separated network for self-supervised depth estimation of all-day images. Specifically, to relieve the negative influence of disturbing terms (illumination, etc.), we partition the information of day and night image pairs into two complementary sub-spaces: private and invariant domains, where the former contains the unique information (illumination, etc.) of day and night images and the latter contains essential shared information (texture, etc.). Meanwhile, to guarantee that the day and night images contain the same information, the domain-separated network takes the day-time images and corresponding night-time images (generated by GAN) as input, and the private and invariant feature extractors are learned by orthogonality and similarity loss, where the domain gap can be alleviated, thus better depth maps can be expected. Meanwhile, the reconstruction and photometric losses are utilized to estimate complementary information and depth maps effectively. Experimental results demonstrate that our approach achieves state-of-the-art depth estimation results for all-day images on the challenging Oxford RobotCar dataset, proving the superiority of our proposed approach. Code and data split are available at https://github.com/LINA-lln/ADDS-DepthNet.

1. Introduction

Self-supervised depth estimation has been applied in a wide range of fields such as augmented reality [3][5], 3D reconstruction [17], SLAM [22][30][31] and scene understanding [1][6] since it does not need large and accurate ground-truth depth labels as supervision. The depth information can be estimated by the implicit supervision provided by the spatial and temporal consistency present in image sequences. Benefiting from the well developed deep learning technology, impressive results have been achieved by Deep Convolution Neural Network(DCNN) based approaches [14][21][24][25], which outperform traditional methods that rely on handcrafted features and exploit camera geometry and/or camera motion for depth and pose estimation.



(a) (b) (c) (d) (e)

Figure 1. Comparison with other approaches on Oxford RobotCar dataset [28]. From left to right: (a) Night Images, (b) Monodepth2 [12], (c) HR-Depth [27], (d) Monodepth2+CycleGAN [44], (e) Ours.

However, most of current DCNN based self-supervised depth estimation approaches [18][36][37][42] mainly solve the problem of depth estimation on day-time images, which are evaluated by day-time benchmarks, such as KITTI [9] and Cityscapes [7]. They fail to generalize well on all-day images due to the large domain shift between day and night images. The night-time images are unstable due to the low visibility and non-uniform illumination arising from multiple and moving lights. Methods [16][19] are proposed by applying a commonly used depth estimation strategy for images captured in low-light conditions. However, the performance is limited due to the unstable visibility. Meanwhile, generative adversarial networks(GAN), such as CycleGAN [44], are also used to solve the problem of depth es-

*Corresponding authors

timation on night-time images by translating information of night-time to day-time in both image levels and feature levels. Unfortunately, due to the inherent domain shift between day and night-time images, it is difficult to obtain natural day-time images or features with GAN using night-time images as input, thus the performance is also limited. Fig. 1 (b) and (c) demonstrate the results of Monodepth2 [12] and HR-Depth [27] of night-time images. Monodepth2[12] is an effective self-supervised depth estimation approach, and HR-Depth[27] make a series of improvements based on Monodepth2[12]. Fig. 1 (d) demonstrate the result Monodepth2 [44] with CycleGAN translated image as input. We can see that the depth details are failed to be estimated due to the non-uniform illumination of night-time images.

For a scene in real-world, the depth is constant if the viewpoint is fixed, while the disturbing terms, such as illumination, varies as time goes, which will disturb the performance of self-supervised depth estimation, especially for night-time images. [8] also proves that texture information plays more important roles on depth estimation than exact color information. To cater to the above issues, we propose a domain-separated network for self-supervised depth estimation of all-day RGB images. The information of day and night image pairs are separated into two complementary sub-spaces: private and invariant domains. Both domains use DCNN to extract features. The private domain contains the unique information (illumination, etc.) of day and night-time images, which will disturb the performance of depth estimation. In contrast, the invariant domain contains invariant information (texture, etc.), which can be used for common depth estimation. Thus the disturbed information can be removed and better depth maps will be obtained.

Meanwhile, unpaired day and night images always contain inconsistent information, which interferes with the separation of private and invariant features. Therefore, the domain-separated network takes a paired of the day-time image and corresponding night-time image (generated by GAN) as input, the private and invariant feature extractors are first utilized to extract private (illumination, etc.) and invariant (texture, etc.) features using orthogonality and similarity losses, which can obtain more effective features for depth estimation of both day and night-time images. Besides, constraints in feature and gram matrices levels are leveraged in orthogonality losses to alleviate the domain gap, thus more effective features and fine-grain depth maps can be obtained. Then, depth maps and corresponding RGB images are reconstructed by decoder modules with reconstruction and photometric losses. Note that real-world day-time and night-time images can be tested directly. As shown in Fig. 1 (e), our approach can effectively relieve the problems of low-visibility and non-uniform illumination, and achieves more appealing results for night-time images.

The main contributions can be summarized as:

- We propose a domain-separated framework for self-supervised depth estimation of all-day images. It can relieve the influence of disturbing terms in depth estimation by separating the all-day information into two complementary sub-spaces: private (illumination, etc.) and invariant (texture, etc.) domains, thus better depth maps can be expected;
- Private and invariant feature extractors with orthogonality and similarity losses are utilized to extract effective and complementary features to estimate depth information. Meanwhile, the reconstruction loss is employed to refine the obtained complementary information (private and invariant information);
- Experimental results on the Oxford RobotCar dataset demonstrate that our framework achieves state-of-the-art depth estimation performance for all-day images, which confirms the superiority of our approach.

2. Related work

2.1. Day-time Depth Estimation

Self-supervised depth estimation has been extensively studied in recent years. [43][11] are the first self-supervised monocular depth estimation approaches which train the depth network along with a separate pose network. Meanwhile, [12][13][15][20] make a series of improvements for outdoor scenes, which are sufficiently evaluated on KITTI dataset [9] and Citycapes dataset [7] subsequently. [23][32][38] outperform better results in indoor scenes.

KITTI [9] and Citycapes [7] datasets only contain day-time images, and all of the above methods are excellently improved for these scenes. However, the self-supervised depth estimation approaches for all-day images have not been well addressed before, and the performance of current approaches on night-time images is limited due to the low-visibility and non-uniform illuminations.

2.2. Night-time Depth Estimation

Approaches have also been proposed for self-supervised depth estimation of night-time images.

[19][26] propose to use additional sensors to estimate depth of night-time images. To estimate all day time depth, [19] utilizes a thermal imaging camera sensor to reduce the influence of low-visibility in the night-time, while [26] adds LiDAR to provide additional information in estimating depth maps at night-time. Meanwhile, using generate adversarial network, [33] and [34] propose effective strategies for depth estimation of night-time images. [33] utilizes a translation network with light effects and uninformative regions that can render realistic night stereo images from day stereo images, and vice versa. During inference, a separate network during the day and night is used to estimate

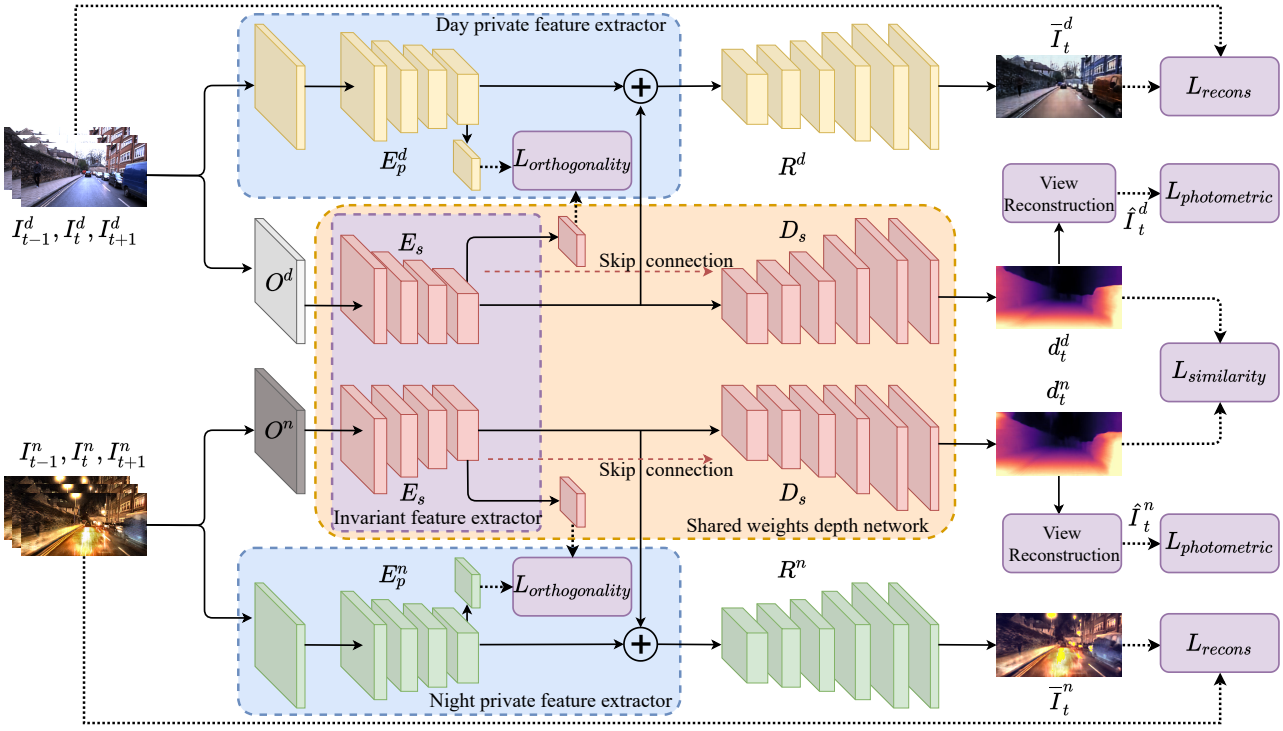


Figure 2. Overview of the network architecture. The network architecture includes three parts: Shared weights depth network (orange area), Day private branch (yellow structure) and night private branch (green structure). Day-time and night-time images are the input of the shared weights depth network, which extracts the invariant features first, and then estimates corresponding depth maps. Meanwhile, the day private feature extractor and night private feature extractor (blue area) extract the private features of day and night, respectively, which are constrained by orthogonality loss to get complementary features. And the private and invariant features are added to reconstruct the original input images with the reconstruction loss. In inference, only operations of O^d, O^n and shared weights depth network are used to estimate depth.

the stereo depth of night-time images. [34] proposes an adversarial domain feature adaptation method to reduce the domain shift between day and night images at the feature level. Finally, independent encoders and a shared decoder are used for the day-time and night-time images during inference.

Though remarkable progress has been achieved, due to the large domain shift between day-time and night-time images, it is difficult to obtain natural day-time images or features with night-time images as input, thus the performances of these approaches are limited.

2.3. Domain Adaptation

Most depth estimation and stereo matching domain adaptation methods mainly focus on the migration between the synthetic domain and the real domain or between different datasets. Most methods usually translate images from one domain to another. To reduce the requirements of real-world images in depth estimation, [2][40][41] explore image translation techniques to generate synthetic labeled data. [29] tackles synthetic to real depth estimation issue by using domain invariant defocus blur as direct supervision. [39] proposes a domain normalization approach of stereo matching that regularizes the distribution of learned representations to allow them to be invariant to domain differences.

Compared with previous approaches, we propose an effective domain separation framework for all-day self-supervised depth estimation, which can effectively handle the problem of domain shift between day and night-time images.

3. Approach

We propose a domain-separated framework to relieve the influence of disturbing terms, which takes day-time images and corresponding night-time images generated by GAN as input, and Fig. 2 demonstrates the pipeline of our proposed domain separated framework.

3.1. Domain Separated Framework

For day-time and night-time images of a scene, the depth information should be consistent, though the illumination of these image pairs is quite different. This means that the essential information of corresponding day-time images and night-time images of a scene should be similar. Here, we

separate the information of day and night-time images into two parts:

$$\begin{aligned} I^d &= I_i^d + I_p^d, \\ I^n &= I_i^n + I_p^n \end{aligned} \tag{1}$$

where I_i^d and I_i^n mean the invariant information of day and night images, which should be similar of the same scene, and I_p^d and I_p^n mean the different private information (illumination, etc.) of day and night images, respectively.

Inspired by [8], the illumination of a scene is different as time goes, while the depth of the scene is constant, thus the illumination components (I_p^d and I_p^n) of a scene play fewer roles in self-supervised depth estimation. As shown in Fig. 2, the proposed domain-separated framework separates the images into two complementary sub-spaces in feature levels (elucidated in Fig.5), and the invariant components are utilized for depth estimation.

Moreover, it is quite difficult to guarantee that the real-world day-time and night-time images of a scene contain the same information except for the private information (illumination, etc.), since there are always moving objects in outdoor scenes. This will mislead the network to obtain private and invariant components of images. Therefore, CycleGAN[44] is used to translate day-time images to night-time images, where the day-time and corresponding generated night-time images are regarded as input image pairs. It ensures that the invariant information is consistent, and all objects are in the same position, reducing the loss of essential information during the process of separating private information. Note that other GANs can also be used here.

Inspired by [4], our domain-separated framework uses two network branches to extract the private and invariant information of an image in feature levels, respectively. Given the input day image sequences $\{I_{t-1}^d, I_t^d, I_{t+1}^d\}$ and the corresponding generated night images sequences $\{I_{t-1}^n, I_t^n, I_{t+1}^n\}$, where t represents the t-th frame image arranged in chronological order, the day private feature extractor E_p^d and night private feature extractor E_p^n are used to extract private features of day-time images and night-time images f_p^d and f_p^n, respectively. The invariant feature extractor E_i^d and E_i^n are utilized to extract invariant features of day-time and night-time images f_i^d and f_i^n, respectively. Since the input day-time and night-time images contains same essential information, E_i^d and E_i^n are weight-shared, which is defined as E_s. Then the feature extraction process can be formulated as:

$$\begin{aligned} f_{p_t}^d &= E_p^d(I_t^d), f_{p_t}^n = E_p^n(I_t^n) \\ f_{i_t}^d &= E_s(I_t^d), f_{i_t}^n = E_s(I_t^n) \end{aligned} \tag{2}$$

where t in the subscript denotes the t-th frame of day and night-time.

Then, decoders are used to reconstruct the corresponding depth maps of day and night-time images. As shown in Fig. 2, the red decoder D_s represents the depth recovery module of shared weights depth network, and the yellow decoder R^d and green decoder R^n denote the reconstructed feature restoration branch. The process of the depth map and image reconstruction can be formulated as:

$$\begin{aligned} \overline{I}_t^d &= R^d(f_{p_t}^d + f_{i_t}^d) \\ \overline{I}_t^n &= R^n(f_{p_t}^n + f_{i_t}^n) \\ d_t^d &= D_s(f_{i_t}^d) \\ d_t^n &= D_s(f_{i_t}^d) \end{aligned} \tag{3}$$

where $\overline{I}_t^d$ and $\overline{I}_t^n$ are the reconstructed images of t-th frame by R^d and R^n, and d_t^d and d_t^n are the corresponding depth maps estimated by D_s.

3.2. Loss function

To obtain private and invariant features and well estimate depth information of all-day images in a self-supervised manner, different losses are leveraged here, including reconstruction loss, similarity loss, orthogonality loss and photometric loss.

3.2.1 Reconstruction Loss

The private and invariant features are complementary information which can be used to reconstruct the original RGB images. Hence, we use reconstruction loss to refine the domain separated framework, which is defined as:

$$\begin{aligned} L_{recons} = &\frac{1}{N}\sum_x(\overline{I}_{tx}^d - I_{tx}^d)^2 + \frac{1}{N^2}(\sum_x(\overline{I}_{tx}^d - I_{tx}^d))^2 \\ &+ \frac{1}{N}\sum_x(\overline{I}_{tx}^n - I_{tx}^n)^2 + \frac{1}{N^2}(\sum_x(\overline{I}_{tx}^n - I_{tx}^n))^2 \end{aligned} \tag{4}$$

where $x \in [1, N]$, N is the pixel number of I_t^n and I_t^d.

3.2.2 Similarity Loss

The proposed domain separated framework takes day-time images and corresponding generated night-time images (CycleGAN [44]) as input, and the estimated depth maps of day-time and night-time images should be consistent. Due to the inherent advantages of the day-time image in depth estimation, the estimated depth of the night image is expected to be as close as possible to the day-time, that is, the depth of the day-time image is used as a pseudo-label to constrain the depth of the night-time image. So the similarity loss is defined as:

$$L_{simi} = \frac{1}{N}\sum_x(d_{tx}^n - \widetilde{d}_{tx}^d)^2 \tag{5}$$

where $x \in [1, N]$, N is the pixel number of d_t^n and d_t^d, x is the x-th pixel. $\widetilde{d}_t^d$ means that the gradient of d_t^d is cut off during back propagation.

3.2.3 Orthogonality Loss

As discussed above, the private and invariant features of an image are complementary and completely different. Therefore, two types of orthogonality losses are utilized here to guarantee the private and invariant features are completely different.

Direct feature orthogonality loss: the private and invariant feature extractors obtains 3-D private and invariant features (f_p and f_i) which have large sizes. To reduce the complexity, we first use a convolution layer with a kernel size of 1×1 to reduce the size of obtained private and invariant features (v_p and v_i), then we straighten the reduced features into 1-D feature vectors. Finally, we calculate the inner product (orthogonality loss) between the private and invariant feature vectors, which is defined as L_f.

Gram matrices orthogonality loss: inspired by style transfer [10], Gram matrix is widely used in style transfer to represent the style of the features. The private and invariant features can be considered to have different styles. Hence, we first calculate the Gram matrices (η_p and η_i) of private and invariant features, then straighten them to 1-D feature vectors, thus the orthogonality loss between these vectors can be calculated, which is defined as L_g.

The process of L_f and L_g can be defined as:

$$\begin{aligned} v_{i_t}^d &= C_{rs}^d(f_{i_t}^d), v_{p_t}^d = C_{rp}^d(f_{p_t}^d) \\ v_{i_t}^n &= C_{rs}^n(f_{i_t}^n), v_{p_t}^n = C_{rp}^n(f_{p_t}^n) \end{aligned} \tag{6}$$

where C_{rs}^d, C_{rp}^d, C_{rs}^n and C_{rp}^n are the 1×1 convolution operation for invariant and private features of day-time and night-time images, respectively.

$$\begin{aligned} L_{ortho} &= L_f + L_g \\ L_f &= V(v_{i_t}^d) \cdot V(v_{p_t}^d) + V(v_{i_t}^n) \cdot V(v_{p_t}^n) \\ L_g &= V(\eta_{i_t}^d) \cdot V(\eta_{p_t}^d) + V(\eta_{i_t}^n) \cdot V(\eta_{p_t}^n) \end{aligned} \tag{7}$$

where V is the operation which convert multi-dimensional features to 1-D features.

3.2.4 Photometric Loss

Following Monodepth2[12], photometric loss are utilized in the self-supervised depth estimation process. Photometric loss L_{pm} can be formulated as the same as [12]:

$$\begin{aligned} \hat{I}_t^d &= P(d_t^d, pose_{(t-1,t)}^d, I_{(t-1)}^d) \\ \hat{I}_t^n &= P(d_t^n, pose_{(t-1,t)}^n, I_{(t-1)}^n) \\ L_{pm} &= \frac{\alpha}{2}(1 - \text{SSIM}(\hat{I}_t^d, I_t^d)) + (1-\alpha)\|\hat{I}_t^d - I_t^d\|_1 \\ &+ \frac{\alpha}{2}(1 - \text{SSIM}(\hat{I}_t^n, I_t^n)) + (1-\alpha)\|\hat{I}_t^n - I_t^n\|_1 \end{aligned} \tag{8}$$

where $pose_{(t-1,t)}$ is the pose estimation process, and following [12], here we use method [35] for pose estimation, $\hat{I}$ is the reprojection image, and $\alpha = 0.85$ which is set empirically (same as [12]).

3.2.5 Total Loss

The total training loss of the network is

$$L_{total} = \lambda_1 L_{recons} + \lambda_2 L_{simi} + \lambda_3 L_{ortho} + \lambda_4 L_{pm} \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weight parameters. In this paper, we set $\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = \lambda_4 = 1$, empirically.

3.3. Inference Process

For day-time image, the output d_t^d is $D_s(E_s(O^d(I_t^d)))$; for night-time d_t^n is $D_s(E_s(O^n(I_t^n)))$. Except for the first convolution layer, the remaining parameters of the depth estimation during the day and night are all shared. In inference, only operations of O^d, O^n and shared weights depth network are used to estimate depth.

4. Experiments

In this section, following [34], we compare the performance of our method with state-of-the-art approaches on Oxford RobotCar dataset[28], which is split to adapt all day time monocular image depth estimation.

4.1. Oxford RobotCar Dataset

The KITTI [9] and Cityscapes [7] datasets are widely used in depth estimation task. However, only day-time images are contained in these datasets, which cannot meet the requirements for all-day depth estimation. Therefore, we choose the Oxford RobotCar dataset [28] as the training and testing dataset. It is a large outdoor-driving dataset that contains images at various times captured in one year, including day and night times. Following [34], we use the left images with the resolution of 960×1280 (collected by the front stereo-camera (Bumblebee XB3)) for self-supervised depth estimation. Sequence "2014-12-09-13-21-02" and "2014-12-16-18-44-24" are used for day-time and night-time training, respectively. Both training data are selected from the first 5 splits. The testing data are collected from the other splits of the Oxford RobotCar dataset, which contains 451 day-times images and 411 night-times images. We use the depth data captured by the front LMS-151 depth sensors as the ground truth in the testing phase. The images are first center-cropped to 640×1280, and then resized to 256×512 as the inputs of the network.

4.2. Implementation Details

First, we use CycleGAN[44] to translate day-time images to night-time images, then generated images and the day-time images are regarded as image pairs which are the input of the proposed domain-separated network.

Method (**test at night**)	Max depth	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [12](day)	40m	0.477	5.389	9.163	0.466	0.351	0.635	0.826
Monodepth2 [12](night)	40m	0.661	25.213	12.187	0.553	0.551	0.849	0.914
Monodepth2+CycleGAN [44]	40m	0.246	2.870	7.377	0.289	0.672	0.890	0.950
HR-Depth [27]	40m	0.512	5.800	8.726	0.484	0.388	0.666	0.827
ADFA[1] [34]	40m	**0.201**	2.575	7.172	0.278	**0.735**	0.883	0.942
Ours	40m	0.233	**2.344**	**6.859**	**0.270**	0.631	**0.908**	**0.962**
Monodepth2 [12](day)	60m	0.432	5.366	11.267	0.463	0.361	0.653	0.839
Monodepth2 [12](night)	60m	0.580	21.446	12.771	0.521	0.552	0.840	0.920
Monodepth2+CycleGAN [44]	60m	0.244	3.202	9.427	0.306	0.644	0.872	0.946
HR-Depth [27]	60m	0.462	5.660	11.009	0.477	0.374	0.670	0.842
ADFA[1] [34]	60m	0.233	3.783	10.089	0.319	**0.668**	0.844	0.924
Ours	60m	**0.231**	**2.674**	**8.800**	**0.286**	0.620	**0.892**	**0.956**
Method (**test at day**)	Max depth	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [12](day)	40m	0.117	0.673	3.747	0.161	0.867	0.973	0.991
Monodepth2 [12](night)	40m	0.306	2.313	5.468	0.325	0.545	0.842	0.937
HR-Depth [27]	40m	0.121	0.732	3.947	0.166	0.848	0.970	0.991
Ours	40m	**0.109**	**0.584**	**3.578**	**0.153**	**0.880**	**0.976**	**0.992**
Monodepth2 [12](day)	60m	0.124	0.931	5.208	0.178	0.844	0.963	**0.989**
Monodepth2 [12](night)	60m	0.294	2.533	7.278	0.338	0.541	0.831	0.934
HR-Depth [27]	60m	0.129	1.013	5.468	0.184	0.825	0.958	**0.989**
Ours	60m	**0.115**	**0.794**	**4.855**	**0.168**	**0.863**	**0.967**	**0.989**

Table 1. Quantitative comparison with state-of-the-art methods. Higher value is better for the last three columns, lower value is better for others. Monodepth2[12](day) and HR-Depth[27] mean training with day-time data of the Oxford dataset and testing on the night and day test set. Monodepth2[12](night) means training with night-time data of the Oxford dataset and testing on the night and day test set. CycleGAN[44] means translating night-time Oxford images into day-time images and then using a day-time trained Monodepth2 model to estimate depth from these translated images the same as [34]. The best results are highlighted.

The network is trained 20 epochs in end-to-end manner with Adam optimizer (β_1=0.9, β_2=0.999). We set batch size as 6, the initial learning rate as 1e-4 for the first 15 epochs, and the learning rate is set as 1e-5 for the remaining epochs. In inference, except for the first convolution layer, the day and night branches share weights in the depth estimation part.

4.3. Quantitative Results

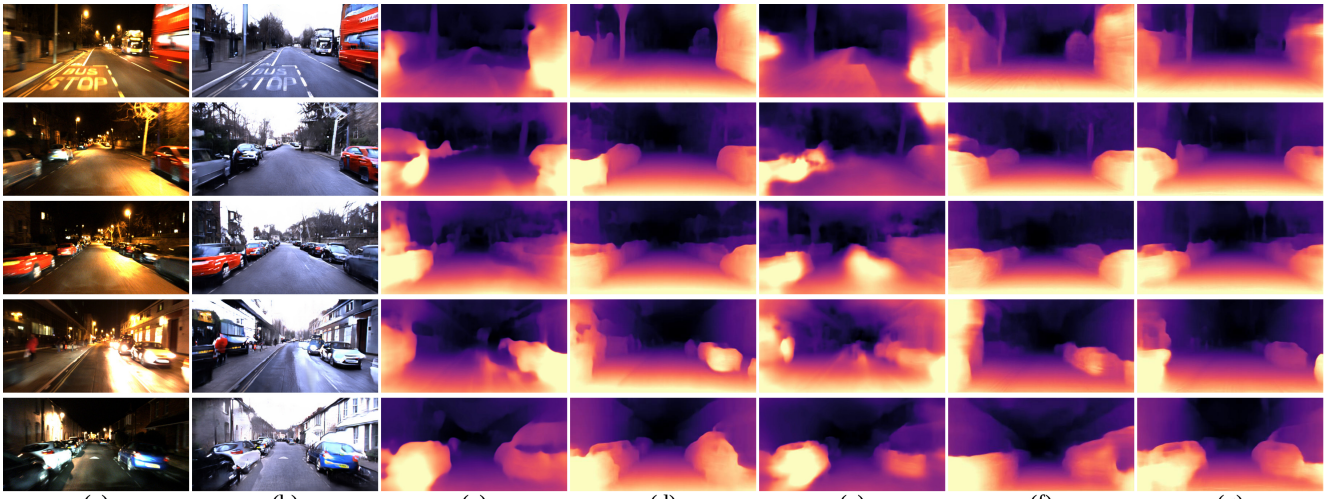
Table. 1 demonstrates the quantitative comparison results between our approach and state-of-the-art approaches. Following [34], we evaluate the performance with two depth ranges: within 40m and 60m. In Table. 1, Monodepth2 [12](day) means the results trained with day-time images of the Oxford dataset, while Monodepth2 [12](night) means the results trained with night-time images of the Oxford dataset. Monodepth2+CycleGAN [44] means the results of translating night-time Oxford images into day-time images and then use a day-time trained Monodepth2 [12] model to estimate depth from these translated images the same as [34].

As shown in Table. 1, Monodepth2 [12] is an effective self-supervised depth estimation approach, which works well for day-time images. However, the performances are limited on night-time images for all models trained by day-time and night-time images. Due to the non-uniform illumination of night-time images, areas that are too bright and too dark will cause varying degrees loss of information, which leads to the fact that training directly on images at night cannot get absolutely good results. Meanwhile, although Monodepth2+CycleGAN[44] and HR-Depth [27] can improve the depth estimation results of the night images to a certain extent, the performances are also limited due to the non-uniform illumination of night-time images. ADFA [34] reduces the domain shift between day and night images at the feature level, but the performance is limited by day-time results. As shown in Table. 1, our domain-separated framework can effectively relieve the influence of disturbing terms, which can improve the depth estimation performance for all-day images. Almost all performance metrics in the depth ranges of 40m and 60m for day and night images can be largely improved by our approach, which prove the superiority of our method.

4.4. Qualitative Results

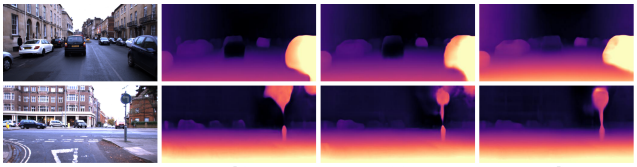
The qualitative comparison results of night-time images are shown in Fig. 3, where (b) shows the translated day-times images of CycleGAN [44] from night-time images, (c) shows the results of Monodepth2 [12] trained with day-time images and tested with night-time images, (d) shows the results of Monodpeth2 [12] trained with day-time im-

[1]Note that the test set of ADFA [34] is not available, our test set is not exactly the same as the test set of ADFA [34].



(a) (b) (c) (d) (e) (f) (g)

Figure 3. Qualitative comparison with other state-of-the-art methods at night. From left to right: (a) Night Images, (b) Fake Day Images translated by CycleGAN[44], (c) Monodepth2[12], (d) Monodepth2+CycleGAN[44], (e) HR-Depth[27], (f) ADFA[34], (g) Ours.



(a) (b) (c) (d)

Figure 4. Qualitative comparison with other state-of-the-art methods at day. From left to right: (a) Day Images, (b) Monodepth2 [12](day), (c) HR-Depth [27], (d) Ours.

ages and tested with generated day-time images (CycleGAN [44]). Compared with (c), it is obvious that (d) obtained better visual results, which prove that CycleGAN works positively for night-time depth estimation. (f) shows the result of ADFA [34] which uses generative adversarial strategy in feature level of night-time images, which cannot fully restore the contour of the objects. Comparing with (c) to (f), more visual apparelling results can be obtained by our approach, which proves the effectiveness of our method.

Fig.4 demonstrates the qualitative comparison results of day-time images. We can see that more depth details can be recovered by our approach, which clarifies the ability of our method for day-time images.

4.5. Ablation Study

4.5.1 Private and Invariant Features

Fig. 5 demonstrates the private and invariant features obtained by the private and invariant feature extractors of day and night-time images, where the first column shows the corresponding input images, and the remaining columns from left to right are top 10 feature maps that contain more information. Row (a) and (b) are private features of the day and night-time images, while row (c) and (d) are corresponding invariant features. It is obvious to see that feature maps in Fig. 5 (a) and (b) contain non-regular and smooth information with fewer structures, which is similar to the illumination information of images. Feature maps in Fig. 5 (c) and (d) contain regular and texture information with obvious structures, which can represent the invariant of scenes, proving that our approach can separate the private (illumination, etc.) and invariant (texture, etc.) information effectively.

4.5.2 Analysis of Input Data and Losses

Unpaired data vs. paired data: The results of U and P in Table. 2 show the quantitative results of our method with unpaired and paired images as input. In specific, we use the day-time and night-time images captured in the same roads from Oxford RobotCar dataset as unpaired data (U), and we use day-time images and the corresponding generated night-time images (generated by GAN) as paired data (P). We can see that results obtained by paired data outperform results obtained by unpaired data, which is mainly because that inconsistent information exists in unpaired images since they are captured at different times, though on the same roads. Hence, we use paired images in this paper.

Reconstruction loss: The private and invariant features are complementary, which should contain all the information of the original images. Therefore, we use reconstruction loss. Table.2 PR shows the quantitative results of our approach with reconstruction loss. Comparing with P, PR produces great improvement to the depth estimation of night-time images. However, the result of the day-time images is slightly worse, because the invariant feature extractor is shared during the day and night, and no other constraints are used.

Orthogonality loss: Orthogonality loss is used to guarantee that private and invariant features can be separated by the private and invariant extractors orthogonally, which

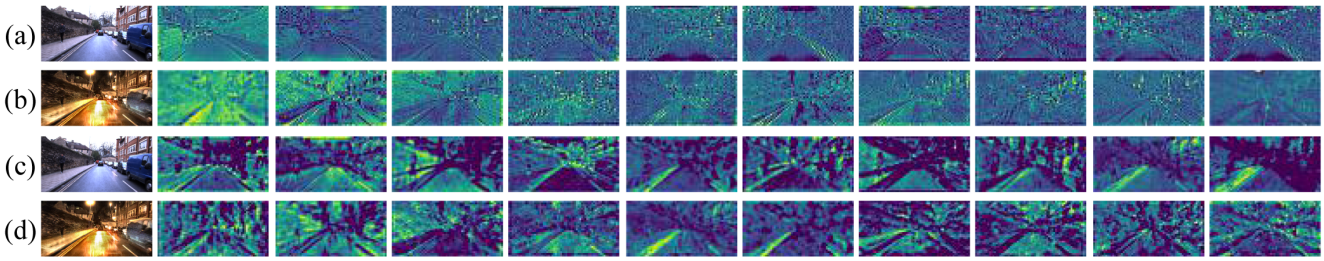


Figure 5. Visualization of Convolution Features. From left to right: (a) Day-time Private Features, (b)Night-time Private Features, (c)Day-time Invariant Features, (d)Night-time Invariant Features. The first column shows the corresponding input images, and the remaining columns from left to right are top 10 feature maps that contain more information.

Method (**night**)	Paired	L_{recons}	L_f	L_g	L_{simi}	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
U						0.573	18.577	11.189	0.524	0.569	0.807	0.897
P	√					0.429	15.183	11.401	0.422	0.589	0.862	0.942
PR	√	√				0.357	10.699	10.385	0.377	0.611	0.884	0.946
PRF	√	√	√			0.251	2.993	8.173	0.299	0.606	0.884	0.949
$PRFG$	√	√	√	√		0.231	2.453	7.327	0.282	0.662	0.900	0.956
$PRFGS$	√	√	√	√	√	0.233	2.344	6.859	0.270	0.631	0.908	0.962
Method (**day**)	Paired	L_{recons}	L_f	L_g	L_{simi}	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
U						0.280	4.509	6.359	0.297	0.661	0.873	0.949
P	√					0.117	0.758	3.737	0.170	0.874	0.967	0.988
PR	√	√				0.131	1.355	3.937	0.178	0.848	0.967	0.990
PRF	√	√	√			0.108	0.569	3.535	0.152	0.879	0.977	0.992
$PRFG$	√	√	√	√		0.109	0.580	3.518	0.152	0.891	0.976	0.991
$PRFGS$	√	√	√	√	√	0.109	0.584	3.578	0.153	0.880	0.976	0.992

Table 2. The table shows the quantitative results of the proposed losses and input data. All experiments are tested within the depth range of 40m.

is composed of L_f and L_g. Table.2. PRF and $PRFG$ show the quantitative results of our approach with L_f and L_g, respectively. Compared with PR, PRF (added L_f loss) can greatly improve the performance of depth estimation night-time images, which also works positively for day-time images. Meanwhile, $PRFG$ (added L_g loss) can help the network achieves better performance on night-time images while maintaining the performance of day-time images, thereby further improving the performance of depth estimation for all-day images.

Similarity loss: The estimated depth maps should be similar for the input paired images, because consistent information is contained. Hence, similarity loss is employed in our approach. Since the depth estimation process of the day-time usually achieves better results than night, we use the day-time depth as a pseudo label so that the depth of the paired night image should be close to the day-time. Table.2. $PRFGS$ demonstrates the quantitative results of our approach with similarity loss, which shows that the similarity constraint can further improve the depth estimation result of nigh-time images while maintaining the performance of day-time images, thus proving the effectiveness of the similarity loss.

5. Conclusion

In this paper, to relieve the problem of low-visibility and non-uniform illumination in self-supervised depth estimation of all-day images, we propose an effective domain separated framework, which separates the images into two complementary sub-spaces in feature levels, including private (illumination, etc.) and invariant (texture, etc.) domains. The invariant (texture, etc.) features are employed for depth estimation, thus the influence of disturbing terms, such as low-visibility and non-uniform illumination in images, can be relieved, and effective depth information can be obtained. To alleviate the inconsistent information of day and night images, the domain-separated network takes the day-time images and the corresponding generated night-time images (GAN) as input. Meanwhile, orthogonality, similarity and reconstruction losses are utilized to separate and constrain the private and invariant features effectively, thus better depth estimation results can be expected. Note that the proposed approach is fully self-supervised and can be trained end-to-end, which can adapt to both the day and night images. Experiments on the challenging Oxford RobotCar dataset demonstrate that our framework achieves state-of-the-art results for all-day images.

Acknowledgements This work is supported in part by Robotics and Autonomous Driving Lab of Baidu Research. This work is also supported in part by the National Key R&D Program of China under Grant (No.2018YFB1305900) and the National Natural Science Foundation of China under Grant (No.61836015).

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[3] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *arXiv preprint arXiv:1608.06019*, 2016.

[5] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51(1):341–377, 2011.

[6] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[10] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[13] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020.

[16] Sunghoon Im, Hae-Gon Jeon, and In So Kweon. Robust depth estimation from auto bracketed images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2946–2954, 2018.

[17] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[18] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. Multispectral transfer network: Unsupervised depth estimation for all-day vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In *European Conference on Computer Vision (ECCV)*, 2020.

[21] Lina Liu, Yiyi Liao, Yue Wang, Andreas Geiger, and Yong Liu. Learning steering kernels for guided depth completion. *IEEE Transactions on Image Processing*, 30:2850–2861, 2021.

[22] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2136–2144, 2021.

[23] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. *arXiv preprint arXiv:2103.15483*, 2021.

[24] Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Wenping Wang. Multi-view depth estimation using epipolar spatio-temporal network. *arXiv preprint arXiv:2011.13118*, 2020.

[25] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *European Conference on Computer Vision*, pages 640–657. Springer, 2020.

[26] Yawen Lu and Guoyu Lu. An alternative of lidar in nighttime: Unsupervised depth estimation based on single ther-

mal image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3833–3843, 2021.

[27] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *arXiv preprint arXiv:2012.07356*, 2020.

[28] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[29] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1071–1080, 2020.

[30] Raul Mur-Artal and Juan 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.

[31] A Alan B Pritsker. *Introduction to Simulation and SLAM II*. Halsted Press, 1984.

[32] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Aashish Sharma, Loong-Fah Cheong, Lionel Heng, and Robby T Tan. Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions. In *2020 International Conference on 3D Vision (3DV)*, pages 23–31. IEEE, 2020.

[34] 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 *European Conference on Computer Vision*, pages 443–459. Springer, 2020.

[35] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[36] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[37] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[38] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *ECCV*, 2020.

[39] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020.

[40] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019.

[41] Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless Fowlkes. Domain decluttering: simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3330–3340, 2020.

[42] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[43] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.