Unsupervised Monocular Depth Estimation in Highly Complex Environments

Chaoqiang Zhao, Yang Tang, Senior Member, IEEE, Qiyu Sun

Abstract—Previous unsupervised monocular depth estimation methods mainly focus on the day-time scenario, and their frameworks are driven by warped photometric consistency. While in some challenging environments, like night, rainy night or snowy winter, the photometry of the same pixel on different frames is inconsistent because of the complex lighting and reflection, so that the day-time unsupervised frameworks cannot be directly applied to these complex scenarios. In this paper, we investigate the problem of unsupervised monocular depth estimation in certain highly complex scenarios. We address this challenging problem by using domain adaptation, and a unified image transfer-based adaptation framework is proposed based on monocular videos in this paper. The depth model trained on day-time scenarios is adapted to different complex scenarios. Instead of adapting the whole depth network, we just consider the encoder network for lower computational complexity. The depth models adapted by the proposed framework to different scenarios share the same decoder, which is practical. Constraints on both feature space and output space promote the framework to learn the key features for depth decoding, and the smoothness loss is introduced into the adaptation framework for better depth estimation performance. Extensive experiments show the effectiveness of the proposed unsupervised framework in estimating the dense depth map from the night-time, rainy night-time and snowy winter images.

Index Terms—Unsupervised estimation, domain adaptation, monocular depth estimation, night, rainy night.

I. INTRODUCTION

Depth is one of the most important information for autonomous systems in perceiving their surroundings and their own states [1], [2]. Therefore, the accurate estimation of the depth information from monocular images has become a hot topic in recent years [3], [4]. Structure from motion and stereo matching are two main ways to recover the depth information based on the geometric relationship between images, and these methods are widely used in traditional SLAM methods to map the environments [5], [6]. With the development of deep learning, using deep neural networks to estimate the pixel-level dense depth from only a single image is becoming possible and has attracted much attention. Recently, different kinds of deep learning-based monocular depth estimation framework have been proposed, including supervised methods and unsupervised methods [7], [8]. Because of the costly ground truth, geometric constraints are gradually replacing ground truth for the training of depth networks, and the unsupervised framework has become a promising direction for monocular depth estimation [8]–[10].

In the unsupervised framework of monocular depth estimation [11], [12], the geometric constraints between adjacent images are considered to supervise the network training. Therefore, only monocular image sequences and camera parameters are needed during the training process. This unsupervised framework is mainly composed of two deep neural networks, including a depth network to regress the dense depth from single images and a pose network to estimate the pose between two frames. Based on the estimated depth map and pose, the geometric relationship between images is built on projection function. The mainly supervised signal is calculated from the photometric error of corresponding pixels between adjacent images by using view reconstruction [12].

However, the above unsupervised framework [11], [12] suffers from two big limitations, the static scenario hypothesis and the photometric consistency hypothesis. For the static scenario hypothesis, since the pixels on moving objects do not satisfy the projection function of camera ego-motion, which leads to incorrect calculation of the loss during training, thereby affecting the accuracy of depth network [13]. Incorporating semantic information into the unsupervised framework is an effective way to recognize moving objects and eliminate their influence, and relevant results have emerged [13], [14]. For the photometric consistency hypothesis, since the
training of the unsupervised framework relies heavily on the photometric error, the photometric consistency of the same pixel on different images is crucial to the overall framework \cite{15}. Since all objects are illuminated by the same light source (sun) during the day, the photometric consistency assumption is basically valid, just like that in traditional direct visual odometer \cite{16}, \cite{17}. Therefore, almost all of the current unsupervised monocular depth estimation methods are trained and tested on day-time images. When the environment changes into other highly complex conditions, like night, snowy winter and especially rainy night, the photometric consistency assumption is not valid any more. Complex static and dynamic light sources (street light and car light) at night \cite{18}, as well as the reflections of road and camera lens caused by rain in rainy night-time scenario, lead to huge photometric changes of the same pixel on different frames. Moreover, the camera exposure and gain in outdoor snowy scenario affect the photometric consistency between frames. The problem of estimating monocular depth in such varying environments is challenging but practical and important. Meanwhile, the perception of changing and complex environments is crucial for autonomous systems \cite{19}, \cite{20}, like robots and autonomous driving cars, and this problem has only received some initial attention \cite{15}, \cite{18}.

Because of the limitations of unsupervised framework in complex scenarios, we tackle this challenging problem by using domain adaptation framework. Instead of training the unsupervised framework on the images from complex environments, we adapt the model trained by day-time images to other complex environments, thereby circumventing the photometric inconsistency in complex environments and achieving satisfactory depth estimation in an unsupervised way in the highly complex environments. Instead of adapting an overall depth network, we only adapt the encoder of the depth network by following \cite{18}. An additional encoder is designed to encode the images of complex environments, and after adaptation, the encoded features come from the same feature space as the features of day-time images encoded by day-time encoder. The adapted encoders for different scenarios share the same decoder for monocular depth estimation in different complex scenarios. Therefore, this method not only reduces computational complexity, but also facilitates practical applications: switching different encoders for adapting various environments. Different from \cite{18} using adversarial domain adaptation in feature space for night-time depth estimation, we propose to adopt the image transfer-based domain adaptation framework and constrain the training from both feature space and output space, which is more stable and accurate \cite{21}–\cite{24}. Besides, the proposed image transfer-based domain feature adaptation (ITDFA) framework not only can be used for nighttime depth estimation, but also shows outstanding performance in more challenging rainy night-time and ordinary snowy winter scenarios, as shown in Fig. \ref{fig:map}. Moreover, the models obtained by ITDFA can extend to adapt to new scenarios, which demonstrates the ability of ITDFA to learn and extract the key features for depth decoding.

In summary, in this paper, we analyze and tackle the unsupervised monocular depth estimation problem in three typical and challenging scenarios (night, rainy night and snowy winter), including proposing the ITDFA framework, constructing novel training/testing sets on different scenarios, digging into the continuous adaptation ability of the ITDFA, and exploring the influence of image transfer model on ITDFA. Our main contributions are as follows:

- This paper analyzes the major reason for the limited performance and application of the current unsupervised monocular depth estimation framework, and we tackle the problem of unsupervised monocular depth estimation in highly complex environments, which is very practical.
- An image transfer-based unsupervised domain feature adaptation framework is proposed for estimating depth from the highly complex scenes, like night, snowy winter, and rainy night. The night-time depth model obtained by ITDFA is further adapted to the rainy night-time scenario, which demonstrates the ability of the proposed framework to learn and extract key features.
- Extensive experiments on RobotCar dataset \cite{25} show the effectiveness and applicability of our proposed method in highly complex environments.

II. RELATED WORK

In this section, we introduce the popular unsupervised monocular depth estimation framework \cite{12}, which is trained on monocular sequences. Firstly, many recent research results for improving this unsupervised framework is briefly reviewed, from the perspectives of occlusions, static scenario hypothesis and photometric consistency hypothesis. Then, we review the framework combined with domain adaptation, in which depth models are trained on synthetic datasets and then adapted to real-world scenarios through domain adaptation.

**Unsupervised framework.** To circumvent the need for costly ground truth, Zhou et al. \cite{12} propose to use geometric constraints between frames instead of ground truth to train depth network. Their framework contains a depth network for monocular depth estimation and a pose network for inter-frame pose estimation. Then, based on the projection function established by the estimated pose and depth, the view reconstruction is designed to warp and construct the target frame from its adjacent frame. The photometric error between the warped and real target images is used to supervise the training process, so that the depth and pose networks are trained in an unsupervised manner. To improve the accuracy of depth estimation, several novel loss functions and network frameworks are proposed, which are well reviewed in \cite{26}.

**Occlusions and moving objects.** Occlusions and moving objects affect the pixel correspondence between images, thus impacting the photometric loss during training and resulting in the limited performance of the depth network. A number of methods \cite{12}, \cite{27}–\cite{29} design a mask or mask network to estimate the regions that violate the projection, so as to reduce the effect of these regions on the training process. Since the mask network is jointly trained with pose and depth networks in an unsupervised manner, this method cannot completely address the influence of occlusions and moving objects. Considering that occlusions are caused by visual field...
changes, Godard et al. [11] propose a minimization of the view reconstruction error to deal with these occlusions, which obtains occluded information by using complementary information from different visual fields. Therefore, their method effectively reduces the influence of occlusions on training and improves the performance of monocular depth estimation. To tackle the effect of dynamic objects, many methods [13], [14], [30] identify and eliminate dynamic objects in the scene by using semantic segmentation or target detection during training. Klingner et al. [14] select the dynamic-class object by using semantic segmentation and adopt the projection error to determine if the dynamic object is moving. Finally, the dynamic regions are eliminated during calculating the photometric error. Instead of removing dynamic regions on images, Casser et al. [30] and Lee et al. [13] accurately model the ego-motion of moving objects by using pose networks. For view reconstruction in unsupervised training, they construct not only static objects but also moving objects based on the estimated ego-motion of objects and the camera. Therefore, their methods do well in highly dynamic scenarios and break the limitation of static scenario hypothesis in unsupervised monocular framework.

**Photometric inconsistency.** Photometric inconsistency is another major reason for the limited performance and application of the unsupervised framework. Similar to the drawbacks faced by direct visual odometry [31], the above unsupervised framework is also limited by the photometric consistency hypothesis. Affected by the camera optical vignetting and exposure controls, the corresponding pixels between day-time images face the problem of photometric inconsistency [17], which results in the performance degradation of the unsupervised framework in outdoor scenarios. To deal with the photometric inconsistency, Yang et al. [32] propose to estimate the brightness transformation parameters to align the photometry of images during training.

Moreover, for some highly complex environments, like night-time environment and more challenging rainy night-time environment, due to the complex lighting conditions, e.g., street lamps, car lights, and especially the reflection of light from the road caused by rain, the unsupervised framework shows unsatisfied accuracy and robustness when applying to such scenarios. The unsupervised monocular depth estimation in such complex scenarios is a largely under explored domain, and only a few methods for night-time monocular depth estimation have been proposed most recently [15], [18]. To overcome the photometric consistency of images, Spencer et al. [15] propose to use dense feature representation of images for unsupervised training. A deep neural network is proposed to extract the feature maps of images, which are consistent under different lighting conditions. Since the corresponding features between different images are consistent and unaffected by light, this unsupervised framework can well adapt to night-time scenarios. However, since the whole framework is unsupervised, their framework still needs the help of photometric error during training. Different from [15], Vankadari et al. [18] regard this challenge as a domain adaptation problem. The depth and pose networks are trained on day-time scenarios by following [11] at first. Then, an additional encoder is designed to encode the night-time images, and an adversarial domain feature adaptation method is used to adapt the features encoded by the day-time encoder and night-time encoder. Finally, the combination of night-time encoder and day-time decoder is used for night-time monocular depth estimation. Since the output of the encoder are multi-scale high-dimensional feature maps, they design multiple discriminators to constrain each scale feature map. However, if any of the discriminators do not converge, the overall performance of their framework will be affected [33]. In addition, adjusting the adversarial framework consisting of multiple discriminators and a generator is extremely complex, which influences its stability for applying to other scenarios [33], [34]. Moreover, although adversarial learning helps to reduce the distance between the distributions of day and night feature spaces, the key features for depth decoding are not valued because the decoder is not involved in their domain adaptation process.

**Domain adaptation.** In monocular depth estimation, domain adaptation algorithms are mainly applied to adapt the model trained on synthetic datasets to real-world datasets [20], [35]–[37]. Compared with the ground truth obtained by different sensors in the real-world, the ground truth obtained from virtual environments is cheaper and easier. The depth model is trained on synthetic and real-world datasets and supervised by the ground truth of synthetic datasets. Since supervised training can get more cues than unsupervised training, this method achieves better accuracy on monocular depth estimation than unsupervised methods, and it provides a new way to circumvent the need for costly ground truth at the same time. While for the night-time scenario, the above framework cannot work because it is difficult to generate synthetic night-times images that can capture all the vagaries of real-world night conditions [18], let alone rainy night-time scenarios and even more complex scenarios. Therefore, to tackle the unsupervised monocular depth estimation in highly complex environments, we use unsupervised domain adaptation to transfer the model trained on day-time images to work for night-time and rainy night-time images. To increase the applicability and reduce the computational complexity, our model only adapts the encoders during training by following [18].

Compared with the most similar work [18], our model ITDFA does not need to consider the stability of adversarial learning, which is conducive to the stability and practicability of the framework. Meanwhile, Vankadari et al. [18] only consider the adversarial constraint on feature space, while this paper constrains the training of the encoder from both feature space and output space. The constraints on output space help encoder to focus on learning the key features of depth decoding. Hence, our model achieves a better performance than [18] in night-time scenario. Moreover, [18] only addresses the unsupervised monocular depth estimation on night-time images, and their model suffers from the instability for other scenarios because of the adversarial learning with multiple discriminators, while our framework can also do well in more challenging rainy night-time scenario and snowy winter scenario.
III. METHODS

In this section, we will introduce the ITDFA framework, which is designed for the unsupervised monocular depth estimation in highly complex environments.

A. Basic pre-trained models

1) Depth model: We use the famous unsupervised framework, monodepth2 [11], to acquire the trained depth network in an unsupervised manner. Monodepth2 [11] has been widely used as the basic unsupervised framework in this field because of its high practicability and accuracy [14], [15], [38]. After using as the basic unsupervised framework in this field because of its high practicability and accuracy [14], [15], [38]. After training the monodepth2 on day-time dataset, the depth model consisting of an encoder $E_d$ and a decoder $D_d$ is used in the proposed ITDFA. The depth model learns a mapping from the day-time images $I_d$ to the pixel-level depth maps $O_d$:

$$O_d = D_d(E_d(I_d)).$$

(1)

2) Image transfer model: Besides, for different complex environments, different pre-trained image transfer models ($G_{d2X}$ and $G_{X2d}$) based on CycleGAN [39] are needed to transfer the images between day-time style $d$ and different complex environmental styles $X$:

$$\begin{cases} I_{d2X} = G_{d2X}(I_d), & \text{day to } X \\ I_{X2d} = G_{X2d}(I_X), & \text{X to day} \end{cases}$$

(2)

where $X$ refers to night $n$, rainy night $r$ and snowy winter $s$. In addition, to verify the continuous transfer ability of the model obtained through domain adaptation, we also train an additional image transfer model between night-time style and rainy night-time style ($G_{n2r}$ and $G_{r2n}$).

B. ITDFA framework

1) Training framework: As shown in Fig. 2 an encoder $E_X$ is designed to encode the features of images from highly complex scenarios to the same feature space as the features of day-time images encoded by the day-time encoder $E_d$. In ITDFA, the pre-trained day-time encoder $E_d$ is used to encode the day-time images and obtain their corresponding feature maps $f$:

$$\begin{cases} f_d = E_d(I_d) \\ f_{X2d} = E_d(\hat{I}_{X2d}) \end{cases}$$

(3)

where $I_d$ refers to the real day-time images, and $\hat{I}_{X2d}$ stands for the fake day-time images generated by CycleGAN model $G_{X2d}$ from the highly complex scenario $X$. The encoder $E_X$ for complex environments has the same network framework as the day-time encoder $E_d$, and it is used to encode the real and fake images of highly complex scenarios and obtain their feature maps $f$:

$$\begin{cases} f_X = E_X(I_X) \\ f_{d2X} = E_X(\hat{I}_{d2X}) \end{cases}$$

(4)

where $I_X$ refers to the real images from highly complex scenarios, and $\hat{I}_{d2X}$ stands for the fake images transferred from day-time scenario $d$. The pre-trained decoder $D_d$ is used to decode the features from $E_d$ and $E_X$ and obtain their corresponding depth maps $O$:

$$\begin{cases} O_d = D_d(f_d) \\ O_{X2d} = D_d(f_{X2d}) \\ O_X = D_d(f_X) \\ O_{d2X} = D_d(f_{d2X}) \end{cases}$$

(5)

During training, the weights of $E_d$ and $D_d$ are fixed, and only $E_X$ is updated.

Fig. 2. The framework of ITDFA for unsupervised monocular depth estimation in highly complex environments. $I_d$ and $I_X$ refer to the image from day-time scenario and one of the complex scenarios, and $O_d$ and $O_X$ stand for their corresponding depth maps. The encoder $E_d$ and decoder $D_d$ trained by day-time images do not update their weights during training.
During testing, the depth map can be estimated from the images of highly complex scenarios in one-step:

\[ O_X = D_d(I_d(I_X)) . \]  

2) **Training losses:** The ITDFA framework is an unsupervised framework, and neither ground truth nor real paired images are used to train the depth model, CycleGAN model and domain adaptation model. As shown in Fig. 2 different constraints are designed to supervise the training process, including the feature consistency loss \( L_{FC} \) on feature space, and the depth consistency loss \( L_{OC} \) as well as smoothness loss \( L_{OS} \) on output (depth) space. Therefore, the overall loss function for training the encoder \( E_X \) is formulated as:

\[ L_{\text{total}} = \alpha L_{FC} + \beta L_{OC} + \gamma L_{OS}. \]  

**Feature consistency loss:** Based on the pre-trained CycleGAN model, we can get the paired images from day-time scenario and highly complex scenario, like \( I_d \) with \( I_{d2X} \) and \( I_X \) with \( I_{X2d} \). Therefore, to promote the consistency of different encoders in feature space, we directly minimize the error of the feature maps, which are encoded by \( E_d \) and \( E_X \) from these image pairs:

\[ L_{FC} = \Psi(f_d,f_{d2X}) + \Psi(f_X,f_{X2d}). \]  

**Depth consistency loss:** Although the above loss can help to promote the consistency of feature space between the two encoders, the ultimate goal is the depth map rather than the feature map, and the contribution of different features to the depth decoding is different. To further constrain the key features of feature maps for depth decoding, we design a depth consistency loss in output space:

\[ L_{OC} = \Psi(O_d,O_{d2X}) + \Psi(O_X,O_{X2d}), \]  

where

\[ \Psi(A,B) = | |A - B||_1. \]

**Smoothness loss:** Moreover, to promote the smoothness of the generated depth map, we propose to utilize the edge-aware smoothness during training, which is widely used in previous unsupervised depth framework [9], [11], [40].

\[ L_{OS} = | \partial_1 O_{d2X} | e^{\partial_1 I_d} + | \partial_2 O_{d2X} | e^{\partial_2 I_d}, \]

where \( O_{d2X} = O_{d2X}/O_{d2X} \) represents the mean-normalized inverse depth. Note that this loss is established between the real day-time image \( I_d \) and the depth map \( O_{d2X} \) of its corresponding transferred images \( I_{d2X} \).

Why not adopt a framework that directly combines image style transfer with monocular depth estimation: the images from complex environments are first transferred to normal day-time style and then use the day-time depth model to estimate the depth. After testing, we find this is a possible way to solve the problem, as shown in Fig. 3 and this is also the reason why our ITDFA framework can work. However, the accuracy of depth estimation relies heavily on the quality of transferred images, so this framework has great instability in practical applications, as shown in Fig. 3. Then, poor real-time performance will also limit the application of this framework because of the two-step process. Moreover, Vankadari et al. [18] have shown that the above framework is not as accurate as the domain adaptive framework.

### IV. Experiments

#### A. Datasets

Since this paper focuses on the unsupervised monocular depth estimation in multiple highly complex environments, we choose the publicly available Oxford RobotCar dataset [25] as our training and testing sets. RobotCar dataset [25] is one of the most famous outdoor datasets, and it contains the image sequences collected in all weather conditions, including rain, night, direct sunlight and snow. The image sequences captured by the left camera of Bumblebee XB3 are used for the experiments of this paper. The images are manipulated to RGB style from the raw recordings with the resolution of 1280x960, and we crop the car-hood of the images and resize them to 512x256. For the day-time and night-time scenarios, we use the sequences from 2014-12-09-13-21-02 and 2014-12-16-18-44-24, which are the same as [18] for a fair comparison. For the rainy night-time and snowy winter scenarios that have not received attention in recent research, we choose the sequences from 2014-12-17-18-18-43 and 2015-02-03-08-45-10.

#### B. Training and testing sets setup

**Training sets:** For the day-time depth model, the 5 splits of day-time sequence are used to train the unsupervised framework [11], and the basic pre-trained depth model, monodepth2 (day), is obtained for ITDFA. To improve the performance of this depth model, 15,000 images are uniformly selected from 5 splits for training, and the training set does not include the images taken while parking. For the image transfer model, 6,000 images of each scenario are selected to obtain the image transfer models between different scenarios by using CycleGAN [39]. The selection of these images follows the rules before. To address the problem of unsupervised monocular depth estimation in night-time, rainy night-time and...
TABLE I
Comparison with the unsupervised depth estimation methods for night-time scenarios. “M” means that the supervisory signals to train the framework comes from monocular sequences.

| Method                        | Supervision | Depth-range (meter) | Abs Rel | Sq Rel | RMSE | RMSE log | δ < 1.25     | δ < 1.25     | δ < 1.25     |
|-------------------------------|-------------|---------------------|---------|--------|------|----------|--------------|--------------|--------------|
| Vankadari et al. [18]         | M           | 40                  | 0.2005  | 2.3750 | 7.172| 0.278    | 0.735        | 0.883        | 0.942        |
| ITDFA [41]                   | M           | 40                  | 0.1656  | 1.3902 | 5.458| 0.226    | 0.764        | 0.912        | 0.964        |

| Method                        | Supervision | Depth-range (meter) | Abs Rel | Sq Rel | RMSE | RMSE log | δ < 1.25     | δ < 1.25     | δ < 1.25     |
|-------------------------------|-------------|---------------------|---------|--------|------|----------|--------------|--------------|--------------|
| Vankadari et al. [18]         | M           | 60                  | 0.2327  | 3.783  | 10.089| 0.319    | 0.668        | 0.844        | 0.924        |
| ITDFA [41]                   | M           | 60                  | 0.1986  | 2.104  | 7.725| 0.260    | 0.690        | 0.802        | 0.955        |

TABLE II
Comparison with the unsupervised depth estimation methods for rainy night-time scenarios.

| Method                        | Supervision | Depth-range (meter) | Abs Rel | Sq Rel | RMSE | RMSE log | δ < 1.25     | δ < 1.25     | δ < 1.25     |
|-------------------------------|-------------|---------------------|---------|--------|------|----------|--------------|--------------|--------------|
| Monodepth2 (d) [11]           | M           | 40                  | 0.4297  | 5.878  | 14.945| 0.386    | 0.616        | 0.837        | 0.915        |
| Monodepth2 (r) [11]           | M           | 40                  | 0.3902  | 50.828 | 14.945| 0.386    | 0.616        | 0.837        | 0.915        |
| ITDFA [41]                   | M           | 40                  | 0.1753  | 1.3891 | 5.146| 0.235    | 0.739        | 0.922        | 0.973        |
| ITDFA [41]                   | M           | 40                  | 0.1879  | 1.2211 | 4.923| 0.246    | 0.697        | 0.918        | 0.974        |

Fig. 4. A qualitative comparison of predicted depth-maps with different experiments. X refers to the night-time scenario (lines 1, 2), rainy night-time scenario (lines 3, 4) and snowy winter scenario (lines 5, 6). The first column includes the images from different complex environments X. “Monodepth2(X)" stands for the monodepth2 model trained on the images from X. The final column shows the results of ITDFA for different complex environments X.

snowy winter scenarios, three image transfer models should be pre-trained for ITDFA: between day and night (Gd2n and Gr2d), between day and rainy night (Gd2r, and Gd2r), and between day and snowy winter (Gd2w, and Gr2w). In addition, to verify the continuous transfer ability of the model obtained through domain adaptation, we also train an additional image transfer model between night and rainy night (Gn2r, and Gd2r). During training the ITDFA, the training sets used for domain adaptation are the same as that of the image transfer model.

Testing sets: During testing, 300 rainy night-time images and 300 snowy winter images are randomly selected from the remaining splits of each sequence to test the model obtained by ITDFA. While for the night-time scenario, the testing set is the same as [18] for a fair comparison. The evaluation metrics used in this paper follow previous methods [11], [12], [26], and we evaluate the depth models from the aspect of error and accuracy.

C. Experimental setup
The experiments are implemented by using Pytorch learning framework [41] on an NVIDIA RTX 2080 Ti GPU. For pre-training the depth model and image transfer model, we use the original setting proposed in [11] and [39] and resize the image resolution to 512x256. For the ITDFA, the framework is trained for 30 epoches using Adam optimizer [42] with a batch size of 1 and an input/output resolution of 512x256. The learning rate is set as 0.0002 during training. The weights of the three loss components are set to α = 1.0, β = 0.01 and γ = 0.01. The experimental setup of ITDFA is the same when applied for different complex scenarios, including the night-time, rainy night-time and snowy winter scenarios. Moreover, the encoders trained for different complex scenarios share the same day-time decoder. Therefore, in practice, this framework only needs to switch the corresponding encoder to cope with the change of environments, which is practical and meaningful for applying in changing environments.

D. Results
1) Annotation: Results of related experiments are shown in Tables [III] and [IV]. To the best of authors’ knowledge, since the proposed work may be the first attempt at solving the monocular depth estimation problem in rainy night-time scenario and snowy winter scenario, and no priors are available in the literature, we compare the results of different models based on the well known unsupervised framework, monodepth2 [11].

In the tables, “Monodepth2 (X)” refers to the depth model trained on the images from X by the monodepth2 framework.
Fig. 5. Estimating the depth from night-time images. “Monodepth2(n)” refers to the depth model trained on night-time images by the unsupervised monocular framework, monodepth2. For night-time scenario, it suffers from complex lighting conditions, like the street lights, car lights, and the reflection of light from the road, which bring big challenges to the unsupervised training framework.

### TABLE III
Comparison with the unsupervised depth estimation methods for snowy winter scenario.

| Method            | Supervision | Depth-range (meter) | Abs Rel | Sq Rel | RMSE | RMSE log | δ < 1.25 | δ < 1.25^2 | δ < 1.25^3 |
|-------------------|-------------|---------------------|---------|--------|------|----------|---------|-----------|-----------|
| Monodepth2 (d)    | M           | 40                  | 0.2794  | 2.3635 | 6.416| 0.342    | 0.627   | 0.835     | 0.939     |
| Monodepth2 (s)    | M           | 40                  | 0.2205  | 2.4044 | 6.406| 0.309    | 0.709   | 0.876     | 0.917     |
| ITDFA_{d2n}       | M           | 40                  | 0.1964  | 1.5744 | 5.275| 0.260    | 0.719   | 0.898     | 0.964     |
| Monodepth2 (d)    | M           | 60                  | 0.2698  | 2.9862 | 7.961| 0.335    | 0.585   | 0.835     | 0.928     |
| Monodepth2 (s)    | M           | 60                  | 0.2399  | 2.9308 | 7.832| 0.330    | 0.662   | 0.856     | 0.929     |
| ITDFA_{d2r}       | M           | 60                  | 0.2119  | 1.9792 | 6.592| 0.281    | 0.683   | 0.878     | 0.955     |

[11], and X refers to day (d), night(n), rainy night (r) and snowy winter (s). ITDFA_{d2X} represents the depth model trained by the proposed ITDFA for the complex scenario X, and these models (ITDFA_{d2n}, ITDFA_{d2r}, ITDFA_{d2s}) are adapted from the same day-time model, “Monodepth2 (d)”. Moreover, ITDFA_{d2n2r} stands for the rainy night-time model adapted from the night-time model, ITDFA_{d2n}, which means that the model is obtained through two domain adaptations by ITDFA. Note that all the models obtained by the proposed ITDFA framework share the same decoder with the “Monodepth2 (d)”. The qualitative results are shown in Figs. 4 and 5.

2) **Night-time depth estimation:** To verify the effectiveness of the proposed ITDFA framework in night-time scenario, we compare our model with the state-of-the-art method [18], which only focuses on unsupervised night-time depth estimation. As shown in Table I for estimating the depth from night-time images, the depth model trained by ITDFA gets a much better performance than the current state-of-the-art method [18] with lower error and higher accuracy.

3) **Rainy night-time depth estimation:** As shown in Table [11] because of the domain drift, the monodepth2(d) does not have a good performance when testing on the rainy night-time images. Compared with the day-time and night-time scenarios, the rainy night-time scenario suffers from not only photometric inconsistency but also the reflection of road and camera caused by rain. Therefore, the monodepth2(r) cannot get an accurate depth estimation because of the limitation of the unsupervised framework [11] in such complex scenario. The rainy night-time depth model trained by ITDFA (ITDFA_{d2r}) shows a much better performance than the monodepth2 models, which proves the effectiveness of the framework proposed in this paper.

Because the domain gap between night and rainy night is smaller than that between day and rainy night, ITDFA_{d2n} shows more accurate depth estimation than the monodepth2(d) on rainy night-time images. To study the influence of different domain gaps on adaptation, we additionally design a two-step adaptation method, in which the day-time depth model is firstly adapted to night-time scenario, and then the night-time model is adapted to rainy night scenario, shown as ITDFA_{d2n2r}. As shown in Table II even if after the second...
round adaptation, ITDFA$_{d2n2r}$ achieves an outstanding accuracy than others in rainy night-time scenario, which means that multi-step adaptation will be helpful for the domain adaptation between large domain gap. Besides, it also proves that the proposed ITDFA framework has the ability to effectively learn the key features for depth decoding. The qualitative results are shown in Figs. 4 and 6.

4) Snowy winter depth estimation: In the snow-covered environment during winter, the unsupervised framework suffers from repeat-texture scenes, and the camera exposure and gain in outdoor snowy scenario affects the photometric consistency between frames, resulting in the limited performance of unsupervised framework. Therefore, as shown in Table III, the adaptive model obtained by ITDFA gets a better performance in snowy winter scenarios than the models trained by monodepth2. The qualitative results are shown in Fig. 4 and Fig. 7, although the model “Monodepth2” looks better than our proposed method in details, the depth of some objects, like cars, is incorrect according to the relative depth. Meanwhile, due to different image brightness caused by the camera exposure and gain, the depth estimation of the similar road is significantly different, as shown in Fig. 4. Quantitative results in Table III prove that our method can achieve good results in snowy winter. ITDFA achieves more robust depth estimation, but monodepth2 outperforms ITDFA in the depth presentation of details.

5) Ablation study: To analyze the effects of each component in the overall loss function $L_{\text{total}}$, Eq. (7), we design a series of ablations to analyze our approach, and quantitative results are shown in Table IV. Experiments show that the constraint of feature space is more effective than that of output space in promoting the consistency of feature maps. The model trained by the constraints from both feature space and output space outperforms the others, which means that the constraints of output space help encoder to focus on learning the key features of depth decoding during training. Moreover, the smoothness loss helps to improve the accuracy of depth estimation in complex environments.

E. Discussion

The ITDFA is an unsupervised framework, and the depth models for highly complex scenarios are trained in a completely unsupervised manner. Note that all the monocular depth models trained by ITDFA for different environments share the same decoder during testing, which has practical significance. For example, in autonomous driving, facing different weather conditions [43], the vehicle can independently switch to the corresponding encoder to obtain better environmental perception.

Although the ITDFA shows an outstanding performance in various highly complex environments, there are still some
shortcomings in this framework. Firstly, the depth and 3D structure of some objects, like cars and trees, are not well decoded by the depth model, as shown in Figs. 5, 6, 7. Meanwhile, the performance of ITDFA will be influenced by the performance of the image transfer model, and higher quality transferred images will result in a better depth model.

To explore the impact of image transfer performance on ITDFA, we select 5 different image transfer models saved from the last 50 epochs of CycleGAN training to participate in the training of ITDFA. The experiments are built on nighttime images, and the image transfer models are trained to transfer the images between day-time style and night-time style. As shown in Table V, different image transfer models lead to different image style transfer performances, which has different effects on the depth model adapted by ITDFA. However, the effect of the image transfer model on ITDFA is limited, and the error metrics of the obtained depth models do not fluctuate greatly. Moreover, all of the depth models get better performance than the state-of-the-art night-time depth estimation method [18], which is a good proof of the effectiveness of our ITDFA framework.

**TABLE V**

| Error Metrics (Lower is better) | Loss | Abs Rel | Sq Rel | RMSE | RMSE log |
|--------------------------------|------|---------|--------|------|----------|
| ITDFA\_IT1                   | 0.1767 | 1.4412 | 5.529  | 0.240|          |
| ITDFA\_IT2                   | 0.1809 | 1.4541 | 5.563  | 0.242|          |
| ITDFA\_IT3                   | 0.1656 | 1.3902 | 5.458  | 0.226|          |
| ITDFA\_IT4                   | 0.1775 | 1.1567 | 5.111  | 0.240|          |
| ITDFA\_IT5                   | 0.1643 | 1.3693 | 5.429  | 0.226|          |
| Vankadari et al. [18]        | 0.2005 | 2.5750 | 7.172  | 0.278|          |

Fig. 7. *Estimating the depth from the images of snowy winter.* “Monodepth2(s)” refers to the depth model trained on snowy winter images by the unsupervised monocular framework, monodepth2. For snowy winter scenario, because the reflection of the snow will increase the brightness of the scene, this will result in the changes of camera exposure and gain, thereby resulting in photometric inconsistency between frames. The first two lines are two examples that ITDFA outperforms monodepth2, and the last two lines are two examples that monodepth2 seems better than ITDFA.

**Fig. 8. An example of application framework.** Multiple encoders share the same decoder to adaptively perceive the depth of different environments.

**F. Application Prospect**

The changing environment has always been a big challenge for the perception of autonomous systems because existing algorithms can only work in certain scenarios. Because of the domain shift, it is difficult to apply a single depth model to all scenarios with satisfactory performance. Note that all of the depth models obtained by ITDFA share the same decoder. Therefore, this paper provides a solution for dealing with the depth perception of changing environments: switching to the corresponding encoder according to the current environment, as shown in Fig 8, which is also more convenient than switching to the corresponding depth models.

**V. Conclusion**

In this paper, we tackle the problem of unsupervised monocular depth estimation in highly complex environments, which is important and practical for autonomous systems. We survey the related research on solving the limitations of
current unsupervised monocular depth estimation framework, and analyze the reason why the unsupervised framework cannot do well in certain highly complex environments. A novel domain adaptation framework, called ITDFA, is proposed in this paper to address the above problem. The proposed ITDFA framework achieves more accurate depth estimation in night-time scenario than the state-of-the-art [18]. Moreover, the performance of ITDFA in the snowy winter scenario and more challenging rainy-time scenario proves the practicability and effectiveness of ITDFA. Therefore, ITDFA is able to provide a way to address the complex environmental change problems faced by monocular depth estimation during the practical application. Furthermore, there are still shortcomings that need to be addressed, like enhancing the depth perception of some small objects in complex environments, which is also a promising direction for future work.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] F. Tian, Y. Gao, Z. Fang, Y. Fang, J. Gu, H. Fujita, and J.-N. Hwang, “Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint,” IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[3] M. Song, S. Lim, and W. Kim, “Monocular depth estimation using laplacian pyramid-based depth residuals,” IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[4] S. Abdullawahab, H. A. Rashwan, M. A. Garcia, M. Jabreel, S. Chambon, and D. Puig, “Adversarial learning for depth and viewpoint estimation from a single image,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 9, pp. 2947–2958, 2020.

[5] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision. Cambridge University Press, 2003.

[6] R. Mur-Artal and J. D. Tardos, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.

[7] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in Neural Information Processing Systems, 2014, pp. 2366–2374.

[8] S. Chen, Z. Pu, X. Fan, and B. Zou, “Fixing defect of photometric loss for self-supervised monocular depth estimation,” IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[9] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 270–279.

[10] M. Poggi, F. Aletti, F. Tosi, and S. Mattoccia, “On the uncertainty of self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3227–3237.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3828–3838.

[12] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1851–1858.

[13] S. Lee, S. Im, S. Lim, and I. S. Kweon, “Learning monocular depth in dynamic scenes via instance-aware projection consistency,” arXiv preprint arXiv:2102.02629, 2021.

[14] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, “Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance,” in Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 582–600.

[15] J. Wang, E. R. Cowie, and S. Hothoi, “DefNet: General monocular depth via simultaneous unsupervised representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14 402–14 413.

[16] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611–625, 2017.

[17] 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, vol. 3, no. 4, pp. 2878–2885, 2018.

[18] M. Vankadari, S. Garg, A. Majumder, S. Kumar, and A. Behera, “Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation,” in Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 443–459.

[19] Y. Tang, C. Zhao, J. Wang, C. Zhang, Q. Sun, W. Zheng, W. Du, F. Qian, and J. Kurths, “An overview of perception and decision-making in autonomous systems in the era of learning,” arXiv preprint arXiv:2001.02319, 2020.

[20] C. Zhang, J. Wang, G. G. Yen, C. Zhao, Q. Sun, Y. Tang, F. Qian, and J. Kurths, “When autonomous systems meet accuracy and transferability through ai: A survey,” Patterns (Cell Press), vol. 1, no. 4, art. No. 100050, 2020.

[21] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, “Cudoco: Pixel-level domain transfer with cross-domain consistency,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1791–1800.

[22] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” in International Conference on Machine Learning (ICML). PMLR, 2018, pp. 1989–1998.

[23] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, “Diversify and match: A domain adaptive representation learning paradigm for object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12 456–12 465.

[24] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, “Fully convolutional adaptation networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6810–6818.

[25] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000km: The Oxford RobotCar Dataset,” The International Journal of Robotics Research (IJRR), vol. 36, no. 1, pp. 3–15, 2017.

[26] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, “Monocular depth estimation based on deep learning: An overview,” Science China Technological Sciences, vol. 63, no. 9, pp. 1612–1627, 2020.

[27] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, “Unsupervised scale-consistent depth and ego-motion learning from monocular video,” in Advances in Neural Information Processing Systems, 2019, pp. 35–45.

[28] S. Wang, H. Wang, Y. Liu, and W. Chen, “Unsupervised learning of monocular depth and ego-motion using multiple masks,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 4724–4730.

[29] C. Zhao, G. G. Yen, Q. Sun, C. Zhang, and Y. Tang, “Masked GAN for unsupervised depth and pose prediction with scale consistency,” IEEE Transactions on Neural Networks and Learning Systems, Accepted, 2020.

[30] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Unsupervised monocular depth and ego-motion learning with structure and semantics,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 381–388.

[31] L. von Stumberg, P. Wenzel, Q. Khan, and D. Cremers, “GN-Net: The gauss-newton loss for multi-weather relocalization,” IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 891–897, 2020.

[32] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, “D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1281–1292.

[33] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, “Improved techniques for training gans,” arXiv preprint arXiv:1606.03498, 2017.

[34] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, “Improved techniques for training gans,” arXiv preprint arXiv:1606.03498, 2017.

[35] K. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” arXiv preprint arXiv:1706.08500, 2016.

[36] J. P. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, “Adadepth: Unsupervised content congruent adaptation for depth estimation,” in
[37] S. Zhao, H. Fu, M. Gong, and D. Tao, “Geometry-aware symmetric domain adaptation for monocular depth estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9788–9798.

[38] A. Johnston and G. 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), 2020, pp. 4756–4765.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

[40] C. Wang, J. M. Buenaposada, Z. Rui, and S. Lucey, “Learning depth from monocular videos using direct methods,” in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp. 2022–2030.

[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[43] W.-T. Chu, X.-Y. Zheng, and D.-S. Ding, “Camera as weather sensor: Estimating weather information from single images,” Journal of Visual Communication and Image Representation, vol. 46, pp. 233–249, 2017.