2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

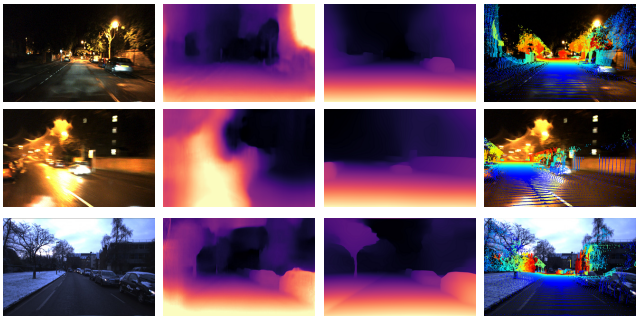
Self-supervised Monocular Depth Estimation in Challenging Environments Based on Illumination Compensation PoseNet

Shengyu Hou, Wenjie Song, Rongchuan Wang, Meiling Wang, Yi Yang, Mengyin Fu

***Abstract*— Self-supervised depth estimation has attracted much attention due to its ability to improve the 3D perception capabilities of unmanned systems. However, existing unsupervised frameworks rely on the assumption of photometric consistency, which may not hold in challenging environments such as night-time, rainy nights, or snowy winters due to complex lighting and reflections, resulting in inconsistent photometry across different frames for the same pixel. To address this problem, we propose a self-supervised monocular depth estimation unified framework that can handle these complex scenarios, which has the following characteristics: (1) an Illumination Compensation PoseNet (ICP) is designed, which is based on the classic Phong illumination theory and compensates for lighting changes in adjacent frames by estimating per-pixel transformations; (2) a Dual-Axis Transformer (DAT) block is proposed as the backbone network of the depth encoder, which infers the depth of local repeat-texture areas through spatial-channel dual-dimensional global context information of images. Experimental results demonstrate that our approach achieves state-of-the-art depth estimation results in complex environments on the challenging Oxford RobotCar dataset.**

I. Introduction

Depth estimation has wide applications in many artificial intelligence tasks, from augmented reality to 3D reconstruction, from autonomous driving to robot positioning and mapping. Compared to active sensing sensors such as Time-Of-Flight (ToF) and Laser Imaging Detection and Ranging (LIDAR), cameras have the unique advantages of low cost and high information richness, making image-based depth estimation a promising research direction. Among learning-based depth estimation methods, self-supervised monocular depth estimation[1], [2], [3] has been widely applied due to its lack of expensive ground truth depth labeling and its independence from the long baseline constraint of binocular vision. In addition, with the efforts of [4], [5], [6], self-supervised depth estimation methods have achieved comparable performance to supervised methods on widely used benchmarks such as KITTI[7], Cityscapes[8], etc. Unfortunately, these methods usually only solve the depth estimation problem under relatively friendly conditions. In some challenging environments, such as night-time, rainy nights, or snowy winters, due to complex lighting and reflections, the luminance of the same pixel in adjacent frames is not consistent, making it hard for day-time unsupervised frameworks to be directly applied to these challenging scenes. However, in the field of autonomous driving, it is crucial to ensure that vehicles have accurate depth perception capabilities under the aforementioned adverse weather conditions.



(a) Input image (b) Monodepth2 (c) Ours (d) Ground Truth

Fig. 1. The depth predictions of the proposed method against the baseline method at night, rainy night and snowy winter. Although the adverse conditions such as non-uniform point sources and reflections of rain and snow on the camera and road violate the assumption of photometric loss, our method can successfully estimate accurate depth maps.

In fact, there are two main challenges in these complex scenes as shown in Fig. 1 : (i) night and rainy night scenes often suffer from varying point sources (such as street lights and car headlights) and glare caused by rain reflecting off the camera and road, resulting in large variations in luminance between different frames for the same pixel, leading to erroneous pixel matching and depth estimation. (ii) In winter snow environments, there are local repeat-texture scenes that lead to inaccurate pixel matching, and camera exposure in outdoor snow scenes also affects inter-frame photometric consistency. For the first challenge, existing unsupervised depth estimation methods[15], [2], [16] rely heavily on photometric error during training, and the consistency of photometry for the same pixel in adjacent frames is extremely important. However, when the environment changes to other highly complex conditions, such as night-time, winter snow, and especially rainy nights, the assumption of photometric consistency is no longer valid. For the second challenge of local repeating texture scenes caused by winter snow, existing CNN-based frameworks cannot cope with the challenges posed by local repeat-texture due to a lack of global perception and long-term relationships between model pixels. However, conventional Transformer-based frameworks[19], [20], [21] only consider pixel long-term dependencies in the spatial dimension, and global information in the channel dimension is also important for pixel matching in local repeating textures.

To address the above two problems, we propose a simple

This work was partly supported by Program for National Natural Science Foundation of China (Grant No. 62373052, 92370203,62233002), Youth Talent Promotion project of China Association for science and technology, Beijing Institute of Technology Research Fund Program for Young Scholars.
The authors are with the School of Automation, Beijing Institute of Technology, Beijing 100081, P.R.China. Corresponding author: Wenjie Song (email: songwj@bit.edu.cn). M. Fu is also with School of Automation, Nanjing University of Science and Technology, Nanjing 210094, P.R.China.

but effective self-supervised monocular depth estimation unified framework that can handle these complex scenarios. Firstly, for the problem of violating the photometric consistency assumption caused by complex point sources during night-time and rainy nights, we propose an Illumination Compensation PoseNet (ICP) based on classical Phong illumination theory, which compensates for the illumination changes between adjacent frames by estimating pixel-wise transformations. Secondly, we propose a Dual-Axis Transformer (DAT) block as the backbone network of the depth encoder to address the challenge posed by unsupervised frameworks caused by erroneous pixel matching between adjacent frames due to local repeat-texture present in winter snow environments. The Dual-Axis Transformer block utilizes the spatial-channel two-dimensional global self-attention information of winter snow images to infer the depth of local repeat-texture regions. In addition, in order to achieve communication between the spatial-channel axes to capture the strong correlation between two-dimensional features, the weights of queries and keys are shared across branches in DAT, which also helps reduce the size of the model. In short, the contributions of our paper can be summarized as follows:

1) A self-supervised unified framework for monocular depth estimation under challenging environments is proposed, which does not require the introduction of additional encoders like models based on transfer learning, and achieves state-of-the-art results on multiple challenging scenarios in the RobotCar dataset.
2) An illumination compensation PoseNet (ICP) based on the classic Phong illumination model is proposed to compensate for pixel-level brightness variations caused by point light sources in complex scenes. It significantly boosts the estimation performance of the influence area of point light sources at night/rainy night.
3) A Dual-Axis Transformer (DAT) block is proposed as the backbone of the depth encoder, which infers the depth of local repeat-texture areas through spatial-channel dual-dimensional global context information of images.

II. Background

Using a single RGB image to obtain accurate depth information is challenging, and the development of deep learning has alleviated this difficulty. Eigen[10] first proposed an end-to-end supervised learning method, which consists of two-scale networks. Subsequently, supervised learning-based models have continuously promoted the development of technology through various improvements [11], [12], [13]. However, acquiring a large number of ground truth depth value annotations is expensive, and self-supervised learning methods have recently been shown to be a promising direction to circumvent this major limitation. Garg et al. [14] have made pioneering explorations in the field of self-supervised depth estimation in 2016, proposed a stereo framework that considers multiple geometric constraints losses. Instead of using stereo image pairs, Zhou et al.[15] proposed using consecutive monocular images to train and estimate self-motion and depth. On this basis, Monodepth2[2] introduces automatic masking and minimum reprojection loss to solve the problem of moving objects and occlusion. Tian et al.[16] takes advantage of the quadtree constraint to optimize the depth estimation network. Due to the lack of global long-term dependencies in the CNN-based framework, the performance of depth estimation in locally repeat-texture regions is limited.

ViTs have recently shown excellent results on various visual tasks[17], [18], thanks to their ability to establish long-range relationships between pixels, thereby forming a global receptive field. Some works[19], [20], [21] also solved the problem of monocular depth estimation by using the Transformer architecture. Such as [19] employed a convolution-free Swin Transformer as an image feature extractor for depth estimation, [20] proposed a light-weight self-supervised monocular depth estimation model with a hybrid CNN and Transformer architecture.

The above methods all use photometric loss as the main supervision signal, and assume that the illumination of adjacent frames is consistent, but this is not valid at night-time and rainy nights. Some works attempts to solve the problem from the perspective of hardware by estimating night-time depth through thermal images[9], [22], but thermal cameras have relatively low resolution. Meanwhile, [23] and [24] are based on adversarial learning, treating night-time depth estimation as a domain adaptation problem, training an independent encoder to generate 'day-like' features from night-time images. ADDS-DepthNet[25] proposed a domain-separated network for self-supervised depth estimation of night-time images. RNW[26] enhances the brightness and contrast of images while maintaining brightness consistency through a mapping consistency image enhancement module to address the problem of low visibility in the dark. All of these methods either require the introduction of additional encoders[23], [24] or require learning prior feature distributions[26] from reference images. In contrast, we propose a unified framework for night-time, winter snow, and rainy nights depth estimation based on illumination compensation PoseNet, without the need to train additional transfer encoders or learn additional features.

III. Method

A. Problem Formulation

In self-supervised depth estimation, the learning problem can be viewed as a view reconstruction problem. It uses a DepthNet f_D to learn depth information D_t from the input RGB target image I_t, and an additional PoseNet f_P to estimate relative pose $T_{t\to s}$ between source image I_s and target image I_t. The estimated depth and relative pose can be used to acquire a per-pixel correspondence between an arbitrary point p_t in I_t and another point p_s in I_s by:

$$p_s \sim K T_{t\to s} D_t(p_t) K^{-1} p_t, \tag{1}$$

where $K \in R^{3\times 3}$ denotes the camera intrinsic parameter. Then, we use the differentiable bilinear sampling operation

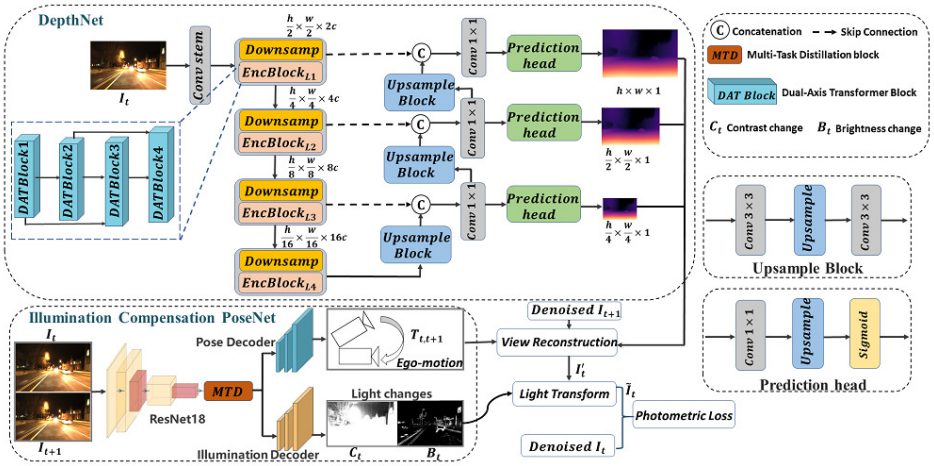


Fig. 2. Architecture overview of this work, which including: DepthNet and Illumination Compensation PoseNet. Illumination Compensation PoseNet is a Shared encoder-Dual decoder network, where the Pose Decoder estimates the pose transformation of adjacent frames, and the Illumination Decoder estimates the pixel-wise illumination change of adjacent frames.

[29] $s(\cdot,\cdot)$ to reconstruct I_t from I_s:

$$I'_t = s(I_s, p_s), \tag{2}$$

Moreover, unlike existing unsupervised frameworks, due to the lack of brightness consistency between the target frame I_t and the reconstructed frame I'_t during challenging environments, which violates the photometric loss requirement, we compensate for the brightness difference of I'_t in PoseNet by using a illumination decoder based on the classic Phong illumination model, which can be expressed as:

$$\tilde{I}_t \approx C_t \odot I'_t + B_t, \tag{3}$$

where C_t represents the contrast(scale) change and B_t represents the brightness (shift) change, and $\tilde{I}_t$ represents a reconstructed image with illumination consistent with the target image I_t after brightness transformation.

B. DepthNet

As shown in Fig. 2 , the proposed framework consists of a DepthNet and a Illumination Compensation PoseNet. Our model backbone DepthNet is based on the previous Uformer, which has four stages of generating different scales. Each stage of the encoder has a similar structure, which consists of L_i Dual Attention Transformer (DAT) blocks and a down-sampling/up-sampling layer.

1) Dual-Axis Transformer Block: Existing unsupervised depth estimation frameworks based on Transformer only consider pixel long-term dependencies in the spatial dimension, but global information in the channel dimension is also important for addressing the problem of pixel mis-matching during view synthesis caused by locally repeated textures in outdoor snow environments. Therefore, to infer the depth of local repeat-texture areas by global context information, the proposed Dual Attention Transformer block can not only obtain rich long range dependencies, but also take into account the spatial-channel dimension, and reduce the computational complexity to linear in the spatial dimension. The DAT block has an important attribute is that it can be shared keys-queries

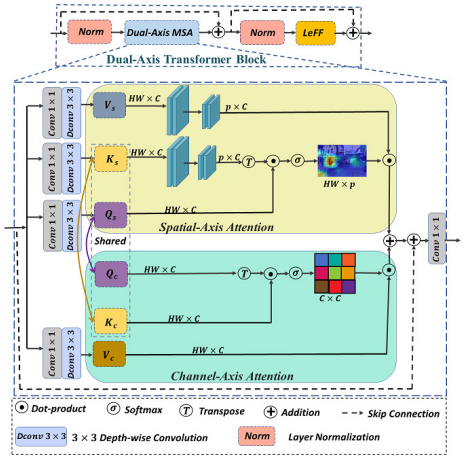


Fig. 3. Illustration of Dual-Axis Transformer Block (DAT), which can not only infer the depth of local repeat-texture areas through spatial-channel dual-axis global context information, but also learn complementary properties through shared keyword queries between spatial-channel axes.

pair between spatial-channel axis attention, which can reduce computation and interactive information.

As shown in Fig. 3, given a layer normalized tensor X of shape $HW \times C$,where $H \times W = N$ denotes the spatial dimension and C is the number of channels. DAT block projects the X into Query (Q), Key (K), and Value (V) enriched with local context by 1×1 convolutions and 3×3 depth-wise convolutions. Among them, the former can aggregate pixel-level cross-channel local context, and the latter can aggregate channel-level cross-spatial local context. The generated covariance weights of Q and K are shared across spatial-channel axis attention modules in DAT block and V is different from each other for spatial-channel axis attention modules. The projection process can be represented as:

$$\begin{cases} Q_s^{N\times C} = XW_{share}^{Q}, K_s^{N\times C} = XW_{share}^{K}, V_s^{N\times C} = XW_s^{V} \\ Q_c^{N\times C} = XW_{share}^{Q}, K_c^{N\times C} = XW_{share}^{K}, V_c^{N\times C} = XW_c^{V} \end{cases} \tag{4}$$

where W_{share}^{Q}, W_{share}^{K} are the learnable parameters for shared queries, shared keys and W_s^V , W_c^V are the learnable parameters for spatial attention module and channel attention

module.

Spatial Self-Attention. For spatial dimension self-attention, we first introduc a spatial-reduction self-attention (SSA) layer to reduce the complexity from $O(N^2)$ to $O(Np)$, where p is the reduction ratio of the spatial self-attention layers and $p \ll N$. First, the K_s and V_s layers are reshaped from $N \times C$ into a lower-dimensiona sequence of size $p \times C$ before the attention operation, which are respectively represented as $\hat{K}_s$ and $\hat{V}_s$. Second, the spatial-channel shared query Q_s and the reshaped K_s are multiplied, followed by softmax to generate the self-information matrix in the global spatial dimension, which focus on spatial interaction:

$$A_s \in \mathbb{R}^{N \times p} = Softmax\left(\frac{Q_s \times \hat{K}_s^{\top}}{\sqrt{d_{head}}}\right), \tag{5}$$

Third, the reshaped V_s is multiplied with similarities map to generate the final spatial self-attention map. The proposed spatial self-attention can be expressed as:

$$X_s \in \mathbb{R}^{N \times C} = Softmax\left(\frac{Q_s \times \hat{K}_s^{\top}}{\sqrt{d_{head}}}\right) \cdot \hat{V}_s, \tag{6}$$

where Q_s, $\hat{K}_s$, $\hat{V}_s$ are shared queries, reshaped shared keys, and reshaped spatial value, respectively, and $\sqrt{d_{head}}$ is the dimension of each head.

Channel Self-Attention. Self-attention in the channel dimension can provide deeper global correlation information for pixel matching in locally repetitive texture scenes. Compared with the spatial self-attention module, the keys K_c and values V_c of the channel self-attention module can realize linear calculation without being reconstructed into a lower dimensional matrix. Same as the spatial self-attention module, the shared queries Q_c and shared keys K_c are adopted, the values V_c are calculated independently. Then, Q_c will be reshaped to $Q_c \in \mathbb{R}^{C \times N}$, Q_c and K_c are employed to generate the crossed-channels global context attention maps $A_c \in \mathbb{R}^{C \times C}$ by the dot product, which focus on channel interaction:

$$A_c \in \mathbb{R}^{C \times C} = Softmax\left(\frac{Q_c \times K_c}{\sqrt{d_{head}}}\right), \tag{7}$$

Then we use dot product for the channel self-attention map A_c and the original feature V_c to obtain the final output:

$$X_c \in \mathbb{R}^{N \times C} = V_c \cdot Softmax\left(\frac{Q_c \times K_c}{\sqrt{d_{head}}}\right), \tag{8}$$

Finally, the outputs of the two-dimensional self-attention module are added and fused through convolutional blocks and skip connections to obtain a feature representation that includes local-global and spatial channel aspects. For the above spatial and channel self-attention, we use the multi head self attention(MSA) to divide the dimension into multiple heads, and the number of heads from the first to fourth stages is set to [1,2,4,8] to ensure that each head dimension is the same.

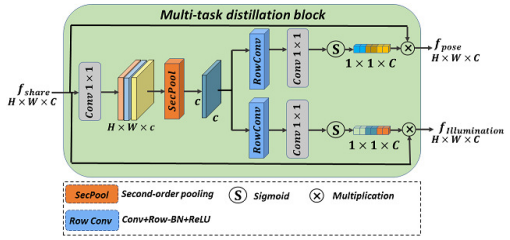


Fig. 4. Illustration of Multi-Task Distillation(MTD) block, which can distill the intermediate features of different tasks to suppress the negative transfer issue.

(a) Input image (b) Contrast (c) Brightness

Fig. 5. Visualisations of estimated light changes.

C. Illumination Compensation PoseNet

In order to accurately estimate the contrast(scale) change C_t and brightness(shift) change B_t between consecutive frames(from I_t to I_{t+1}) through a learning-based method, we propose a multi-task joint learning PoseNet based on shared encoder and dual decoder, which is inspired by the classic Phong illumination theory in computer graphics. Similar to [2], [3], we chose the ResNet18 lightweight network as the shared encoder for the PoseNet, which allows the adjacent frames $[I_t, I_{t+1}]$ to feed into the network. The decoder part consists of the pose decoder which estimates the relative pose of adjacent frames and the illumination decoder which estimates the brightness change of adjacent frames.

In addition, to address the issue of performance degradation in individual tasks due to shared intermediate features in multi-task learning, we propose a Multi-Task Distillation(MTD) block at the bottleneck layer to extract the most relevant channel features for different tasks. Fig. 4 shows the structure of our MTD block, which takes shared encoder downsampling features $f_{share} \in \mathbb{R}^{H \times W \times C}$ as input and output high-order long-sequence dependent features for different tasks. Specifically, to reduce computational complexity, we first use a 1×1 convolution to reduce the input feature dimension size to $H \times W \times c$ $(c < C)$. Then, inspired by the recent success of higher-order statistics[32], we use global second-order pooling to learn $c \times c$ covariance representations along the channel dimensions, where i^{th} row indicates statistical dependency of channel i with all channels. Next, for the shared covariance matrix, we use two parallel row-level convolutional normalization block and convolutional block to generate statistical weight vectors for different tasks, where the activation function of the former uses ReLU and the latter uses Sigmoid. Finally, perform dot product on statistical weight vectors and input f_{share} to obtain the final specific bottleneck layer features of different tasks.

After the MTD obtains the high-order long-range depen-

dent features of adjacent frames, the pose decoder composed of four convolutional layers estimates the relative pose with 6-DOF vector of adjacent frames. And the illumination decoder composed of four convolutional layers with skip connections estimates the per-pixel illumination change between two consecutive frames. As shown in Fig. 5, the illumination change image is a dual-channel image which is stacked by a contrast image C_t and a brightness image B_t.

D. Loss Function

Photometric Loss Similar to [1], [27], we use SSIM[28] and L1 as our photometric re-projection error to optimize the depth and pose networks, which is defined as:

$$p(I_t,\tilde{I}_t) = \frac{\alpha}{2}(1 - SSIM(I_t,\tilde{I}_t)) + (1-\alpha)\|I_t - \tilde{I}_t\|_1, \quad (9)$$

where $\alpha = 0.85$. In order to better handle occlusion, we follow [2] to minimize photometric error for each pixel in all source images. Our final photometric error can be denoted as:

$$L_p = \min_t p(I_t,\tilde{I}_t), \quad (10)$$

Smoothness Loss Moreover, we follow previous works[1] by applying edge-aware smoothness error to regularize the disparities in texture-less gradient regions:

$$L_s = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|}, \quad (11)$$

where ∂_x and ∂_y are the image gradient symbols along the horizontal and vertical directions, respectively.

Total Loss The total loss is the weighted sum of photometric loss and smoothing loss at each scale $k \in \{1, \frac{1}{2}, \frac{1}{4}\}$:

$$L_{total} = \frac{1}{3}\sum_k (L_p + \lambda L_s), \quad (12)$$

where k is the different scale output by the depth decoder, and λ represents the weight of the smoothing loss term.

IV. Experiment

Since this article focuses on unsupervised monocular depth estimation in highly complex scenes, we chose the publicly available Oxford RobotCar[30] dataset as our training and testing sets. Oxford RobotCar[30] dataset captures a variety of road conditions combining weather, traffic, and pedestrians, and includes more than 20 million images captured by six on-board cameras, as well as terrain data collected from laser ranging, GPS, and inertial navigation. We first converted the image data from the original record to an RGB style with a resolution of 1280×960 using an official toolbox, and then cropped the front hood in batches to adjust the resolution to 512×256. For night scenes, we extracted training and testing data from sequences 2014-12-10-18-10-50 and 2015-02-03-19-43-11 to form the RobotCar-Night dataset. For the rainy night scene, we extracted training and testing data from sequences 2014-11-21-16-07-03 and 2014-12-17-18-18-43, forming the RobotCar-Rainy Night dataset. For the winter snow scene, we only extracted training and testing data from a single sequence, 2015-02-03-08-45-10, to form the RobotCar-Snowy Winter dataset.

A. Implementation Details

Our models are implemented in PyTorch trained on i9-13900K CPU and NVIDIA RTX4090 GPU in mini-batches of 8. We adopt Adam optimizer with the hyper-parameter $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. We set the initial learning rate as $1e-4$, and the learning rate is decreased by half for every $1/5$ of total iterations during training. The experimental settings for the above parameters are the same in different complex scenarios, including night-time, rainy night, and winter snow scenarios. According to our experiments, 120 epochs are enough for convergence on RobotCar-Night and RobotCar-Rainy Night datasets, and 80 epochs are enough for convergence on RobotCar-Snowy Winter dataset. The weights of the smoothness constraint loss components are set to $\lambda = 0.1$. In the encoding layers of the four down-sampling stages, the number of DAT blocks is set to [4, 6, 6, 8], and the number of channels is set to [48, 96, 192, 384].

Evaluation metrics. For quantitative evaluation, we have selected several universal evaluation indicators to evaluate the performance of our model:

$$\begin{cases} \text{Abs Rel} = \frac{1}{|D|}\sum_{d\in D}\frac{|d-d^*|}{d^*} \\ \text{Sq Rel} = \frac{1}{|D|}\sum_{d\in D}\frac{|d-d^*|^2}{d^*} \\ \text{RMSE} = \sqrt{\frac{1}{|D|}\sum_{d\in D}||d-d^*||^2} \\ \text{RMSE log} = \sqrt{\frac{1}{|D|}\sum_{d\in D}||\log d - \log d^*||^2} \\ \delta_t = \frac{1}{|D|}\left|\left\{d \in D \max(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25^t\right\}\right| \end{cases} \quad (13)$$

where d and d^* represent ground truth and predicted values of the depth maps, respectively. And D represents a set of valid ground truth depth values in an image, $|.|$ returns the number of elements in the input set.

B. Comparison with SOTA Methods

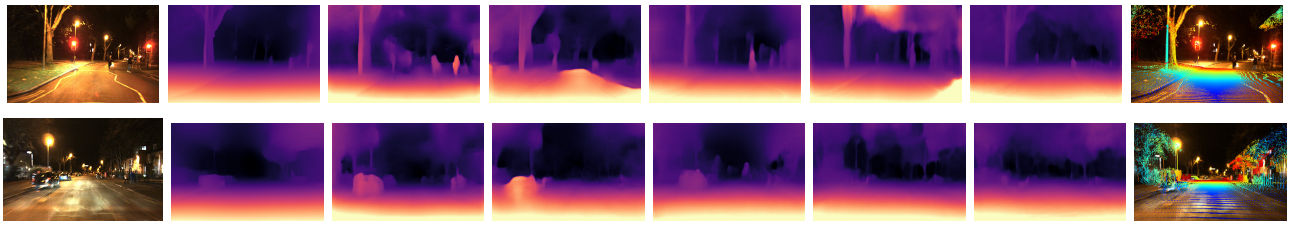
In this section, we compare the current best-performing and most well-known monocular depth estimation methods, including the most well-known Monodepth2[2], the recently proposed SwinDepth[19] and Lite-Mono[20] based on Transformer backbone, as well as the ADDS-DepthNet[25], RNW[26], and Steps[31] that are suitable for night-time scenes.

Results on RobotCar-Night. Fig. 6 and the first row of Table I show the qualitative and quantitative experimental comparison results of our method in night scenes. As shown in Fig. 6, the changing point sources in night scenes not only lead to inconsistent brightness in different regions, but also easily lead to luminance changes in adjacent frames, violating the assumption of consistent brightness in adjacent frames in monocular depth estimation. Compared to other methods, because we introduce luminance compensation for adjacent frames, we can compensate for the negative effects caused by point sources, resulting in better qualitative visualization results. In terms of quantitative results, as shown in Table I, our method achieves the best performance in all metrics, and for the representative metrics Abs Rel and

TABLE I

QUANTITATIVE COMPARISONS WITH DIFFERENT DEPTH METHODS ON CHALLENGING ROBOTCAR DATASETS. THE BEST MEASUREMENTS ARE IN BOLD. FOR ABS REL, SQ REL, RMSE, $RMSE_{log}$, THE LOWER IS BETTER. FOR $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, THE HIGHER IS BETTER.

Method	Supervision	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
RobotCar-Night								
SwinDepth[19]ICRA'23	M	0.588	23.743	13.259	0.563	0.532	0.827	0.910
Monodepth2[2]ICCV'19	MS	0.537	19.865	12.178	0.516	0.558	0.846	0.916
Lite-Mono[20]CVPR'23	M	0.450	13.859	10.656	0.443	0.583	0.861	0.925
ADDS-DepthNet[25]ICCV'21	M	0.264	2.635	8.968	0.304	0.563	0.835	0.949
RNW[26]ICCV'21	M	0.184	1.868	7.142	0.240	0.731	0.911	0.963
Steps[31]ICRA'23	M	0.177	1.695	6.935	0.253	0.736	0.917	0.968
Ours	M	**0.129**	**0.976**	**5.482**	**0.195**	**0.844**	**0.938**	**0.980**
RobotCar-Rainy Night								
SwinDepth[19]ICRA'23	M	0.581	37.366	16.628	0.452	0.575	0.853	0.914
Monodepth2[2]ICCV'19	MS	0.569	42.826	14.758	0.461	0.592	0.841	0.918
Lite-Mono[20]CVPR'23	M	0.290	7.483	9.264	0.383	0.675	0.865	0.937
ADDS-DepthNet[25]ICCV'21	M	0.346	20.749	12.635	0.432	0.628	0.857	0.926
RNW[26]ICCV'21	M	0.223	12.754	8.690	0.351	0.672	0.887	0.943
Steps[31]ICRA'23	M	0.194	1.975	7.328	0.287	0.716	0.902	0.953
Ours	M	**0.140**	**1.083**	**5.538**	**0.216**	**0.830**	**0.935**	**0.974**
RobotCar-Snowy Winter								
SwinDepth[19]ICRA'23	M	0.231	2.658	6.540	0.312	0.714	0.877	0.936
Monodepth2[2]ICCV'19	MS	0.214	2.373	6.247	0.295	0.721	0.884	0.932
Lite-Mono[20]CVPR'23	M	0.158	1.383	5.382	0.225	0.826	0.928	0.973
ADDS-DepthNet[25]ICCV'21	M	0.176	1.763	5.539	0.262	0.791	0.914	0.955
RNW[26]ICCV'21	M	0.195	1.934	5.898	0.273	0.745	0.906	0.947
Steps[31]ICRA'23	M	0.327	5.264	7.496	0.362	0.683	0.869	0.928
Ours	M	**0.133**	**1.042**	**5.174**	**0.195**	**0.857**	**0.942**	**0.986**



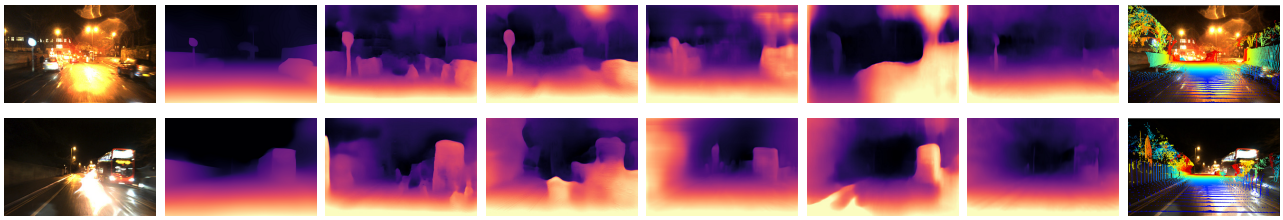
(a) Input (b) Ours (c) Steps (d) RNW (e) ADDS (f) Mono2 (g) Lite (h) Ground Truth

Fig. 6. Qualitative comparison of the state-of-the-art depth estimation methods on RobotCar-Night dataset.

$\delta < 1.25$, we are 29.3% and 15.3% lower than the second-place Steps method, respectively.

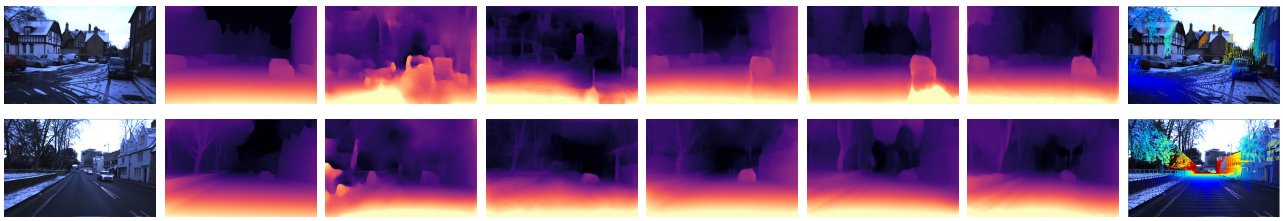
Results on RobotCar-Rainy Night. Fig. 7 and the second row of Table I show the qualitative and quantitative experimental results of the rainy night scenario compared to the optimal comparison method. Compared to night scenes, rainy night scenes not only have problems with luminance inconsistencies, but also have problems with reflections of rain on roads and cameras, resulting in glow/glare, flare, and uneven distribution of light. As shown in Fig. 7, most comparison methods cannot obtain accurate depth estimation in the glare/flare areas caused by rain, resulting in large black holes and distortions. In contrast, our model is not affected by rainfall and can obtain clearer and more accurate results in corresponding areas. From the quantitative results in Table I, we can see that our model achieves the best score in all metrics for rainy night scenes compared to other methods, with a 27.8% and 15.9% lower Abs Rel and $\delta < 1.25$ respectively compared to the second-place Steps[31].

Results on RobotCar-Snowy Winter. In the winter snow environment, the unsupervised framework suffers from repeated texture scenes, which affects image reconstruction based on the photometric consistency assumption. In addition, the exposure of the camera is affected by the snow environment, resulting in insufficient exposure, which limits the performance of the unsupervised framework. As shown in Fig. 8, due to the influence of repeated textures caused by snow, Steps[31], a previously well-performing night-time scene depth estimation model, no longer has an advantage in the snow environment. In contrast, Lite-Mono[20], which benefits from the Transformer architecture, can represent and reconstruct repetitive textures from a global perspective due to the addition of global attention. Therefore, it performs better in the snow environment. However, because Lite-Mono only performs long-sequence reconstruction in the spatial perspective, our model adds a spatial global attention mechanism on top of it to further improve the performance of monocular depth estimation in the snow environment. The quantitative results in Table I also prove this point, as our model achieves the best scores in all indicators.



(a) Input (b) Ours (c) Steps (d) RNW (e) ADDS (f) Mono2 (g) Lite (h) Ground Truth

Fig. 7. Qualitative comparison of the state-of-the-art depth estimation methods on RobotCar-Rainy Night dataset.



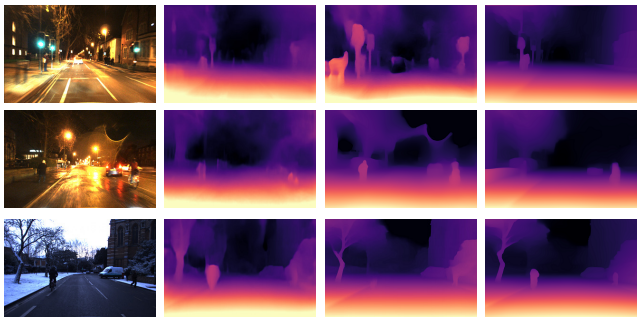
(a) Input (b) Ours (c) Steps (d) RNW (e) ADDS (f) Mono2 (g) Lite (h) Ground Truth

Fig. 8. Qualitative comparison of the state-of-the-art depth estimation methods on RobotCar-Snowy Winter dataset.

TABLE II

QUANTITATIVE RESULTS FOR ABLATION STUDY ON CHALLENGING ROBOTCAR DATASETS. THE BEST MEASUREMENTS ARE IN BOLD. FOR ABS REL, SQ REL, RMSE, $RMSE_{log}$, THE LOWER IS BETTER. FOR $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, THE HIGHER IS BETTER.

Method	Supervision	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
RobotCar-Night								
Baseline	M	0.474	16.823	11.715	0.491	0.546	0.852	0.917
Baseline+DAT	M	0.392	7.763	9.528	0.415	0.627	0.873	0.944
Baseline+DAT+ICP	M	**0.129**	**0.976**	**5.482**	**0.195**	**0.844**	**0.938**	**0.980**
RobotCar-Rainy Night								
Baseline	M	0.247	8.634	9.581	0.362	0.695	0.868	0.932
Baseline+DAT	M	0.211	5.638	7.709	0.323	0.712	0.894	0.945
Baseline+DAT+ICP	M	**0.140**	**1.083**	**5.538**	**0.216**	**0.830**	**0.935**	**0.974**
RobotCar-Snowy Winter								
Baseline	M	0.166	1.574	5.825	0.257	0.788	0.913	0.952
Baseline+DAT	M	0.142	1.281	5.352	0.216	0.834	0.935·	0.979
Baseline+DAT+ICP	M	**0.133**	**1.042**	**5.174**	**0.195**	**0.857**	**0.942**	**0.986**



(a) Input image (b) base (c) b+DAT (d) c+ICP

Fig. 9. Ablation study of qualitative depth estimation on challenging RobotCar datasets, from top to bottom, they are night-time, rainy night, and winter snow scene.

C. Ablation Study

To better understand each component module, we evaluated three different variants of our model on the RobotCar dataset with three highly complex scenarios for ablation studies. We use a Tranformer encoder backbone with only a traditional spatial axis as the baseline model. Based on the baseline model, we add each functional module in turn to evaluate its contribution to the overall performance of the proposed method. As shown in the qualitative results of Fig. 9, for the winter snow scene, compared to the baseline model, the addition of the DAT module can solve the problem of inaccurate reconstruction of locally repeated textures caused by snow through global features in both spatial and channel dimensions. For the challenge of severe changes in adjacent frame photometry and glow/glare caused by point sources and rain, which violates the assumption of adjacent frame photometric consistency in unsupervised depth estimation, the qualitative results in Fig. 9 visually demonstrate that adding the ICP module can compensate for the performance degradation caused by this challenge. The quantitative results in Table II reflect the effectiveness of the proposed DAT and ICP modules for different complex scenarios. For example, for the representative metric Abs Rel, the addition of the ICP module reduces it by 67% and 33.6% for night-time and rainy night scenarios, respectively, while the addition of the DAT module reduces it by 14.5% for snowy environments.

V. Conclusion

To conclude, we present a unique self-supervised monocular depth estimation framework that can be used in challenging environments(night-time, rainy night and snow winter). We achieve state-of-the-art results when compared to previous models by utilising a novel Illumination Compensation PoseNet that compensates for lighting changes in adjacent frames by estimating per-pixel transformations. In addition, a Dual-Axis Transformer (DAT) block as the backbone network of the depth encoder, which infers the depth of local repeat-texture areas through spatial-channel dual-dimensional global context information and self-similarity of images. Future work will focus on the depth perception problem of small objects in complex environments.

References

[1] 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*, 2017, pp. 270–279.

[2] 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*, 2019, pp. 3828–3838.

[3] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan,"Hr-depth: High resolution self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2294–2301.

[4] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.

[5] Y. Almalioglu, M. R. U. Saputra, P. P. De Gusmao, A. Markham, and N. Trigoni,"Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5474–5480.

[6] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9250–9256.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[9] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[10] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[11] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[12] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1592–1599.

[13] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66–75.

[14] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 740–756.

[15] 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*, 2017, pp. 1851–1858.

[16] 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*, vol. 32, no. 4, pp. 1751–1766, 2021.

[17] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1601–1610.

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[19] D. Shim and H. J. Kim, "Swindepth: Unsupervised depth estimation using monocular sequences via swin transformer and densely cascaded network," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4983-4990.

[20] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.

[21] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 668–678.

[22] Y. Lu and G. Lu, "An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3833–3843.

[23] A. Sharma, L.-F. Cheong, L. Heng, and R. T. Tan, "Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 23–31.

[24] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 443–459.

[25] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 737–12 746.

[26] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 055–16 064.

[27] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[30] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[31] Y. Zheng, C. Zhong, P. Li, H.-a. Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang *et al.*, "Steps: Joint self-supervised nighttime image enhancement and depth estimation," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4916–4923.

[32] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 3024–3033.