

Always Clear Depth: Robust Monocular Depth Estimation Under Adverse Weather

Kui Jiang¹, Jing Cao¹, Zhaocheng Yu¹, Junjun Jiang^{1*} and Jingchun Zhou²

¹Harbin Institute of Technology

²Dalian Maritime University

jiangkui@hit.edu.com, {caojing,yuzhaocheng}@stu.hit.edu.cn, jiangjunjun@hit.edu.cn, zhoujingchun@dmlu.edu.cn

Abstract

Monocular depth estimation is critical for applications such as autonomous driving and scene reconstruction. While existing methods perform well under normal scenarios, their performance declines in adverse weather, due to challenging domain shifts and difficulties in extracting scene information. To address this issue, we present a robust monocular depth estimation method called **ACDepth** from the perspective of high-quality training data generation and domain adaptation. Specifically, we introduce a one-step diffusion model for generating samples that simulate adverse weather conditions, constructing a multi-tuple degradation dataset during training. To ensure the quality of the generated degradation samples, we employ LoRA adapters to fine-tune the generation weights of diffusion model. Additionally, we integrate circular consistency loss and adversarial training to guarantee the fidelity and naturalness of the scene contents. Furthermore, we elaborate on a multi-granularity knowledge distillation strategy (MKD) that encourages the student network to absorb knowledge from both the teacher model and pretrained Depth Anything V2. This strategy guides the student model in learning degradation-agnostic scene information from various degradation inputs. In particular, we introduce an ordinal guidance distillation mechanism (OGD) that encourages the network to focus on uncertain regions through differential ranking, leading to a more precise depth estimation. Experimental results demonstrate that our ACDepth surpasses md4all-DD by 2.50% for night scene and 2.61% for rainy scene on the nuScenes dataset in terms of the absRel metric.

1 Introduction

Monocular depth estimation (MDE) is a fundamental task in computer vision that aims to predict the depth from a single image. It has wide-ranging applications in autonomous

driving [Schön *et al.*, 2021; Xue *et al.*, 2020], robot navigation [Häne *et al.*, 2011], and 3D reconstruction [Yu and Gallup, 2014; Yin *et al.*, 2022]. MDE methods can be roughly divided into supervised methods and self-supervised methods [Arampatzakis *et al.*, 2023]. The former [Bhat *et al.*, 2023; Ranftl *et al.*, 2020; Ranftl *et al.*, 2021] directly learns depth mapping from the RGB input with paired samples (depth maps or 3D sensors like LiDAR). While achieving impressive performance, these methods heavily rely on high-quality ground truth data, which is time-consuming and labor-intensive for collection. By contrast, the latter [Godard *et al.*, 2017; Godard *et al.*, 2019; Zhou *et al.*, 2017] uses only image sequences captured by a single camera and the corresponding camera parameters, and employs two assumptions (photometric constancy and rigid motion) [Godard *et al.*, 2019] to produce the supervised signal. Specifically, these methods assume that scene information remains largely unchanged as the viewpoint shifts, relying on the geometric consistency between consecutive frames to provide depth supervision. While existing self-supervised approaches [Zhao *et al.*, 2022; Zhang *et al.*, 2023; Lyu *et al.*, 2021] have shown significant success in outdoor scenes, their performance deteriorates under adverse conditions, such as low-light and rain-haze conditions. In these situations, insufficient lighting and significant motion perturbations (reflections of raindrops and streetlights) violate the assumptions mentioned above, leading to an unreliable estimation.

Recent studies [Gasperini *et al.*, 2023; Saunders *et al.*, 2023; Tosi *et al.*, 2025] have explored ways to enhance the robustness of models under adverse conditions. Some technologies elaborate modules to promote robustness in specific scenes, such as nighttime and rain-haze conditions [Shi *et al.*, 2023; Zheng *et al.*, 2023; Yang *et al.*, 2024a]. However, due to the limitation for scene-specific representation, these methods struggle to generalize to complex and diverse environments, frequently encountered in real-world applications. To promote generalization, some efforts harmonize the merits of distillation learning, contrastive learning, and data augmentation strategies to create more generalized models that improve model performance across varied scenes [Wang *et al.*, 2024b; Wang *et al.*, 2024a; Saunders *et al.*, 2023]. For example, md4all [Gasperini *et al.*, 2023] generates the [normal, degraded] sample pairs with a GAN-based model and uses distillation learning to extract supervised signals from clear

*Corresponding Author

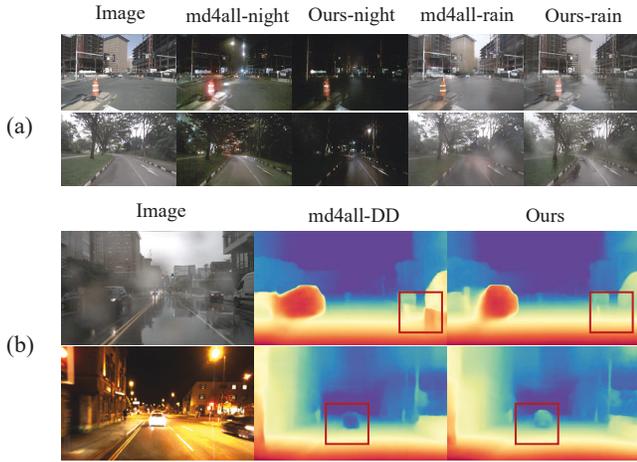


Figure 1: **(a) Comparison of training data in challenging scenes:** Compared to the data generation method used in md4all [Gasperini *et al.*, 2023], the samples generated by our approach are more realistic, providing a better simulation of challenging real-world conditions. **(b) Estimation results in challenging scenes:** Our method consistently produces more accurate results than the existing md4all [Gasperini *et al.*, 2023] method, particularly in handling complex issues such as ground water reflections and nighttime object recognition.

depth maps to train the student model. However, the degraded characteristics in the translated image are unrealistic and unnatural (shown in Figure. 1(a)), which significantly affects the generalization from normal scenarios to real adverse weather. In addition, existing methods are often vulnerable to disturbances in complex scenes due to slack constraints between the teacher and student models, leading to incomplete knowledge transfer. Consequently, as shown in Figure. 1(b), md4all suffers from nonnegligible performance decline when there are significant domain differences between scenes, making it far from generalizing to various degraded scenarios.

Overall, two critical issues are imperative to train a generalizable and reliable depth estimation model: (1) the lack of a high-quality multi-tuple dataset covering diverse degradation types; (2) slack constraints between the teacher and student model, resulting in incomplete knowledge transfer.

To mitigate the problem of data scarcity, the author in [Tosi *et al.*, 2025] utilize depth maps and text descriptions as control conditions to generate [normal, degraded] sample pairs with T2I-Adapter [Mou *et al.*, 2024]. This provides an alternative solution with diffusion models for multi-tuple dataset generation [Zhu *et al.*, 2017; Parmar *et al.*, 2024; Sauer *et al.*, 2025]. However, the translated image shows obvious inconsistencies and less authenticity of scene contents, which is unacceptable for training generalizable depth estimation models.

To address these issues, we propose to optimize the data generation and domain adaptation learning, and construct the degradation-agnostic robust monocular depth estimation method (ACDepth). Specifically, we fully explore the potential generation capability of diffusion model and employ LoRA adapters to encourage the network to generalize to diverse scene generation. Meanwhile, we integrate circular

consistency loss and adversarial training to guarantee the naturalness and consistency of translated images. To achieve full transfer and alignment of capabilities, we propose the multi-granularity knowledge distillation strategy (MKD), which borrows priors from the teacher model to provide comprehensive supervision and guidance to the student model. Besides the commonly used feature and result distillation learning, we pioneer the ordinal guidance distillation mechanism (OGD). In summary, the contributions are as follows:

- We propose a practical multi-tuple degradation dataset generation scheme, and develop a novel robust monocular depth estimation framework, termed as ACDepth for high-quality depth estimation under adverse weather.
- We propose the multi-granularity knowledge distillation strategy (MKD) to achieve the complete transfer and alignment of capabilities from the teacher model to student model. In addition, we introduce the ordinal guidance distillation mechanism (OGD) to heartens the network to focus on uncertain regions through differential ranking.
- Extensive experiments demonstrate the effectiveness of ACDepth, surpassing md4all-DD by 2.50% for night scene and 2.61% for rainy scene on the nuScenes dataset in terms of absRel metric.

2 Related Work

2.1 Monocular Depth Estimation

Before the advent of deep neural networks, traditional depth estimation methods primarily rely on handcrafted priors to explore the limited physical and geometric properties. Deep learning methods [Sun *et al.*, 2012; Liu *et al.*, 2015b; Liu *et al.*, 2015a] have emerged as a preferable alternative due to their ability to learn generalizable priors from large-scale data, such as depth maps from LiDAR or RGB-D cameras. However, due to the high cost to obtain the high-quality annotation, self-supervised learning technologies, deriving depth information from stereo pairs [Garg *et al.*, 2016; Godard *et al.*, 2017] or video sequences [Zhou *et al.*, 2017; Godard *et al.*, 2019; Bian *et al.*, 2019] have drawn growing interest. Unfortunately, a significant portion of the effort is focused on the normal scene. The depth estimation under adverse weather, such as low-light and rain-haze conditions, is barely explored, which is the common scenario in autonomous driving. Some studies [Zheng *et al.*, 2023; Wang *et al.*, 2021; Guo *et al.*, 2020] divide depth estimation in adverse scenes into denoising and estimation, but these methods show poor generalization under unknown degradations. Further, researchers achieve more robust depth estimation through data generation and knowledge distillation [Zhu *et al.*, 2023; Gasperini *et al.*, 2023], and integrate multi-level contrastive learning with diffusion model for robust feature representation [Wang *et al.*, 2024b]. However, the data quality and completeness of constraints in these methods significantly affect the generalization from normal scenarios to real adverse weather.

2.2 Distillation Learning

Early distillation learning methods primarily focus on model compression and acceleration. A classic example of distillation learning is to guide the training of student models by softening the output probability distribution of the teacher model [Hinton, 2015]. Subsequently, several studies have successfully applied distillation learning to monocular depth estimation. For example, Song et al. [Song and Yoon, 2022] selectively distill stereo knowledge for monocular depth estimation, using learned binary masks to pick the best disparity or pixel-wise depth map. More recently, md4all [Gasperini *et al.*, 2023] trains the teacher model on clear samples by self-supervised learning, and transfers the ability or priors to the student model under adverse weather. However, during the distillation process, the student model struggles to fully reproduce the ability of the teacher model, relying solely on depth-derived pseudo-labels or specific priors. To address these challenges, we propose a multi-granularity knowledge distillation strategy that enhances the knowledge transfer process by borrowing priors from multiple teachers to provide comprehensive supervision and guidance.

3 Method

3.1 Preliminary

MDE aims to estimate the depth map D_t from a single RGB image I_t . Restricted by high-quality paired samples in real-world scenarios, especially for adverse weather, the self-supervised learning strategy [Zhu *et al.*, 2023; Wang *et al.*, 2024b] provides an alternative solution to obtain supervised signals of the target image I_t from adjacent frames $I_{t'} \in [I_{t-1}, I_{t+1}]$. Specifically, the pose network and depth network are jointly optimized. The former estimates the relative pose of camera motion $T_{t \rightarrow t'}$, from the target image to the adjacent frame ($I_{t'}$). Combining with the intrinsic parameters K of camera, it allows the network to synthesize the reconstruction image ($I_{t' \rightarrow t}$) of I_t using the adjacent frame $I_{t'}$, depicted as:

$$I_{t' \rightarrow t} = I_{t'} \langle \text{proj}(D_t, T_{t \rightarrow t'}, K) \rangle, \quad (1)$$

where the $\langle \rangle$ denotes the pixel sampling operator, and we constrain the depth by calculating the photometric reconstruction loss between I_t and $I_{t' \rightarrow t}$, formulated by:

$$L_p = \min_{t'} pe(I_t, I_{t' \rightarrow t}), \quad (2)$$

$$pe(I_a, I_b) = \frac{\theta}{2} (1 - SSIM(I_a, I_b)) + (1 - \theta) |I_a - I_b|, \quad (3)$$

where SSIM is the structural similarity index measure. Compared with the pixel-wised L2 loss, it can better reflect the structural similarity between images. Additionally, edge-aware smoothness loss is also used to constrain the continuity of the depth:

$$L_e(D) = |\partial_x D'| e^{\partial_x I} + |\partial_y D'| e^{\partial_y I}, \quad (4)$$

where D' refers to the normalized inverse depth of D . ∂_x and ∂_y represent the horizontal and vertical gradients. Similar to [Gasperini *et al.*, 2023], the above theoretical foundations of self-supervised learning are used to train the teacher model in this work.

3.2 Overall Architecture

Our ultimate goal is to achieve the robust MDE under adverse weather. Thus, we elaborate an ACDepth approach, as shown in Figure. 2, which involves the data generation and robust model training. In the first part, we utilize the LoRA adapters to fine-tune the pretrained diffusion model, where the circular consistency loss and adversarial training are used to guarantee the naturalness and consistency of the translated image. In this way, the trained generator can produce a multi-tuple degradation dataset under different weather conditions. And then, ACDepth takes the multi-tuple degradation dataset as input, and proposes the multi-granularity knowledge distillation strategy (MKD) to borrow priors from the teacher model to optimize the student network. The process is depicted as:

$$L = L_d + \lambda_1 L_r + \lambda_2 L_c, \quad (5)$$

where L_d denotes the distillation loss between the teacher and student models, L_r refers to the ordinal guidance distillation between the Depth Anything V2 model and our ACDepth, and L_c is the feature consistency loss between the teacher encoder and the student encoder. λ_1 and λ_2 are the weight parameters to balance the loss components. Detailed explanation of losses (L_d , L_r and L_c) is provided below.

3.3 Data Generation

Given a normal input (e_i), which is captured under good lighting and visibility conditions, the previous studies employ GAN-based or diffusion-based models to generate the degraded images (h_i^c). c represents various weather scenes, such as rain, night, and fog. However, besides requiring a large number of real multi-tuple pairs for training, the translated images generated by these technologies [Gasperini *et al.*, 2023; Tosi *et al.*, 2025] exhibit obvious differences and unnaturalness from the real samples, leading to poor generalization of depth estimation models across different scenes.

Inspired by recent image translation technologies [Parmar *et al.*, 2024], we explore the content generation capability of stable diffusion, while employing adversarial learning and cycle consistency loss to train LoRA [Hu *et al.*, 2021] adapters to promote the naturalness and consistency of translated images. Specifically, we take e_i as input, along with the corresponding text prompt P_c to learn the specific LoRA adapters for each scene transformation, which can complete the conversion from the source domain to the target domain. The aforementioned process is depicted as:

$$h_i^c = F_c(SDT(e_i, P_c)), \quad (6)$$

where $SDT(\cdot)$ refers to the Stable Diffusion Turbo model, and $F_c(\cdot)$ is the translator that transforms the normal sample into the corresponding adverse scene with the condition c and text prompt P_c . Detailed experiments related to data generation can be found in the appendix.

3.4 Robust Model Training

To achieve full transfer and alignment of capabilities between the teacher model and student model, we propose the multi-granularity knowledge distillation strategy (MKD) to achieve the robust model training of the student model, detailed as follows.

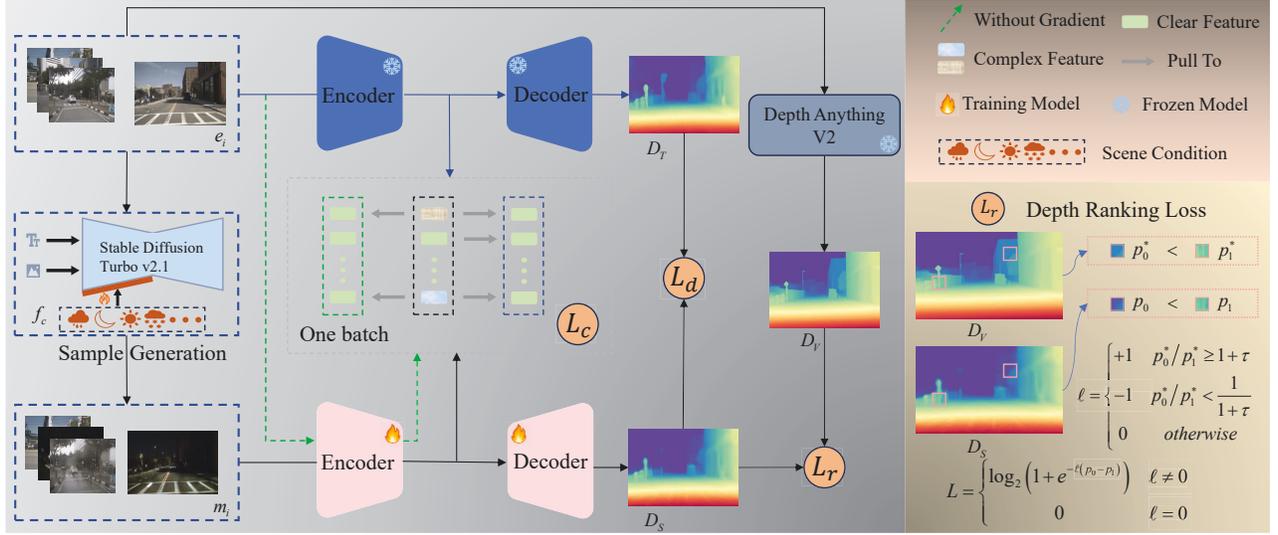


Figure 2: **Overview of our ACDepth for robust monocular depth estimation.** The teacher model is trained on simple samples using self-supervised learning, and the student model is trained on a mixed dataset of simple and complex samples using distillation learning. To provide the student model with supervisory signals beyond those from the teacher model, we designed a depth ranking loss L_r leveraging ordinal information from the Depth Anything V2 model. To improve the student model’s generalization across diverse scenarios, we incorporated a feature constraint loss L_c .

Distillation Learning. Similar to the existing technology [Gasparini *et al.*, 2023], the commonly used multi-scale feature distillation loss between the teacher model ($F_T(\cdot)$) and student model ($F_S(\cdot)$) is employed to facilitate the perception of the student model to the adverse weather. For the given normal input e_i and the produced degraded sample (h_i^c), the distillation loss function is defined as:

$$L_d = \frac{1}{S} \sum_{s=1}^S \frac{1}{N_s} \sum_{j=1}^{N_s} \frac{|F_T(e_i)_{j_s} - F_S(m_i)_{j_s}|}{F_S(m_i)_{j_s}}, \quad (7)$$

where S denotes the number of different scales and m_i represents a sample randomly selected from the mixed training set.

Ordinal Guidance Distillation. 1) **Uncertain Region Definition:** We use the output from Depth Anything V2 model as the supervisory signal because it ensures accurate depth prediction while maintaining acceptable inference speed. D_T and D_S represent the inverse depth of the teacher model and student model, respectively, and D_v represents the output depth of Depth Anything V2. We first identify the regions with significant discrepancies between D_T and D_S for refinement:

$$\bar{D} = |D_T - D_S|, \quad (8)$$

where \bar{D} reflects the difference between the output of the teacher model and the student model. We normalize the depth difference \bar{D} to obtain \hat{D} , which takes values between 0 and 1. We define U as the region of focus for the loss, and the region of focus is determined by the following equation:

$$U = \begin{cases} 1 & d_i > \gamma \\ 0 & d_i \leq \gamma, \end{cases} \quad (9)$$

where γ is the threshold for dividing the key regions, set to 95% of \hat{D} . These key regions correspond to areas with impaired perception in challenging scenes, as shown in Fig 3(e).

2) **Depth Ordinal Strategy:** We sample a pixel from $U * D_S$, with the corresponding depth value p_0 , and another pixel from $(\sim U) * D_S$, with the corresponding depth value p_1 . The ranking loss [Xian *et al.*, 2020; Sun *et al.*, 2023] for the pair $[p_0, p_1]$ is computed as follows:

$$\psi(p_0, p_1) = \begin{cases} \log_2(1 + e^{-\ell(p_0 - p_1)}) & \ell \neq 0 \\ \ell(p_0 - p_1)^2 & \ell = 0, \end{cases} \quad (10)$$

The ordinal label ℓ is calculated as follows:

$$\ell = \begin{cases} +1 & \frac{p_0^*}{p_1^*} \geq 1 + \tau \\ -1 & \frac{p_0^*}{p_1^*} < \frac{1}{1 + \tau} \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

where τ is a hyperparameter used to control the selection of pixel pairs for the sorting. p_0^* is sampled from the region $U * D_v$, and p_1^* is sampled from the complementary region $(\sim U) * D_v$. These two pixels, $[p_0^*, p_1^*]$, form an ordinal pair. The ranking loss aims to enhance the student model’s performance by focusing on poorly predicted regions. However, we found that the constraints imposed by the distillation loss sometimes conflicted with the ranking loss. To mitigate this issue, after identifying each ordinal pair, we compute the average depth value of the surrounding 5×5 pixel region for each sample point and use this average value to calculate the ranking loss. Considering the increase in computational overhead, we control the number of selected points in each iteration by setting $\tau = 0.15$ to select fewer sample points. Similarly, we set the proportion of sample pairs selected from U to 0.05. Finally, we randomly sample ordinal pairs from the

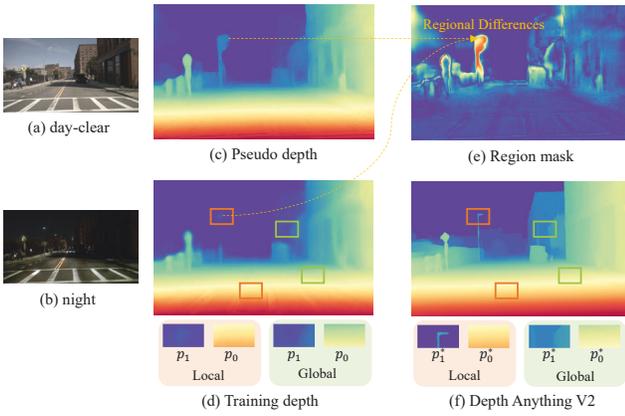


Figure 3: Ordinal Pair Sampling: During training, we use two strategies to efficiently sample ordinal pairs. First, we employ the teacher and student models to respectively predict the depth maps (c and d) from the normal (a) and degraded (b) input, and then compute the pixel-wise errors (e) among them. Based on (e), we sample the local (uncertain regions with large errors) and global (random sampling regions) patches from the depth maps (d) and (f) to compute the total ordinal pair loss.

global depth map by a ratio of 0.01, and the ranking loss calculated from these sampled points is used to refine the global depth information. The sampling process is illustrated in Figure. 3. Thus, our final loss becomes:

$$L_r = \frac{1}{|Z_g|} \sum_{p_0, p_1 \in Z_g} \psi(p_0, p_1) + \frac{1}{|Z_l|} \sum_{p_0, p_1 \in Z_l} \psi(p_0, p_1), \quad (12)$$

where Z_g and Z_l represent the global and local ordinal sampling sample sets, respectively.

Feature Consistency Constraint. Since a robust model can provide the precise perception on both the normal and degraded scenarios, we propose the feature consistency constraint to mitigate the impact of image degradation, thus generalizing to various adverse weather. Specifically, we dynamically adjust the feature alignment strategy based on the input image e_i . If e_i is the degraded image h_i^c , we minimize the error between the feature maps $F_S(h)$ and $F_T(e)$, as well as the error between $F_S(h)$ and $F_S(e)$, where F_T and F_S represent the feature extraction by the teacher and student models, respectively. If e_i is a clear sample e , we minimize the error between the feature maps $F_T(e)$ and $F_S(e)$, respectively extracted by the teacher and student models. This form can be written as:

$$L_c = \begin{cases} f(F_S(h), F_T(e)) + f(F_S(h), F_S(e)) & \text{if } i = h \\ f(F_S(e), F_T(e)) & \text{if } i = e, \end{cases} \quad (13)$$

where f represents L1 loss for aligning features, and the horizontal line denotes that no gradient backpropagation is applied. The above process enables the student model to not only acquire semantic features from the teacher model but also progressively improve its feature extraction capabilities across both normal and challenging scenarios, ensuring a seamless transition from clear to complex environments.

4 Experiments

4.1 Datasets

In this study, the commonly used nuScenes [Caesar *et al.*, 2020] and RobotCar [Maddern *et al.*, 2017] datasets are used for training and comparison. NuScenes is a comprehensive dataset featuring diverse outdoor scenes, specifically tailored for autonomous driving research, where the multi-frame sequence is supplemented with detailed radar annotations. Following [Gasparini *et al.*, 2023], we adopt 15,129 generated samples (day-clear, day-rain, night) for training and 6,019 samples (including 4449 day-clear, 1088 rain, and 602 night) for testing. RobotCar is a large outdoor dataset collected in Oxford, UK. Following [Gasparini *et al.*, 2023], we adopt 16,563 generated samples (day, night) for training and 1,411 samples (including 702 day 709 night) for testing. During the data preprocessing stage, images in the nuScenes dataset are resized to 320×576 , while those in the RobotCar dataset are resized to 320×544 . We record metrics within a range of 0.1m to 50m for RobotCar and 0.1m to 80m for nuScenes.

4.2 Implementation Details

All experiments are conducted on the same ResNet18 architecture [He *et al.*, 2016]. We train the student model and teacher model on a single NVIDIA 3090 GPU with a batch size of 16, using the Adam optimizer. We set initial learning rate to $5e-4$, reducing it by a factor of 0.1 every 15 epoch. The student model are trained for 25 epochs. Following the experimental protocol of [Gasparini *et al.*, 2023], we maintain identical hyperparameter settings for self-supervised learning. Through experimental validation of different parameter combinations, the weights for the loss functions are set to $\lambda_1 = 0.01$, $\lambda_2 = 0.02$. Considering both the effectiveness of supervision signals and training efficiency, we select the small version of Depth Anything V2 [Yang *et al.*, 2024b] as our supervision prior for ordinal guidance distillation. Further details on the model training procedure and dataset translation process are provided in the appendix.

4.3 Comparison with SoTAs

In this section, we evaluate our model on the nuScenes and RobotCar datasets. Qualitative and quantitative comparisons are shown below.

Comparison on the nuScenes. In Table 1, we compare our method with existing depth estimation models, including the typical MDE technologies (Monodepth2 [Godard *et al.*, 2019], PackNet-SfM [Guizilini *et al.*, 2020]) and robust depth estimation approaches (RNW [Wang *et al.*, 2021], md4all [Gasparini *et al.*, 2023] and DMMDE [Tosi *et al.*, 2025]). As shown in Table 1, our proposed ACDepth shows significant competitiveness, particularly for the night condition, respectively reducing 2.50% and 4.49% in terms of absRel and RMSE metrics, when compared to robust depth estimation method md4all-DD. DMMDE [Tosi *et al.*, 2025] uses different depth models as prior information to guide image translation, and achieves improvements over md4all-DD. However, due to the scale-and-shift-invariant loss, the scaling relation of depth map is ignored and corrupted during alignment, degrading the final performance. For better convincing,

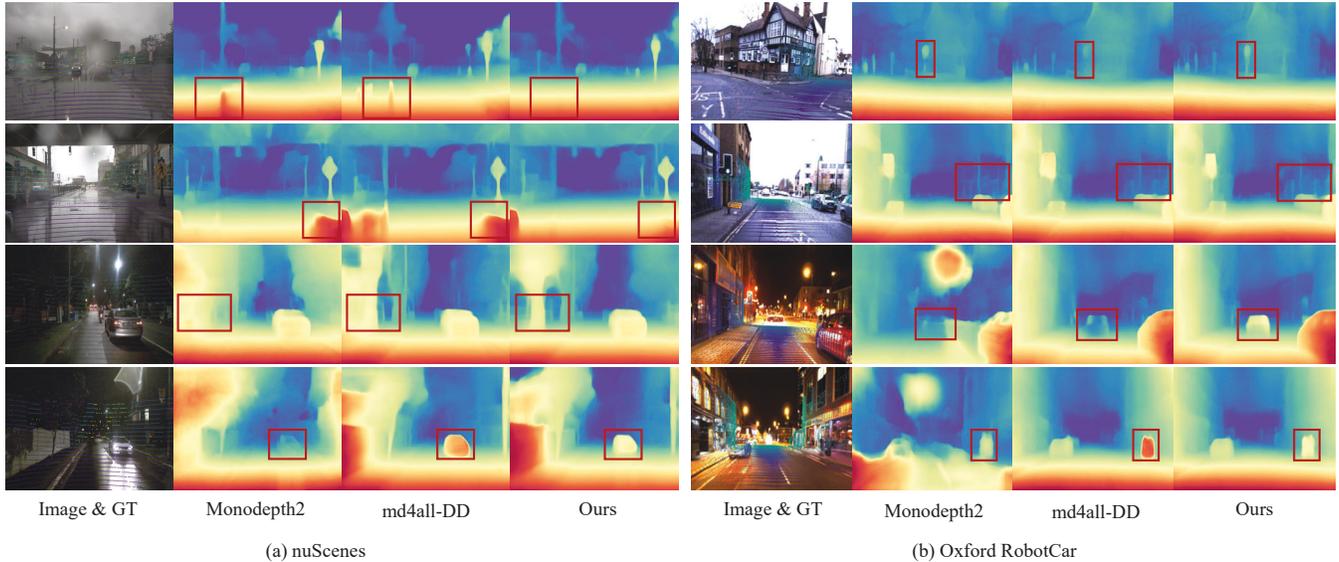


Figure 4: Qualitative results on nuScenes [Caesar *et al.*, 2020] and RobotCar [Maddern *et al.*, 2017]. We compare the ACDepth approach with md4all-DD and MonoDepth2, all of which use the same backbone. To better illustrate the results, the real point cloud is projected onto the original image, and no ground truth (GT) is required during training.

Method	sup.	tr.data	day-clear – nuScenes			night – nuScenes			day-rain – nuScenes		
			absRel	RMSE	δ_1	absRel	RMSE	δ_1	absRel	RMSE	δ_1
Monodepth2 [Godard <i>et al.</i> , 2019]	M*	a: dnr	0.1477	6.771	85.25	2.3332	32.940	10.54	0.4114	9.442	60.58
Monodepth2 [Godard <i>et al.</i> , 2019]	M*	d	0.1374	6.692	85.00	0.2828	9.729	51.83	0.1727	7.743	77.57
PackNet-SfM [Guizilini <i>et al.</i> , 2020]	Mv	d	0.1567	7.230	82.64	0.2617	11.063	56.64	0.1645	8.288	77.07
RNW [Wang <i>et al.</i> , 2021]	M*	dn	0.2872	9.185	56.21	0.3333	10.098	43.72	0.2952	9.341	57.21
md4all-AD [Gasparini <i>et al.</i> , 2023]	Mv	dT \ddagger (nr)	0.1523	6.853	83.11	0.2187	9.003	68.84	0.1601	7.832	78.97
md4all-DD [Gasparini <i>et al.</i> , 2023]	Mv	dT \ddagger (nr)	0.1366	6.452	84.61	0.1921	8.507	<u>71.07</u>	0.1414	7.228	80.98
DMMDE v1 [Tosi <i>et al.</i> , 2025]	Ms	dT \ddagger (nr)	0.1370	<u>6.318</u>	85.05	<u>0.1880</u>	<u>8.432</u>	69.94	0.1470	7.345	79.59
DMMDE v2 [Tosi <i>et al.</i> , 2025]	Ms	dT \ddagger (nr)	0.1400	6.573	83.51	0.1970	8.826	69.65	0.1430	7.317	80.28
DMMDE v3 [Tosi <i>et al.</i> , 2025]	Ms	dT \ddagger (nr)	0.1280	6.449	84.03	0.1910	8.433	71.14	<u>0.1390</u>	<u>7.129</u>	81.36
ACDepth (Ours)	Mv	dT(nr)	<u>0.1340</u>	6.284	<u>85.07</u>	0.1873	8.125	71.14	0.1377	6.970	<u>81.32</u>

Table 1: Evaluation of self-supervised methods on nuScenes [Caesar *et al.*, 2020] validation set. v1,v2,v3: depth maps from md4all-DD, DPT, Depth Anything in [Tosi *et al.*, 2025]. Supervisions (sup.): M: via monocular videos, *: test-time median-scaling via LiDAR, v: weak velocity, s: test-time scaling via LSE criterion [Ranftl *et al.*, 2020]. Training data (tr.data): d: day-clear, n: night, r: rain, a: all. T \ddagger : Translated via GAN, T \ddagger : Translated via Diffusion. Visual support: **1st** and 2nd.

Figure. 4(a) presents qualitative results. md4all-DD is particularly sensitive to water reflections in rainy scenes. For example, in the first and second rows of Figure. 4(a), the obvious errors of depth estimation are observed in the regions of water surface reflections. In addition for nighttime scenes, md4all-DD is hard to predict the real depth of regions hidden in the darkness. In contrast, our ACDepth method can produce reasonable and reliable predictions of depth information in night scenes, such as trees and cars under low-light conditions.

Comparison on the RobotCar. To further verify the effectiveness, we compare our ACDepth with existing state-of-the-art approaches on the RobotCar benchmark. As expected in Table 2, our method achieves the best performance in almost all metrics. Compared to md4all-DD, our approach reduces RMSE by 8.80% on the standard benchmark and 11.99% on the challenging benchmark, highlighting the superiority of our approach. To demonstrate the effectiveness of our method, Figure. 4(b) provides the visual comparisons, showing that our ACDepth can produce more precise depth maps

over those of md4all-DD on both day and night scenes. We speculate that these considerable improvement of this study stems from the more reliable data generation and comprehensive prior constraints, which contribute more to the robust model training and optimization.

4.4 Ablation Study

In this section, we conduct detailed ablation experiments on nuScenes and RobotCar datasets to demonstrate the individual effectiveness of the proposed components.

Evaluation on Major Design Components. Our baseline is trained on daytime scenes [Gasparini *et al.*, 2023]. Based on this, we introduce the distillation learning (DL) to promote training with more robust feature representation, reducing the absRel metric by 22.16% for night scene and 10.69% for rainy scene on the nuScenes. Then we construct another model with additional ordinal guidance distillation (OGD), which provides more precise supervision signals for degraded conditions. By contrast, it achieves the comprehensive im-

Method	sup.	tr.data	day – RobotCar				night – RobotCar			
			absRel	sqRel	RMSE	δ_1	absRel	sqRel	RMSE	δ_1
Monodepth2 [Godard <i>et al.</i> , 2019]	M*	d	0.1196	0.670	<u>3.164</u>	86.38	0.3029	1.724	5.038	45.88
DeFeatNet [Spencer <i>et al.</i> , 2020]	M*	a: dn	0.2470	2.980	7.884	65.00	0.3340	4.589	8.606	58.60
ADIDS [Liu <i>et al.</i> , 2021]	M*	a: dn	0.2390	2.089	6.743	61.40	0.2870	2.569	7.985	49.00
RNW [Wang <i>et al.</i> , 2021]	M*	a: dn	0.2970	2.608	7.996	43.10	0.1850	1.710	6.549	73.30
WSGD [Vankadari <i>et al.</i> , 2023]	M*	a: dn	0.1760	1.603	6.036	75.00	0.1740	1.637	6.302	75.40
md4all-DD [Gasperini <i>et al.</i> , 2023]	Mv	dT \ddagger (n)	0.1128	0.648	3.206	87.13	0.1219	0.784	<u>3.604</u>	84.86
DMMDE v1 [Tosi <i>et al.</i> , 2025]	Mv	dT \ddagger (n)	0.1190	0.676	3.239	87.20	0.1390	<u>0.739</u>	3.700	82.46
DMMDE v2 [Tosi <i>et al.</i> , 2025]	Mv	dT \ddagger (n)	0.1230	0.724	3.333	86.62	0.1330	0.824	3.712	83.95
DMMDE v3 [Tosi <i>et al.</i> , 2025]	Mv	dT \ddagger (n)	0.1190	0.728	3.287	<u>87.17</u>	0.1290	0.751	3.661	83.68
ACDepth (Ours)	Mv	dT(n)	0.1107	0.591	3.084	88.03	0.1206	0.690	3.432	<u>84.47</u>

Table 2: Evaluation of self-supervised works on the RobotCar [Maddern *et al.*, 2017] test set. Trailing 0 added to the values from [Vankadari *et al.*, 2023] and [Tosi *et al.*, 2025]. Notation from Table 1.

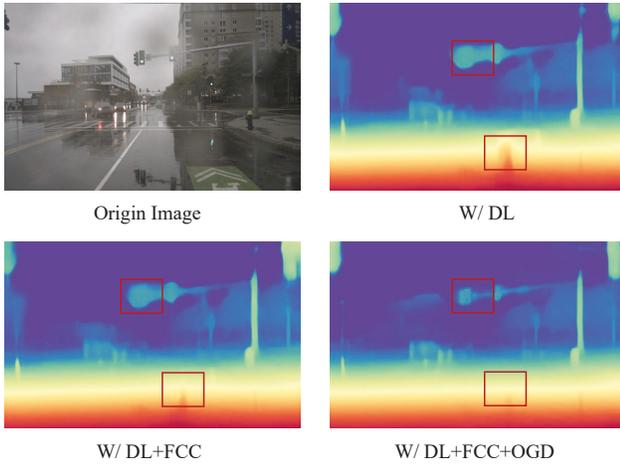


Figure 5: Visualization of ablation study on the distillation learning, feature consistency constraint and ordinal guidance distillation.

	DL	OGD	FCC	day-clear		night		day-rain	
				absRel	RMSE	absRel	RMSE	absRel	RMSE
N	✓			0.1333	6.459	0.2419	10.922	0.1572	7.453
	✓			0.1335	6.408	0.1883	8.398	0.1404	7.092
	✓	✓		0.1325	6.328	0.1879	8.353	0.1414	7.084
	✓	✓	✓	0.1355	6.340	0.1872	8.125	0.1377	6.999
	DL	OGD	FCC	day			night		
				absRel	RMSE	sqRel	absRel	RMSE	sqRel
R	✓			0.1209	3.335	0.723	0.3909	8.227	3.547
	✓			0.1123	3.135	0.631	0.1233	3.476	0.720
	✓	✓		0.1117	3.115	0.615	0.1224	3.458	0.704
	✓	✓	✓	0.1107	3.084	0.591	0.1206	3.432	0.690

Table 3: Ablation study of Design Components. N: nuScenes, R: RobotCar, DL: distillation learning, OGD: ordinal guidance distillation, FCC: feature consistency constraint.

improvements in model performance, reducing the absRel metric by 0.54% for day scene and 0.73% for night scene on the RobotCar. In addition, we introduce the feature consistency constraint (FCC) to evaluate its effect for robust feature representation. As expected, the model with FCC significantly enhances the accuracy of depth estimation in challenging conditions (such as nighttime and rainy scenes), reducing the absRel metric by 0.37% for night scene and 2.62% for rain scene on the nuScenes. We also provide the visual comparisons between the improved versions. As shown in Figure 5, the complete model equipped with DL, OGD and FCC strategies shows significant superiority over its imperfect versions.

These optimization strategies not only provide reliable supervised signals but also guide the network to focus on uncertain regions, resulting in more robust depth estimation.

Evaluation on Ordinal Guidance Distillation. In this section, we investigate the impact of different sampling methods on ordinal guidance distillation, with the experimental results presented in Table 4. Our OGD strategy employs two sampling methods: global and local sampling. We observed that model performance became more reliable as both sampling methods were progressively incorporated. Additionally, our experiments demonstrate that using windowed sampling regions enables the model to learn more robust feature representations, significantly reducing the risk of overfitting.

	G	L	W	day-clear		night		day-rain	
				absRel	RMSE	absRel	RMSE	absRel	RMSE
N				0.1371	6.560	0.1884	8.209	0.1412	7.178
	✓			0.1343	6.384	0.1876	8.161	0.1401	7.004
	✓	✓		0.1330	6.318	0.1875	8.141	0.1377	6.982
	✓	✓	✓	0.1355	6.340	0.1872	8.125	0.1377	6.999
	G	L	W	day			night		
				absRel	RMSE	sqRel	absRel	RMSE	sqRel
R				0.1144	3.184	0.642	0.1230	3.448	0.721
	✓			0.1128	3.157	0.623	0.1228	3.443	0.691
	✓	✓		0.1107	3.095	0.600	0.1223	3.453	0.685
	✓	✓	✓	0.1107	3.084	0.591	0.1206	3.432	0.690

Table 4: Ablation study of sampling method for ranking loss. G: global sampling, L: local sampling, W: window sampling.

5 Conclusion

In this paper, we propose a novel approach named ACDepth for robust monocular depth estimation under adverse weather conditions. In addition, we introduce an elaborate data generation scheme to produce the multi-tuple depth dataset with diverse degradations, which significantly mitigates the problem of data scarcity. Meanwhile, we construct an effective multi-granularity knowledge distillation (MKD) strategy to achieve the robust model training, which facilitates the complete transfer and alignment of capabilities between the teacher model and student model. Extensive experimental results and comprehensive ablation demonstrate the effectiveness of our ACDepth, demonstrating its superior performance compared to SoTA solutions.

Acknowledgments

This research was financially supported by the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006), the National Natural Science Foundation of China (U23B2009) and Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-24-09).

References

- [Arampatzakis *et al.*, 2023] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2396–2414, 2023.
- [Bhat *et al.*, 2023] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023.
- [Bian *et al.*, 2019] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in Neural Information Processing Systems*, 32, 2019.
- [Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.
- [Garg *et al.*, 2016] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016.
- [Gasperini *et al.*, 2023] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *ICCV*, pages 8177–8186, 2023.
- [Godard *et al.*, 2017] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017.
- [Godard *et al.*, 2019] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.
- [Guizilini *et al.*, 2020] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020.
- [Guo *et al.*, 2020] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pages 1780–1789, 2020.
- [Häne *et al.*, 2011] Christian Häne, Christopher Zach, Jongwoo Lim, Ananth Ranganathan, and Marc Pollefeys. Stereo depth map fusion for robot navigation. In *IROS*, pages 1618–1625, 2011.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv*, 2015.
- [Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*, 2021.
- [Liu *et al.*, 2015a] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.
- [Liu *et al.*, 2015b] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:2024–2039, 2015.
- [Liu *et al.*, 2021] Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *ICCV*, pages 12737–12746, 2021.
- [Lyu *et al.*, 2021] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*, pages 2294–2301, 2021.
- [Maddern *et al.*, 2017] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36:3–15, 2017.
- [Mou *et al.*, 2024] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adaptor: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024.
- [Parmar *et al.*, 2024] Gaurav Parmar, Taesung Park, Srini-vasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv*, 2024.
- [Ranftl *et al.*, 2020] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020.
- [Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *CVPR*, pages 12179–12188, 2021.
- [Sauer *et al.*, 2025] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pages 87–103, 2025.

- [Saunders *et al.*, 2023] Kieran Saunders, George Vogiatzis, and Luis J Manso. Self-supervised monocular depth estimation: Let’s talk about the weather. In *ICCV*, pages 8907–8917, 2023.
- [Schön *et al.*, 2021] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. In *ICCV*, pages 15804–15815, 2021.
- [Shi *et al.*, 2023] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. In *ROBIO*, pages 1–7, 2023.
- [Song and Yoon, 2022] Kyeongseob Song and Kuk-Jin Yoon. Learning monocular depth estimation via selective distillation of stereo knowledge. *arXiv*, 2022.
- [Spencer *et al.*, 2020] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *CVPR*, pages 14402–14413, 2020.
- [Sun *et al.*, 2012] Shaoyuan Sun, Linna Li, and Lin Xi. Depth estimation from monocular infrared images based on bp neural network model. In *CVRS*, pages 237–241, 2012.
- [Sun *et al.*, 2023] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:497–508, 2023.
- [Tosi *et al.*, 2025] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *ECCV*, pages 236–257, 2025.
- [Vankadari *et al.*, 2023] Madhu Vankadari, Stuart Golodetz, Sourav Garg, Sangyun Shin, Andrew Markham, and Niki Trigoni. When the sun goes down: Repairing photometric losses for all-day depth estimation. In *CoRL*, pages 1992–2003, 2023.
- [Wang *et al.*, 2021] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *ICCV*, pages 16055–16064, 2021.
- [Wang *et al.*, 2024a] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *ICRA*, pages 4976–4982, 2024.
- [Wang *et al.*, 2024b] Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. In *ACM MM*, pages 4129–4137, 2024.
- [Xian *et al.*, 2020] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, pages 611–620, 2020.
- [Xue *et al.*, 2020] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *IROS*, pages 2330–2337, 2020.
- [Yang *et al.*, 2024a] Haolin Yang, Chaoqiang Zhao, Lu Sheng, and Yang Tang. Self-supervised monocular depth estimation in the dark: Towards data distribution compensation. In *IJCAI*, pages 1561–1569, 2024.
- [Yang *et al.*, 2024b] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024.
- [Yin *et al.*, 2022] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:6480–6494, 2022.
- [Yu and Gallup, 2014] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. In *CVPR*, pages 3986–3993, 2014.
- [Zhang *et al.*, 2023] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *CVPR*, pages 18537–18546, 2023.
- [Zhao *et al.*, 2022] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *3DV*, pages 668–678, 2022.
- [Zheng *et al.*, 2023] Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, et al. Steps: Joint self-supervised nighttime image enhancement and depth estimation. In *ICRA*, pages 4916–4923, 2023.
- [Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.
- [Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, pages 2223–2232, 2017.
- [Zhu *et al.*, 2023] Ruijie Zhu, Ziyang Song, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Ec-depth: Exploring the consistency of self-supervised monocular depth estimation under challenging scenes. *arXiv*, 2023.