MonoDiffusion: Self-Supervised Monocular Depth Estimation Using Diffusion Model

Shuwei Shao, Zhongcai Pei, Weihai Chen*, Dingchi Sun, Peter C.Y.Chen and Zhengguo Li*, *Fellow, IEEE*

***Abstract*—Over the past few years, self-supervised monocular depth estimation has received widespread attention. Most efforts focus on designing different types of network architectures and loss functions or handling edge cases, for example, occlusion and dynamic objects. In this work, we take another path and propose a novel conditional diffusion-based generative framework for self-supervised monocular depth estimation, dubbed MonoDiffusion. Because the depth ground-truth is unavailable in a self-supervised setting, we develop a new pseudo ground-truth diffusion process to assist the diffusion for training. Instead of diffusing at a fixed high resolution, we perform diffusion in a coarse-to-fine manner that allows for faster inference time without sacrificing accuracy or even better accuracy. Furthermore, we develop a simple yet effective contrastive depth reconstruction mechanism to enhance the denoising ability of model. It is worth noting that the proposed MonoDiffusion has the property of naturally acquiring the depth uncertainty that is essential to be implemented in safety-critical cases. Extensive experiments on the KITTI, Make3D and DIML datasets indicate that our MonoDiffusion outperforms prior state-of-the-art self-supervised competitors. The source code will be publicly available upon the acceptance.**

***Index Terms*—Monocular depth estimation, Conditional diffusion, Self-supervised learning**

I. Introduction

Monocular depth estimation (MDE) is one of the fundamental tasks in the computer vision community, with many applications, *e.g.*, 3D reconstruction, scene understanding and autonomous driving [1]–[3]. Recently, learning-based approaches [4]–[6] have achieved remarkable advances, where the fully supervised MDE [7]–[11] achieves higher accuracy due to the available depth ground-truth. Nevertheless, the depth ground-truth is hard to acquire because of expensive hardware sensors, sensor noise, limited operating capabilities, etc.

Self-supervised MDE [5], [12]–[14] has been proposed as a promising alternative by formulating the MDE as a task of novel view synthesis. When leveraging stereo image pairs, the motion of the camera is available, allowing for the use of a separate depth estimation network in the training phase. When training on monocular videos, an additional pose network is necessary to estimate the camera motion. Nevertheless, self-supervised MDE that depends solely on monocular videos is preferred because of the difficulties in collecting stereo data, for example, intricate configurations and data processing. In order to boost the performance of the task, recent approaches have focused on developing improved loss functions [15]–[17], leveraging geometric constraints [18] and disparity offset refinement [19], incorporating discrete disparity prediction [20] and feature representation learning [21], employing semantic information [22], [23] to deal with dynamic objects and occlusions, and integrating Convolutional Neural Networks (CNNs) with Transformers [24] to implement more powerful network architectures. As a result, they fall under the discriminative-based paradigm.

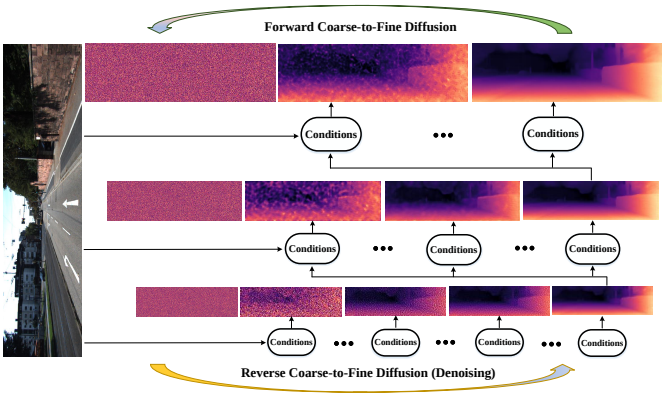

Fig. 1. **Conditional coarse-to-fine diffusion process in our MonoDiffusion.** It consists of three stages at different resolutions. During the two finer stages, the denoised depth prediction from the prior stage is combined with the visual image as conditions to strengthen the denoising process.

Diffusion models [25]–[28] have emerged as highly effective generative models, showcasing impressive success [29]. Based upon these progressions, recent studies [30]–[32] have searched into controllable image synthesis by leveraging external conditions. These approaches now make training more efficient and achieve superior performance across a wide range of generative tasks [33]–[35]. In light of this, there has been a growing surge of interest in introducing diffusion models to downstream discriminative tasks, *e.g.*, face anti-spoofing [36], dense matching [37], [38], pose estimation [39], [40], semantic segementation [41]–[44] and depth estimation [45]–[49]. They have achieved fruitful outcomes but are limited to a supervised setting where the target ground-truth is available. Given the impressive generative capabilities of diffusion models, it is natural to inquire how efficacious these models are for self-

This work was supported in part by the National Natural Science Foundation of China under grant 61620106012 and in part by the A*STAR-MTC Programmatic Funds under grant M23L760021.

Shuwei Shao, Zhongcai Pei, Weihai Chen and Dingchi Sun are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. (email: swshao@buaa.edu.cn, peizc@buaa.edu.cn, whchen@buaa.edu.cn, sdc@nuaa.edu.cn)

Peter C.Y.Chen is with the Department of Mechanical Engineering, National University of Singapore, Singapore. (e-mail: mpechenp@nus.edu.sg)

Zhengguo Li is with the VI department, Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore. (e-mail: ezgli@i2r.a-star.edu.sg)

*(Corresponding authors: Weihai Chen and Zhengguo Li.)

Fig. 2. **Illustration of the depth smearing phenomenon, which has been highlighted by the red box.** The depth of the tree silhouettes exhibits severe smearing.

supervised monocular depth estimation, especially considering that the target depth ground-truth is unavailable during training and therefore standard forward diffusion is not applicable. Moreover, in contrast to prior discriminative-based self-supervised depth estimation methods [15]–[20], introducing diffusion models enjoys the following several benefits: i) The powerful generative capabilities hold the potential of deriving higher-quality depth predictions; ii) Reframing self-supervised monocular depth estimation as a conditional denoising process offers a promising avenue to push the boundaries of the task to greater heights; iii) Natural awareness of the prediction uncertainty enables safety-critical applications.

In this paper, we give the answer and propose MonoDiffusion, a novel conditional diffusion-based generative framework for self-supervised monocular depth estimation. During training, we introduce a new pseudo ground-truth diffusion process to implement the forward diffusion in a coarse-to-fine manner, which consists of three stages. Specifically, we first pre-train a standard depth model in a self-supervised fashion, which we designate as a teacher model. At the initial stage, we append the Gaussian noise governed by a noise schedule [26] to the coarsest pseudo ground-truth generated by the teacher model to obtain noisy samples. Next, we embed the noisy samples to acquire noise embeddings and merge them with the visual conditions extracted from the image and the time embeddings acquired from the current timestep. These merged features are then input into a noise predictor to obtain a noise prediction. The following two stages are similar to the initial stage. The differences are that we perform diffusion on the finer pseudo ground-truth and further encode the denoised depth prediction from the previous stage to acquire geometric conditions, which are combined with the image conditions as visual conditions for the next stage. From coarse to fine, the number of total forward and reverse diffusion steps decreases. Compared with the traditional fixed high-resolution diffusion, such a coarse-to-fine diffusion manner (Fig. 1) allows for faster inference time without compromising accuracy or even better accuracy. As a by-product, the teacher model also enables us to impose a knowledge distillation loss on the denoised depth prediction to improve the results. In order to alleviate the negative impact of depth errors in pseudo ground-truth, we leverage a multi-view check filter [19] to filter out erroneous depth. The depth of a scene point, when predicted using images from different views, is expected to be identical or very close after being transformed to the same view. If it is not, this indicates that the depth data contains large errors and should be discarded. Furthermore, we observe that the denoised depth prediction tends to display depth smearing in boundary regions, as shown in Fig. 2. To alleviate this, we develop a simple yet effective contrastive depth reconstruction mechanism to strengthen the model denoising ability through fully exploiting the contextual information.

To summarize, our contributions are listed as follows:

- We investigate the utility and efficacy of diffusion models for self-supervised monocular depth estimation, in which the standard forward diffusion is not practical due to the absence of depth ground-truth.
- We introduce a generative framework, termed as MonoDiffusion, by formulating self-supervised monocular depth estimation as a conditional denoising process, providing a compelling pathway to push the boundaries of the task to newer heights.
- We devise a coarse-to-fine pseudo ground-truth diffusion process to realize the forward diffusion and a contrastive depth reconstruction mechanism to further reinforce the model denoising ability.
- The proposed MonoDiffusion outperforms previous state-of-the-art self-supervised competitors on the KITTI [50], Make3D [51] and DIML [9] datasets.

II. Related Work

A. Monocular Depth Estimation

MDE aims to predict depth from a single image, which is an ill-posed problem because there are infinitely many 3D scenes that can be projected onto the identical 2D image. The relevant methods are broadly categorized into two groups.

Supervised depth estimation uses depth ground-truth to establish supervision and has achieved outstanding performance. Eigen *et al.* [4] introduced the first attempt of utilizing CNN to perform multi-scale depth estimation. Then, Laina *et al.* [52] utilized the residual CNN [53] to enable network optimization to be easier. Cao *et al.* [54] and Fu *et al.* [55] discretized the full depth range into multiple intervals and chose the optimal interval as depth prediction. Yuan *et al.* [56] developed neural window fully-connected conditional random fields (CRFs) to reduce the computation in conventional CRFs. Shao *et al.* [57] introduced the cross-distillation paradigm to integrate strengths from CNN and Transformer. Liu *et al.* [58] enforced the first-order variational constraint to regularize depth map. Shao *et al.* [59] decoupled depth estimation into two subtasks, surface normal and plane-to-origin distance estimation, to leverage the inherent 3D geometric constraints. Bhat *et al.* [60] combined the relative depth pre-training and metric depth fine-tuning to achieve excellent generalization performance while preserving metric scale. Nonetheless, these approaches depend heavily on the depth ground-truth, which poses challenges in acquisition

due to costly hardware sensors, sensor noise, limited operating capabilities, etc.

Self-supervised depth estimation does not require costly depth ground-truth and employs stereo image pairs or adjacent frames in a video to generate the supervisory signal. As one of the pioneering works, Zhou *et al.* [5] trained a depth network and a pose network via novel view synthesis to estimate depth and 6-degrees of freedom pose from unlabeled video sequence. To handle edge cases such as dynamic objects and occlusion, they developed an explainability mask to remove these regions. Zou *et al.* [15] calculated rigid flow using depth and pose, and enforced a consistency constraint with predicted optical flow as an auxiliary supervisory signal. Godard *et al.* [16] introduced a classical method and devised an auto-masking technique and a per-pixel minimum re-projection loss to better handle dynamic objects along with occlusion. Later, numerous self-supervised MDE approaches have sprung out. Bian *et al.* [17] developed a geometry consistency loss to obtain scale-consistent depth and pose predictions. Liu *et al.* [61] designed a domain-separated network to derive depth maps for both day and night images. Ramamonjisoa *et al.* [62] exploited wavelet decomposition to achieve efficient depth estimation. Johnston *et al.* [20] divided the whole disparity range into multiple intervals and then used the discrete disparity prediction to preserve depth boundaries. Guizilini *et al.* [63] utilized symmetrical packing and unpacking operations to train the compression and decompression of detail-preserving representations. Liu *et al.* [19] utilized self-reference distillation and disparity offset refinement to further improve the performance. Zhang *et al.* [24] developed an efficient hybrid architecture by combining CNN and Transformer. Other approaches involve feature representation learning [21], surface normal [18] and semantic segmentation [23], [64]. In a stark contrast, we introduce the diffusion models and draw on its strong generative capability to produce high-quality depth predictions.

B. Diffusion Models

Diffusion models [25], [65] have been subject to widespread attention due to their potent generation capability. The Denoising Diffusion Probabilistic Model (DDPM) [25] introduced a diffusion model featuring Markovian properties in both its forward and reverse processes. The Denoising Diffusion Implicit Model (DDIM) [26] improved upon DDPM by utilizing non-Markovian chains to substitute the original diffusion process, resulting in faster sampling speeds. Expanding on these breakthroughs, conditional diffusion models have emerged, utilizing auxiliary conditions to implement controlled image synthesis. Saharia *et al.* [31] introduced a versatile framework for image-to-image translation by using the source image as an additional condition through concatenation. In a similar vein, Brooks *et al.* [66] made use of a conditional diffusion model trained with paired image and text instructions for instruction-driven image editing. On the flip side, several studies [67]–[69] have shifted their focus towards enhancing image resolution. The Cascaded Diffusion Model [67], for instance, utilized a cascaded pipeline to gradually boost the resolution of synthesized images via the diffusion denoising process.

C. Diffusion Models for Downstream Tasks

Lately, the impressive capabilities of diffusion models have been broadened to deal with downstream discriminative tasks, involving image segmentation [70]–[72], object detection [73], dense matching [37], [38], pose estimation [39], [40], face anti-spoofing [36] and depth estimation [45]–[49]. Nevertheless, these pioneering studies are restricted to a supervised setting where the target ground-truth is provided. In scenarios where the target ground-truth is unavailable during training, for example, in self-supervised MDE, standard forward diffusion becomes impractical. Thus, the effectiveness and applicability of diffusion models in such scenarios remain unknown. This paper aims to shed light on this issue. Additionally, rather than employing a heavy U-Net with an encoder-decoder structure to predict noise as conducted in [45], we develop a lightweight noise predictor that consists of just three convolutional layers and one embedding layer. In order to further strengthen the computational efficiency, we develop a coarse-to-fine diffusion process, enabling MonoDiffusion to achieve satisfactory results with only three denoising steps at the highest resolution.

III. Preliminaries

Conditional diffusion models. The diffusion models, categorized within generative models, consist of two main types: unconditional models [25], [65] and conditional models [74], [75]. Concretely, unconditional diffusion models focus on acquiring a direct approximation of the data distribution through the learning process. Conditional diffusion models, in contrast, are developed to estimate the data distribution under specific conditions, for example, text prompt [74].

In the conditional diffusion models, the data distribution is approximated by iteratively denoising a data sample from the Gaussian noise. Suppose a sample $\mathbf{x}_0$, the forward diffusion process involves the iterative addition of noise to $\mathbf{x}_0$ and gets the noisy sample $\mathbf{x}_\tau$. Mathematically,

$$q\left(\mathbf{x}_\tau \left|\mathbf{x}_0\right.\right) := \mathcal{N}\left(\mathbf{x}_\tau \left|\sqrt{\overline{\alpha}_\tau}\mathbf{x}_0, (1-\overline{\alpha}_\tau)\mathbf{I}\right.\right), \tag{1}$$

with

$$\overline{\alpha}_\tau := \prod\nolimits_{n=0}^{\tau} \alpha_\tau = \prod\nolimits_{n=0}^{\tau} (1-\beta_n), \tag{2}$$

and

$$\mathbf{x}_\tau = \sqrt{\overline{\alpha}_\tau}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_\tau}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \tag{3}$$

where τ consists of $\mathbf{T}$ steps, β_n stands for the noise variance schedule as DDPM [25] and $\mathcal{N}$ is the Gaussian noise.

As for the denoising process, a noise predictor ϵ_θ learns to reverse the diffusion process and recover $\mathbf{x}_0$ under the guidance of visual conditions $\mathbf{c}$. Each denoising step is approximated by a Gaussian distribution,

$$p_\theta\left(\mathbf{x}_{\tau-1} \left|\mathbf{x}_\tau, \mathbf{c}\right.\right) := \mathcal{N}\left(\mathbf{x}_{\tau-1}; \mu_\theta\left(\mathbf{x}_\tau, \tau, \mathbf{c}\right), \sigma_\tau^2\mathbf{I}\right), \tag{4}$$

where $\mu_\theta\left(\mathbf{x}_\tau, \tau, \mathbf{c}\right)$ is acquired through the linear combination of $\mathbf{x}_\tau$ and predicted noise $\epsilon_\theta\left(\mathbf{x}_\tau, \tau, \mathbf{c}\right)$, and σ_τ^2 stands for the transition variance.

Self-supervised MDE frames the depth estimation as a task of novel view synthesis and typically requires two networks, a depth network and an additional pose network. The depth

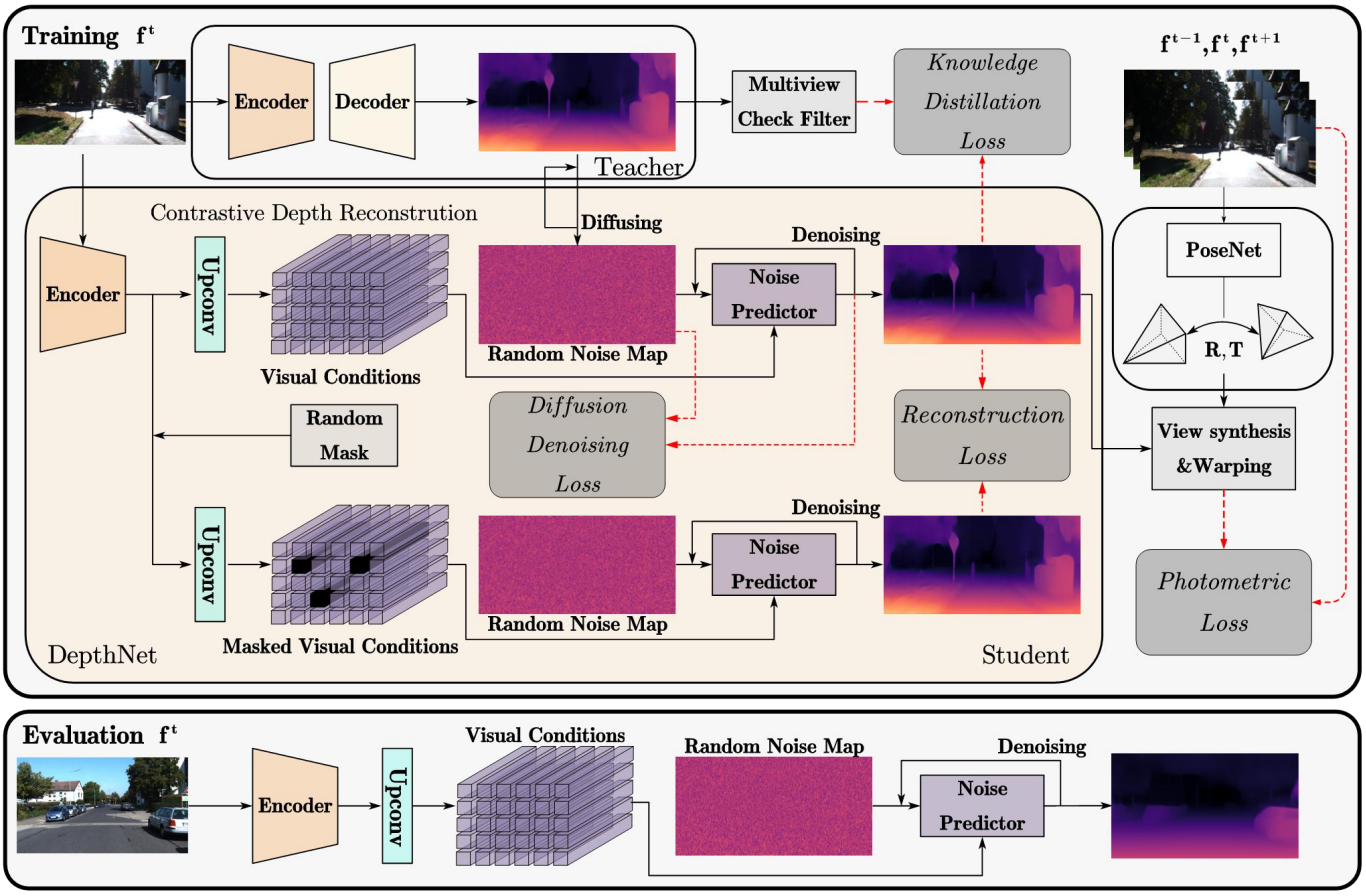

Fig. 3. **An overview of the proposed MonoDiffusion**. To facilitate training, we develop the coarse-to-fine pseudo ground-truth diffusion and contrastive depth reconstruction. The knowledge distillation helps improve denoised depth predictions, while PoseNet and DepthNet are combined to achieve self-supervision. It should be noted that the target frame $\mathbf{f}^t$ is separately paired with the source frames $\mathbf{f}^{t-1}$ and $\mathbf{f}^{t+1}$ to acquire stacked image pairs, which are then fed into PoseNet to generate the corresponding relative pose for each pair. In the evaluation phase, MonoDiffusion starts with a random noise map and iteratively refines depth predictions using the coarse-to-fine denoising approach. The details of coarse-to-fine diffusion process are simplified in the flowchart.

network takes a single RGB image as input and generates a corresponding depth map, while the pose network processes a concatenated pair of images to estimate the 6-degrees of freedom (DOF) relative pose involving three Euler angles and a translation vector between them. To achieve supervision, the predicted depth map is utilized to back-project each pixel into 3D camera space with the camera intrinsic. The acquired 3D point cloud is then transformed to another view utilizing the predicted relative pose. Suppose a target frame $\mathbf{f}^t$ and a source frame $\mathbf{f}^s$, the procedure is defined as

$$\mathbf{p}^{t\to s} = \mathbf{K}\mathbf{M}^{t\to s}\mathbf{D}^t\mathbf{K}^{-1}\mathbf{p}^t, \tag{5}$$

where $\mathbf{p}^{t\to s}$ denotes the mapped pixel coordinate from target view t to source view s, $\mathbf{p}^t$ denotes the pixel coordinate in the target view, $\mathbf{K}$ denotes the camera intrinsic, $\mathbf{M}^{t\to s}$ stands for the transformation matrix from target view to source view, converted from the 6-DOF pose and composed of a rotation matrix and a translation vector, and $\mathbf{D}^t$ denotes the depth map of target frame. Next, we can obtain a warped frame $\mathbf{f}^{s\to t}$ via

$$\mathbf{f}^{s\to t}(\mathbf{p}) = \mathbf{f}^s\left\langle \mathbf{p}^{t\to s}\right\rangle, \tag{6}$$

where $\langle\cdot\rangle$ denotes the warping operation [76]. The appearance difference between $\mathbf{f}^{s\to t}$ and $\mathbf{f}^t$ is used to supervise the whole framework. As in most works, *e.g.*, [16], [77], a combination of L1 loss and structural similarity (SSIM) term [78] measures the dissimilarity in appearance, written as

$$\begin{aligned}\mathcal{L}_{ph} = &\sum_{\mathbf{p}} \kappa \frac{1-\mathrm{SSIM}\left(\mathbf{f}^t(\mathbf{p}), \mathbf{f}^{s\to t}(\mathbf{p})\right)}{2} \\ &+ \sum_{\mathbf{p}} (1-\kappa)\left\|\mathbf{f}^t(\mathbf{p}) - \mathbf{f}^{s\to t}(\mathbf{p})\right\|_1,\end{aligned} \tag{7}$$

where $\mathcal{L}_{ph}$ is referred to as the photometric loss and κ is set to 0.85.

IV. Methodology

A. Overview

Fig.3 illustrates a comprehensive overview of our proposed framework. During the training phase, MonoDiffusion mainly involves the coarse-to-fine pseudo ground-truth diffusion and the contrastive depth reconstruction. The knowledge distillation is integrated into the framework, facilitating more accurate denoised depth predictions. The teacher model based on Lite-Mono [24], which is pre-trained in a self-supervised manner, guides the learning process but is discarded once the training

is completed. The PoseNet is deployed together with DepthNet to achieve self-supervision. During the evaluation phase, MonoDiffusion carries out a coarse-to-fine denoising process, initiating with a random noise map and gradually refining the depth predictions through iterative denoising steps. In order to streamline the flowchart, the details of coarse-to-fine diffusion are omitted. More specifics will be elaborated in the following sections.

B. Training

The diffusion-based methods typically construct the forward diffusion process from the target ground-truth, which however is not available in self-supervised MDE during training. Additionally, the diffusion models recover samples in a progressive manner, necessitating multiple iterations of the model. In order to obtain satisfactory results, previous methods, such as [46], employ a large number of iterations on a fixed high resolution (20 denoising steps in [46]), which significantly increases the computational overhead. In terms of the first issue, we develop a pseudo ground-truth diffusion process. The pseudo ground-truth is generated by a pre-trained teacher model. As for the second problem, we develop a coarse-to-fine diffusion process, where more iterations are performed in the coarser stages to reduce the computational complexity while fewer iterations are executed in the finer stages to capture the fine-grained details. Our MonoDiffusion achieves satisfactory results requiring only three denoising steps at the highest resolution.

Coarse-to-fine pseudo ground-truth diffusion. We begin by pre-training a teacher model based upon Lite-Mono [24], a self-supervised monocular depth estimator. The full diffusion process consists of three stages, performed at 1/4 resolution, 1/2 resolution, and original resolution, respectively.

In the initial stage, we append the Gaussian noise, controlled by a noise schedule [26], to the coarsest pseudo ground-truth generated by the teacher model, achieving a noisy sample $\mathbf{D}_\tau$,

$$q\left(\mathbf{D}_\tau \left|\mathbf{D}_{pseudo}\right.\right) := \mathcal{N}\left(\mathbf{D}_\tau \left|\sqrt{\overline{\alpha}_\tau}\mathbf{D}_{pseudo}, (1-\overline{\alpha}_\tau)\mathbf{I}\right.\right), \quad (8)$$

where $\mathbf{D}_{pseudo}$ corresponds to the pseudo ground-truth. Next, we use a convolutional layer to embed the noisy sample and an embedding layer to embed the current timestep. The resultant noise and time embeddings are fused with the visual conditions extracted from the image. Then, these fused features are input into the noise predictor to derive a noise prediction.

The subsequent stages follow a similar procedure, but with some variations. Specifically, we append the Gaussian noise to the finer pseudo ground-truth. The denoised depth prediction from the previous stage is further encoded via a convolutional layer to obtain geometric conditions, which are combined with the image conditions to constitute visual conditions. The image conditions are varied across different stages and correspond to the 1/4 resolution features, 1/2 resolution features, and original resolution features from the decoder, providing rich multi-scale information. From the initial to last stage, the number of total forward and reverse diffusion steps are set to 250, 200, 150 and 5, 4, 3, respectively.

As a by-product, the teacher model also allows for imposing a knowledge distillation loss on the denoised depth prediction. The knowledge distillation paradigm distills knowledge from the teacher model to the student model, which is beneficial to improve the performance [57]. In order to mitigate the adverse impact of depth errors in the pseudo ground-truth, we adopt a multi-view check filter [19] to filter out erroneous depth. The knowledge distillation loss is thus defined as

$$\mathcal{L}_{KD} = \sum_{\mathbf{p}} \Phi(\mathbf{p}) \odot \left(\mathbf{D}^t\left(\mathbf{p}\right) - \mathbf{D}_{pseudo}\left(\mathbf{p}\right)\right), \quad (9)$$

where $\Phi(\mathbf{p})$ stands for the multi-view check filter, $\odot$ denotes the element-wise multiplication. The core principle of multi-view check is similar to that of the novel view synthesis in achieving self-supervision. To specify, we first obtain the depth maps for both the target frame $\mathbf{f}^t$ and the source frame $\mathbf{f}^s$ using the depth network. These two frames are then stacked along the channel dimension to acquire an image pair, which is fed into the pose network. The pose network derives the 6-DOF relative pose between the two frames, which is subsequently converted into a transformation matrix $\mathbf{M}^{t\to s}$. Next, we draw on Eq. 5 to get each projected point $\mathbf{p}^{t\to s}$ in the source view and acquire a warped depth map,

$$\mathbf{D}^{s\to t}(\mathbf{p}) = \mathbf{D}^s\left\langle \mathbf{p}^{t\to s}\right\rangle, \quad (10)$$

where $\mathbf{D}^s$ stands for the depth map of source frame. Similar to the projection procedure depicted in Eq. 5, we leverage $\mathbf{D}^{s\to t}$ to perform a reprojection procedure to acquire each reprojected 2D point $\widetilde{\mathbf{p}}^{s\to t}$ and reprojected depth map $\widetilde{\mathbf{D}}^{s\to t}(\mathbf{p})$ in the target view through $\mathbf{K}\mathbf{M}^{s\to t}\mathbf{D}^{s\to t}\mathbf{K}^{-1}\mathbf{p}^s$, where $\mathbf{M}^{s\to t}$ is the inverse matrix of $\mathbf{M}^{t\to s}$. Thereafter, a reprojection error e_{repro} and a geometry error e_{geo} are defined as

$$e_{repro} = \left\|\widetilde{\mathbf{p}}^{s\to t} - \mathbf{p}^t\right\|_2, \quad (11)$$

$$e_{geo} = \frac{\left|\widetilde{\mathbf{D}}^{s\to t}\left(\mathbf{p}\right) - \mathbf{D}^t\left(\mathbf{p}\right)\right|}{\mathbf{D}^t\left(\mathbf{p}\right)}, \quad (12)$$

The determination of valid pixels for multi-view check filter is achieved by

$$\{\mathbf{p}\} = \{\mathbf{p}\,|e_{repro} < a\overline{e}_{repro}, e_{geo} < b\overline{e}_{geo}\}, \quad (13)$$

where $\overline{e}_{repro}$ and $\overline{e}_{geo}$ denote the average over all pixels and a and b are set to 4 based on [19]. The teacher model will be discarded once the training phase is completed.

Contrastive depth reconstruction. As presented in Fig. 2, the denoised depth prediction tends to exhibit depth smearing in boundary regions. To mitigate this issue, we develop a simple yet effective contrastive depth reconstruction mechanism to further enhance the denoising ability of MonoDiffusion by fully exploiting the contextual information, taking inspiration from the masked image modeling [79], [80]. Unlike previous methods that randomly mask the input tokens to train the encoder, we aim to train stronger noise predictors in the decoder. Hence, our paradigm is more like a masked "autodecoder".

For this aim, we generate random masks to remove tokens from the multi-scale feature tokens propagated by the encoder. To prevent information leakage, these masks are shared otherwise the masked information may be easily borrowed from

TABLE I

QUANTITATIVE DEPTH COMPARISON ON THE KITTI DATASET WITH THE EIGEN SPLIT [4]. THE DEFAULT RESIZING FOR ALL INPUT IMAGES IS SET TO 640×192, UNLESS OTHERWISE SPECIFIED. THE BEST AND SECOND BEST RESULTS ARE INDICATED IN **BOLD** AND UNDERLINED, RESPECTIVELY. THE FOLLOWING ABBREVIATIONS ARE USED: "M" STANDS FOR KITTI MONOCULAR VIDEOS, "M+SE" INDICATES MONOCULAR VIDEOS WITH ADDED SEMANTIC SEGMENTATION, "M*" REPRESENTS INPUT RESOLUTION OF 1024×320, AND "M†" DENOTES BACKBONES WITHOUT PRE-TRAINING ON IMAGENET [81]. MONODIFFUSION AND MONODIFFUSION-8M ADOPT THE SAME ENCODER AS LITE-MONO AND LITE-MONO-8M, RESPECTIVELY. ASIDE FROM THE DEPTH ERROR AND ACCURACY, WE REPORT THE MODEL SIZE.

Method	Year	Data	Depth Error (↓)				Depth Accuracy (↑)			Model Size (↓)
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Params.
GeoNet [82]	2018	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975	31.6M
DDVO [83]	2018	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974	28.1M
Monodepth2-Res18 [16]	2019	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981	14.3M
Monodepth2-Res50 [16]	2019	M	0.110	0.831	4.642	0.187	0.883	0.962	0.982	32.5M
SGDepth [23]	2020	M+Se	0.113	0.835	4.693	0.191	0.879	0.961	0.981	16.3M
Johnston *et al.* [20]	2020	M	0.111	0.941	4.817	0.189	0.885	0.961	0.981	14.3M+
CADepth-Res18 [84]	2021	M	0.110	0.812	4.686	0.187	0.882	0.962	0.983	18.8M
HR-Depth [85]	2021	M	0.109	0.792	4.632	0.185	0.884	0.962	0.983	14.7M
Lite-HR-Depth [85]	2021	M	0.116	0.845	4.841	0.190	0.866	0.957	0.982	3.1M
R-MSFM3 [86]	2021	M	0.114	0.815	4.712	0.193	0.876	0.959	0.981	3.5M
R-MSFM6 [86]	2021	M	0.112	0.806	4.704	0.191	0.878	0.960	0.981	3.8M
MonoFormer [87]	2023	M	0.108	0.806	4.594	0.184	0.884	0.963	0.983	23.9M+
SRDepth-Res18 [19]	2023	M	0.111	0.762	4.619	0.186	0.877	0.961	0.983	23.3M
Lite-Mono [24]	2023	M	0.107	0.765	4.561	0.183	0.886	0.963	0.983	3.1M
MonoDiffusion (ours)	2024	M	**0.103**	**0.720**	**4.412**	**0.178**	**0.894**	**0.965**	**0.984**	3.1M
Monodepth2-Res18 [16]	2019	M†	0.132	1.044	5.142	0.210	0.845	0.948	0.977	14.3M
Monodepth2-Res50 [16]	2019	M†	0.131	1.023	5.064	0.206	0.849	0.951	0.979	32.5M
R-MSFM3 [86]	2021	M†	0.128	0.965	5.019	0.207	0.853	0.951	0.977	3.5M
R-MSFM6 [86]	2021	M†	0.126	0.944	4.981	0.204	0.857	0.952	0.978	3.8M
Lite-Mono [24]	2023	M†	0.121	0.876	4.918	0.199	0.859	0.953	0.980	3.1M
MonoDiffusion (ours)	2024	M†	**0.117**	**0.828**	**4.764**	**0.192**	**0.867**	**0.958**	**0.982**	3.1M
Monodepth2-Res18 [16]	2019	M*	0.115	0.882	4.701	0.190	0.879	0.961	0.982	14.3M
R-MSFM3 [86]	2021	M*	0.112	0.773	4.581	0.189	0.879	0.960	0.982	3.5M
R-MSFM6 [86]	2021	M*	0.108	0.748	4.470	0.185	0.889	0.963	0.982	3.8M
HR-Depth [85]	2021	M*	0.106	0.755	4.472	0.181	0.892	0.966	**0.984**	14.7M
SRDepth-Res18 [19]	2023	M*	0.106	0.673	4.379	0.180	0.886	0.965	**0.984**	23.3M
Lite-Mono [24]	2023	M*	0.102	0.746	4.444	0.179	0.896	0.965	0.983	3.1M
Lite-Mono-8M [24]	2023	M*	0.097	0.710	4.309	0.174	0.905	0.967	**0.984**	8.7M
MonoDiffusion (ours)	2024	M*	0.098	0.690	4.289	0.174	0.904	0.967	**0.984**	3.1M
MonoDiffusion-8M (ours)	2024	M*	**0.095**	**0.670**	**4.219**	**0.171**	**0.909**	**0.968**	**0.984**	8.8M
MonoViT-tiny [88]	2022	M	0.102	0.733	4.459	0.177	0.895	0.965	**0.984**	10.3M
Lite-Mono-8M [24]	2023	M	0.101	0.729	4.454	0.178	0.897	0.965	0.983	8.7M
MonoDiffusion-8M (ours)	2024	M	**0.099**	**0.702**	**4.385**	**0.176**	**0.899**	**0.966**	**0.984**	8.8M

feature tokens of different resolutions. In practice, we generate a mask with the highest resolution, and leverage nearest neighbor interpolation to acquire a pyramid of masks. These masked feature tokens are further aggregated into masked conditions through learnable layers composed of 3×3 convolutions and upsampling layers. We make use of the masked conditions to reconstruct the denoised depth map guided by the complete conditions, instead of the RGB image in [79], [80].

The reconstruction loss is defined as

$$\mathcal{L}_{rec} = \sum_{\mathbf{p}} \left| \widehat{\mathbf{D}}^{t}(\mathbf{p}) - \mathbf{D}^{t}(\mathbf{p}) \right|, \tag{14}$$

where $\widehat{\mathbf{D}}^{t}$ denotes the denoised depth prediction using masked conditions.

C. Inference

When provided a test image as conditional input, the model begins with a random noise map sampled from the Gaussian distribution and refines its predictions iteratively in a coarse-to-fine manner. We select the DDIM update rule [26] for the sampling. At each sampling step, the fused noise embeddings, time embeddings and visual conditions serve as input to the noise predictors for noise predictions. Then, we make use of the reparameterization trick to calculate the noise maps for the next step. The denoising step will be repeated until reaching the final step.

Due to the stochastic nature of the noise map initialization, regions with high uncertainty in the depth map exhibit diverse outcomes with different noise map initializations. This variability enables us to estimate the uncertainty corresponding to the depth map by employing multiple noise map initializations and calculating the variance. Nonetheless, we observe that the distinction in the initialization of different noise maps is not readily apparent in the numerical results due to the averaging operation, so we opt to employ a single noise map initialization when presenting numerical results, unless otherwise specified.

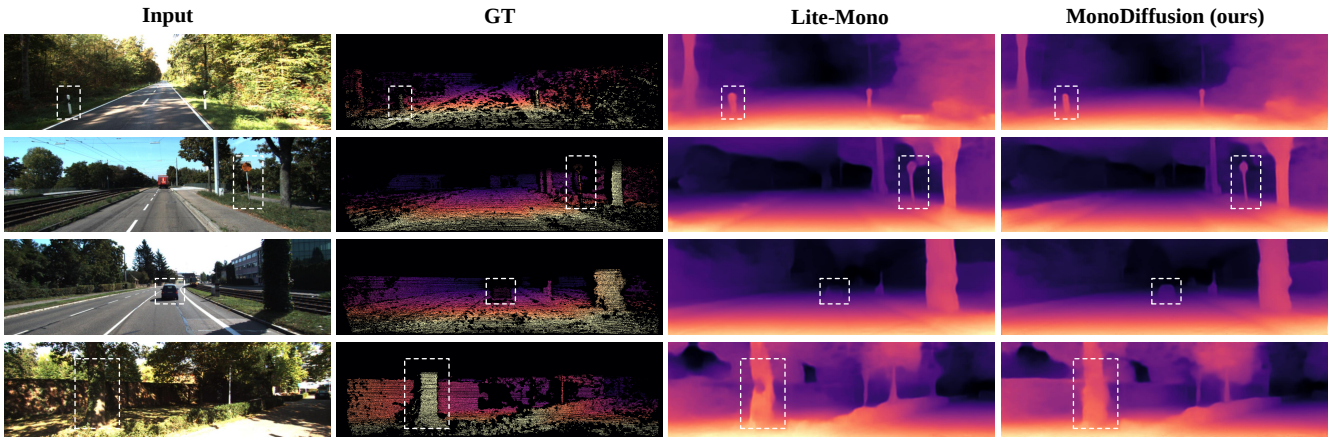

Fig. 4. **Qualitative depth comparison on the KITTI dataset.** The white boxes indicate the regions to emphasize. GT stands for the corresponding ground-truth depth maps. MonoDiffusion predicts depth more accurately than Lite-Mono, preserving object shapes and handling complex cases like shadows more robustly.

TABLE II
QUANTITATIVE DEPTH COMPARISON ON THE KITTI EIGEN SPLIT, WITH IMPROVED GROUND-TRUTH FROM [89].

Method	Year	Data	Depth Error (↓)				Depth Accuracy (↑)			Model Size (↓)
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Params.
GeoNet [82]	2018	M	0.132	0.994	5.240	0.193	0.883	0.953	0.985	31.6M
DDVO [83]	2018	M	0.126	0.866	4.932	0.185	0.851	0.958	0.986	28.1M
Ranjan [90]	2019	M	0.123	0.881	4.834	0.181	0.860	0.959	0.985	-
EPC++ [91]	2019	M	0.120	0.789	4.755	0.177	0.856	0.961	0.987	-
Monodepth2-Res18 [16]	2019	M	0.090	0.545	3.942	0.137	0.914	0.983	0.995	14.3M
Johnston *et al.* [20]	2020	M	0.081	0.484	3.716	0.126	0.927	0.985	0.996	14.3M+
R-MSFM6 [86]	2021	M	0.082	0.432	3.658	0.127	0.924	0.985	0.996	3.8M
Lite-Mono [24]	2023	M	0.077	0.405	3.514	0.120	0.931	0.987	**0.997**	3.1M
MonoDiffusion (ours)	2024	M	**0.073**	**0.377**	**3.451**	**0.115**	**0.935**	**0.988**	**0.997**	3.1M

D. Network Architecture and Overall Loss

Both **depth networks** adopt the prevalent encoder-decoder architecture. The teacher model is Lite-Mono-8M [24]. On the other hand, the student model uses two encoders from Lite-Mono family in different settings, which are the same as Lite-Mono and Lite-Mono-8M, respectively. Its decoder is adapted from [24] and multi-scale feature tokens from the encoder are aggregated into image conditions via 3×3 convolutional layers and upsampling layers. We make use of three noise predictors corresponding to coarse-to-fine diffusion-denoising stages.

The **pose network** design is same to prior works [16], [24], where a pre-trained ResNet18 [53] is employed as the encoder and four convolutional layers with global average pooling are used as the decoder. It receives a stacked pair of images and estimates a 6-DOF relative pose between them.

Diffusion denoising loss. In line with previous work [26], we employ a diffusion denoising loss to supervise the predicted noise,

$$\mathcal{L}_{dd} = \sum_{\mathbf{p}} \left\| \epsilon - \epsilon_\theta \left(\mathbf{D}_\tau, \tau, \mathbf{c} \right) \right\|^2. \tag{15}$$

Edge-aware smoothness loss. As defined in [16], we apply an edge-aware smoothness loss to encourage the smoothness property of depth map,

$$\mathcal{L}_{es} = \sum_{\mathbf{p}} \left| \nabla \mathbf{D}^t \left(\mathbf{p} \right) \right| \odot \exp^{-\left\| \nabla \mathbf{f}^t \left(\mathbf{p} \right) \right\|_1}, \tag{16}$$

where $\nabla \mathbf{D}^t \left(\mathbf{p} \right)$ and $\nabla \mathbf{f}^t \left(\mathbf{p} \right)$ calculates the first-order gradients of depth map and RGB image, respectively.

Overall loss. The full optimization objective is summarized as

$$\mathcal{L}_{overall} = \lambda_1 \mathcal{L}_{ph} + \lambda_2 \mathcal{L}_{KD} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{dd} + \lambda_5 \mathcal{L}_{es}, \tag{17}$$

where λ_1, λ_2, λ_3 and λ_4 are empirically set to 1, 1, 0.1, 1, and 1e-3, respectively. Similar to [16], we employ two source frames in $\mathcal{L}_{ph}$ and the one with the minimum $\mathcal{L}_{ph}$ is chosen to handle occlusion. Moreover, an auto-masking mechanism [16] is used to deal with dynamic objects. Because our coarse-to-fine diffusion process enables the acquisition of depth maps at three different scales, the above loss functions are concurrently applied to these three depth maps. For simplicity, we omit this detail.

V. EXPERIMENT

A. Datasets and Evaluation Metrics

The **KITTI** dataset is composed of 61 stereo road scenarios with the image resolution around 1241×376 pixels. The data collection involves multiple sensors such as cameras, 3D Lidar, GPU/IMU, among others. We adopt the Eigen split [4], which comprises 39180 monocular triplets for training, 4424 images for validation, and 697 images for testing. As in [16], we adopt

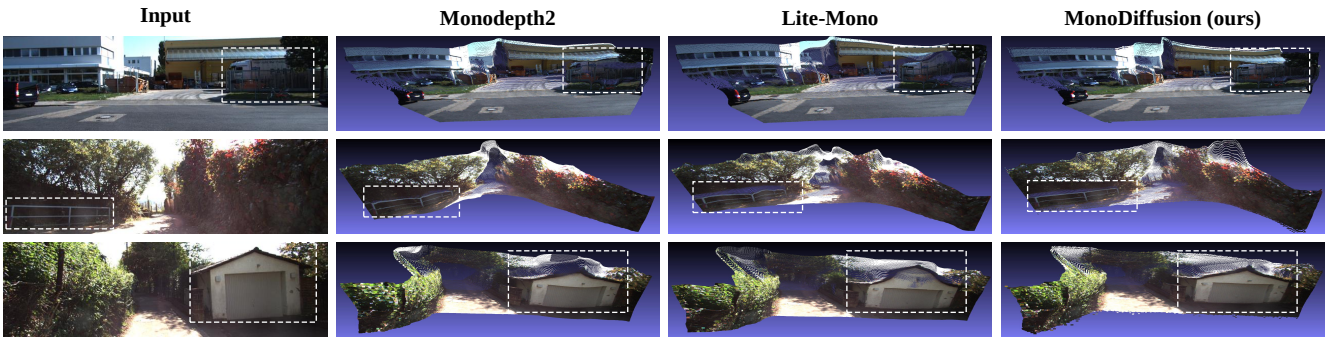

Fig. 5. **Qualitative point cloud results on the KITTI dataset.** The white boxes highlight the regions to focus on. MonoDiffusion reduces distortion and preserves geometry better than Monodepth2 and Lite-Mono, showing the strong generative ability of the diffusion model.

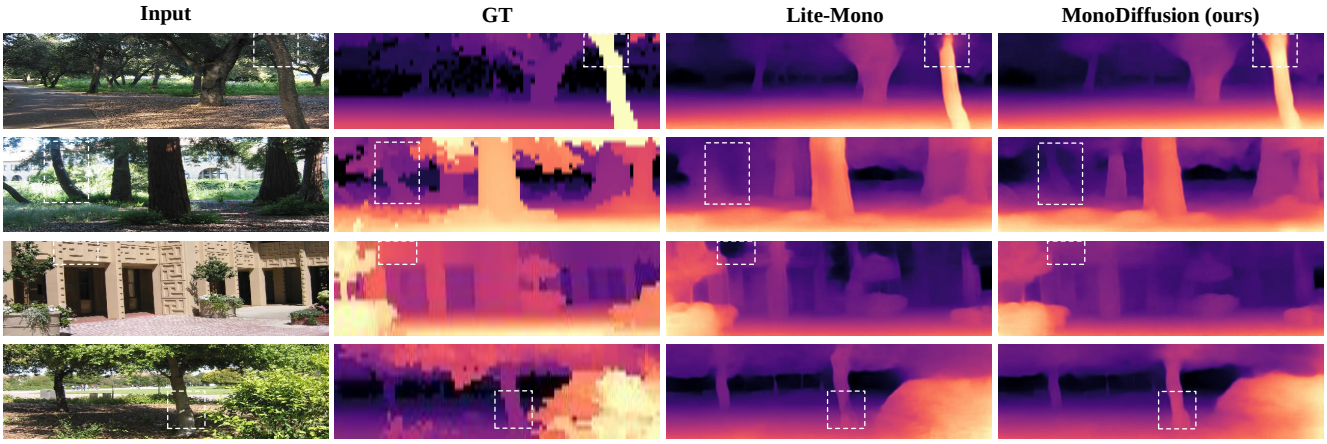

Fig. 6. **Qualitative depth comparison on the Make3D dataset**. The white boxes indicate the regions of interest, and GT refers to the ground-truth depth maps. MonoDiffusion produces more continuous depth maps and accurately separates the outlines of closely positioned trees.

the same intrinsic for all images by averaging the intrinsic of each image. During the evaluation phase, the predicted depth is constrained to the common practice range of $[0, 80]$m.

The **Make3D** dataset consists of 134 test images collected from outdoor scenes and the **DIML** dataset involves 500 test images captured from outdoor scenes. We only leverage these two datasets for a zero-shot generalization study by utilizing models trained on the KITTI dataset.

B. Evaluation Metrics

Similar to [24], we employ the following evaluation metrics in our experiments,

- Abs Rel: $\frac{1}{\|\mathbf{D}\|_1}\sum_{d\in\mathbf{D}}|d_t-d|/d$;
- Sq Rel: $\frac{1}{\|\mathbf{D}\|_1}\sum_{d\in\mathbf{D}}\|d_t-d\|^2/d$;
- RMSE: $\sqrt{\frac{1}{\|\mathbf{D}\|_1}\sum_{d\in\mathbf{D}}\|d_t-d\|^2}$;
- RMSE log: $\sqrt{\frac{1}{\|\mathbf{D}\|_1}\sum_{d\in\mathbf{D}}\|\log d_t-\log d\|^2}$;
- $\delta < thr$: % of d satisfies $\left(\max\left(\frac{d_t}{d},\frac{d}{d_t}\right)=\delta<thr\right)$ for $thr=1.25,1.25^2,1.25^3$.

Abs Rel, Sq Rel, RMSE and RMSE log are depth error metrics and the lower the better. $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$ are depth accuracy metrics and the higher the better. Because the self-supervised MDE system has inherent scale ambiguity, we make use of the median scaling technique introduced by [5] to recover the absolute scale for depth evaluation.

TABLE III
QUANTITATIVE COMPARISON ON THE MAKE3D DATASET. ALL MODELS ARE TRAINED ON KITTI WITH AN IMAGE RESOLUTION OF 640×192.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log↓
Monodepth [92]	0.544	10.940	11.760	0.193
SfMLearner [5]	0.383	5.321	10.470	0.478
DDVO [83]	0.387	4.720	8.090	0.204
Monodepth2 [16]	0.322	3.589	7.417	0.163
R-MSFM6 [86]	0.334	3.285	7.212	0.169
Lite-Mono [24]	0.305	3.060	6.981	0.158
MonoDiffusion	**0.297**	**2.871**	**6.877**	**0.156**

C. Implementation Details

MonoDiffusion is implemented in the PyTorch library and trained on a single NVIDIA TITAN RTX. We use the AdamW optimizer [93] where the weight decay is 1e-2 and the batch size is set to 12. When training models from scratch, an initial learning rate of 5e-4 is adopted along with a cosine learning rate schedule [94]. The training process runs a total number of 35 epochs in this scenario. It is observed that pre-training on the ImageNet [81] accelerates network convergence, resulting

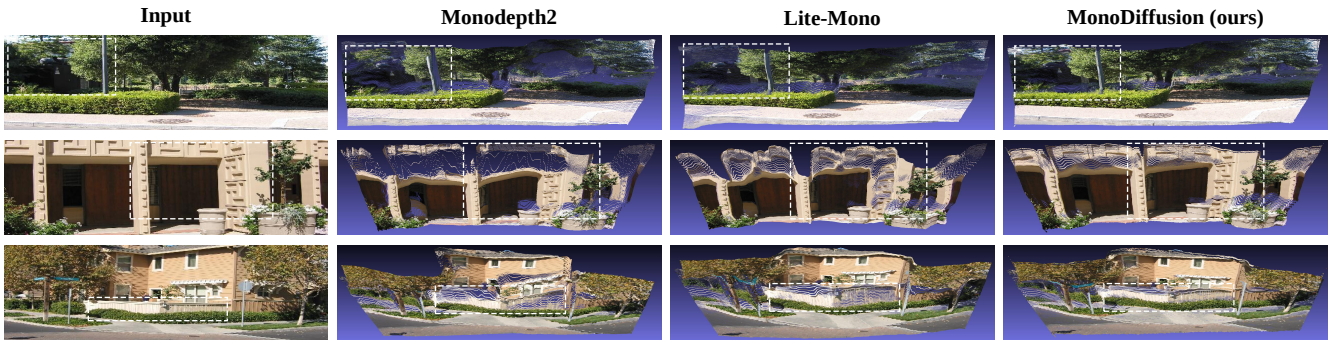

Fig. 7. **Qualitative point cloud results on the Make3D dataset.** The white boxes mark the regions of interest. The compared methods have difficulty with structures like dense foliage, flat walls, and thin railings, while MonoDiffusion demonstrates better performance in reconstructing these features.

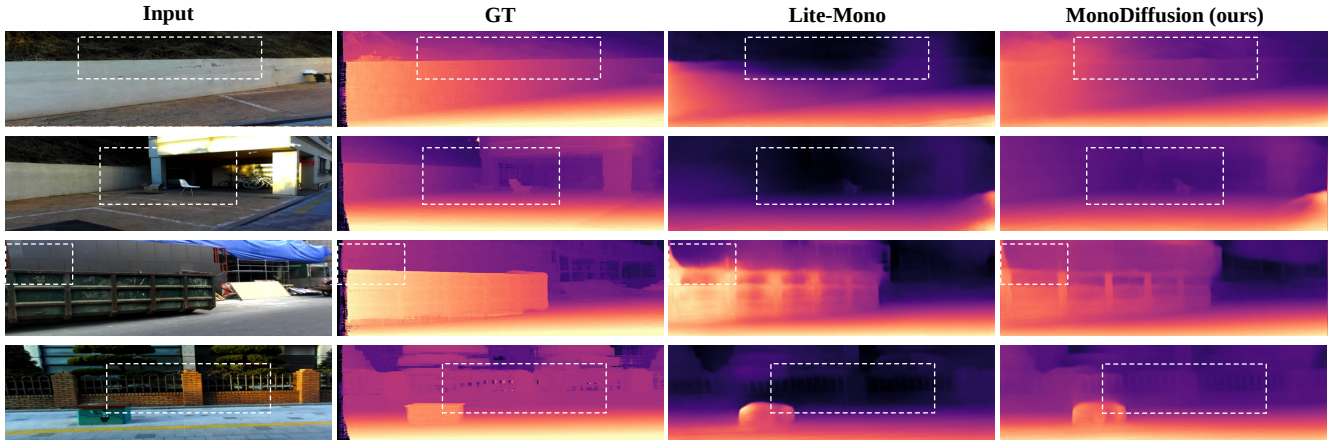

Fig. 8. **Qualitative depth comparison on the DIML dataset**. The white boxes show the regions of interest, and GT represents the ground-truth depth maps. Lite-Mono tends to have black holes in the depth maps, while MonoDiffusion shows better generalization.

in a shorter training time of 30 epochs when loading the pre-trained weights. Moreover, the initial learning rate is adjusted to 1e-4. Following [24], [87], we leverage random horizontal flip as well as random brightness adjustment, saturation adjustment, contrast adjustment and hue jitter with a 50% chance.

D. Comparison to Prior State-of-the-Arts

We compare the proposed MonoDiffusion with prior state-of-the-art competitors on the Eigen split of KITTI benchmark. We mainly focus on training with monocular videos in three settings: 640×192 resolution, without pre-training on ImageNet and 1024×320 resolution. The results are summarized in Table I. As we can see, MonoDiffusion is able to exceed the compared methods in each setting and beats recent Lite-Mono with almost identical number of parameters. Moreover, even compared to semantic segmentation-assisted methods such as SGDepth, MonoDiffusion shows great advantages. In Table II, we show the results on the Eigen split with improved ground-truth. Once again, MonoDiffusion achieves better performance than the compared methods in most metrics.

Fig. 4 shows a qualitative depth comparison. For the depth map predicted by Lite-Mono, the warning pole appears thicker than it actually is, whereas MonoDiffusion maintains its original shape more accurately. Besides, MonoDiffusion is better at predicting the traffic sign and distant car. In a more complex case, in which tree trunk is partially obscured by leaf shadows, the depth map predicted by Lite-Mono exhibits discontinuity, while MonoDiffusion is more robust to such a case. To further show the strengths of MonoDiffusion, we convert depth maps into point clouds and demonstrate the 3D structures in Fig. 5. MonoDiffusion preserves key geometric features and achieves less distortion than the compared methods. For example, the railings in the middle row appear curved in the results of Monodepth2 and Lite-Mono, whereas MonoDiffusion sustains the straight geometric properties. The outstanding performance of

TABLE IV
QUANTITATIVE COMPARISON ON THE DIML DATASET. ALL MODELS ARE TRAINED ON THE KITTI DATASET USING IMAGES WITH A RESOLUTION OF 640×192.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log↓
SfMLearner [5]	0.235	0.427	1.477	0.330
DDVO [83]	0.262	0.492	1.669	0.364
Monodepth2 [16]	0.185	0.298	1.140	0.249
HRDepth [85]	0.183	0.296	1.128	0.248
R-MSFM6 [86]	0.181	0.301	1.132	0.243
Lite-Mono [24]	0.173	0.271	1.108	0.239
MonoDiffusion	**0.166**	**0.256**	**1.084**	**0.232**

TABLE V
ROBUSTNESS SUTDY ON THE CORRUPTED IMAGES OF KITTI DATASET. THE RESULTS OF LITE-MONO [24] ARE ACQUIRED USING PUBLICLY AVAILABLE PRE-TRAINED WEIGHTS.

Corruption Type		Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Clean		Lite-Mono [24]	0.107	0.765	4.561	0.183	0.886	0.963	0.983
		MonoDiffusion	**0.103**	**0.720**	**4.412**	**0.178**	**0.894**	**0.965**	**0.984**
Noise	Speckle Noise	Lite-Mono [24]	0.119	0.838	5.017	0195	0.859	0.956	0.982
		MonoDiffusion	**0.110**	**0.740**	**4.660**	**0.184**	**0.876**	**0.962**	**0.984**
	Shot Noise	Lite-Mono [24]	0.139	1.006	5.613	0.218	0.813	0.940	0.978
		MonoDiffusion	**0.132**	**0.906**	**5.254**	**0.209**	**0.827**	**0.946**	**0.981**
Blur	Motion Blur	Lite-Mono [24]	0.123	0.938	5.220	0.204	0.852	0.950	0.979
		MonoDiffusion	**0.116**	**0.839**	**5.040**	**0.197**	**0.862**	**0.954**	**0.981**
	Zoom Blur	Lite-Mono [24]	0.171	1.435	6.259	0.259	0.759	0.912	0.964
		MonoDiffusion	**0.162**	**1.241**	**5.864**	**0.243**	**0.775**	**0.922**	**0.970**
Digital	Contrast	Lite-Mono [24]	0.128	1.026	5.206	0.206	0.847	0.950	0.979
		MonoDiffusion	**0.118**	**0.890**	**4.990**	**0.196**	**0.862**	**0.956**	**0.981**
	Brightness	Lite-Mono [24]	0.117	0.895	4.936	0.195	0.867	0.956	0.981
		MonoDiffusion	**0.112**	**0.823**	**4.747**	**0.187**	**0.876**	**0.960**	**0.982**
Weather	Fog	Lite-Mono [24]	0.148	1.165	5.565	0.225	0.806	0.935	0.975
		MonoDiffusion	**0.136**	**1.009**	**5.326**	**0.212**	**0.827**	**0.946**	**0.979**
	Snow	Lite-Mono [24]	0.136	1.056	5.501	0.220	0.822	0.941	0.975
		MonoDiffusion	**0.130**	**0.963**	**5.239**	**0.211**	**0.835**	**0.946**	**0.978**

our MonoDiffusion evidences the strong generative capability of diffusion model.

E. Zero-shot Generalization

In Tables III and IV, we validate the generalization ability of MonoDiffusion in a fully zero-shot setting. The models are trained on the KITTI dataset but are evaluated on the Make3D and DIML datasets. The superior results of our MonoDiffusion indicate that it generalizes well to unseen scenarios. In Fig. 6, we demonstrate a qualitative depth comparison on the Make3D dataset. The depth maps of MonoDiffusion demonstrate better continuity, for instance, in the regions of tree truck and wall. Besides, we observe that in the second row, the outlines of the two closely positioned trees are merged together in the depth map from Lite-Mono, whereas MonoDiffusion accurately distinguishes them. In Fig. 7, we further demonstrate a qualitative point cloud comparison on the Make3D dataset. The compared methods struggle with the structures like lush foliage, flat walls and thin railings. By contrast, MonoDiffusion is able to recover these structures reasonably. In Fig. 8, we present a qualitative depth comparison on the DIML dataset. It can be seen that Lite-Mono tends to exhibit black holes in the predicted depth maps, which may result from the significant distribution gap between the KITTI training dataset and the DIML test set. On the other hand, our MonoDiffusion shows better generalization ability.

F. Robustness Study

Model robustness in depth estimation is important, as real-world images are often susceptible to various types of corruption. In such scenarios, it is significant to develop a model that remains resilient. In order to evaluate the robustness of the proposed MonoDiffusion, we follow prior methods [95], [96]

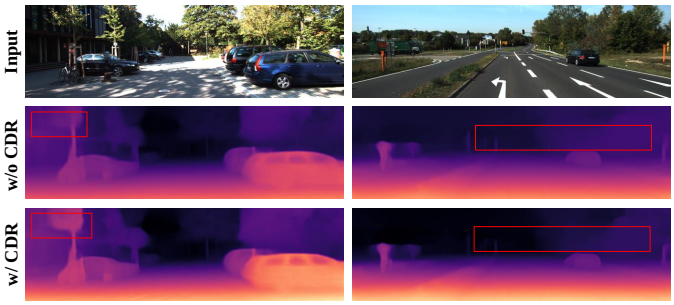

Fig. 9. **Qualitative depth comparison of MonoDiffusion with and without the contrastive depth reconstruction (CDR) mechanism.** The red boxes stand for the highlighted regions. The CDR mechanism alleviates the depth smearing, which improves tree silhouettes and allows for better differentiation between the depths of distant objects, such as the sky, and trees.

and test our model on the corrupted images. To be specific, we append the corruptions including speckle noise, shot noise, motion blur, zoom blur, contrast, brightness, fog and snow to the images, respectively. Table V displays the depth estimation results for the corrupted images of the KITTI test set. We can see that the proposed method achieves higher robustness to all types of corruption than Lite-Mono, making it more suitable for the challenging scenarios.

G. Ablation Study

To better understand the influence of different components in MonoDiffusion on performance, we provide detailed ablation results.

Whole framework (Table VI). We use the Lite-Mono [24] as our baseline (ID 1). The implementation of forward diffusion is hard for self-supervised MDE due to the lack of depth ground-truth. Accordingly, we develop a coarse-to-fine pseudo ground-truth diffusion. The results verify the applicability and

TABLE VI
ABLATION STUDY ON THE WHOLE FRAMEWORK. CFD: COARSE-TO-FINE PSEUDO GROUND-TRUTH DIFFUSION; CDR: CONTRASTIVE DEPTH RECONSTRUCTION; KD: KNOWLEDGE DISTILLATION.

ID	CFD	CDR	KD	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
1				0.107	0.787	4.580	0.182	0.887	0.963	0.983
2	✓			0.105	0.747	4.500	0.180	0.890	0.964	**0.984**
3	✓	✓		0.104	0.733	4.446	0.179	0.893	0.964	**0.984**
4	✓	✓	✓	**0.103**	**0.720**	**4.412**	**0.178**	**0.894**	**0.965**	**0.984**

TABLE VII
EFFECT OF DIFFERENT TEACHER MODELS ON THE MONODIFFUSION. WE USE THE LITE-MONO-8M, LITE-MONO AND LITE-MONO-SMALL [24] AS TEACHERS IN THE TRAINING PHASE, RESPECTIVELY. AMONG THESE THREE TEACHERS, LITE-MONO HAS THE SAME ENCODER AND IDENTICAL NUMBER OF PARAMETERS AS MONODIFFUSION. 8.7 M, 3.1 M AND 2.5 M ARE THE NUMBER OF MODEL PARAMETERS.

Method	Teacher	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Lite-Mono-8M (8.7 M)	-	**0.101**	0.729	4.454	**0.178**	**0.897**	**0.965**	0.983
Lite-Mono (3.1 M)	-	0.107	0.765	4.561	0.183	0.886	0.963	0.983
Lite-Mono-small (2.5 M)	-	0.110	0.802	4.671	0.186	0.879	0.961	0.982
MonoDiffusion (3.1 M)	Lite-Mono-8M	0.103	**0.720**	**4.412**	**0.178**	0.894	**0.965**	**0.984**
MonoDiffusion (3.1 M)	Lite-Mono	0.104	0.737	4.453	**0.178**	0.891	0.964	**0.984**
MonoDiffusion (3.1 M)	Lite-Mono-small	0.105	0.741	4.527	0.180	0.888	0.964	**0.984**

TABLE VIII
COMPARISON OF COMPUTATIONAL EFFICIENCY. INF.: INFERENCE TIME; FHD: FIXED HIGH-RESOLUTION PSEUDO GROUND-TRUTH DIFFUSION; CFD: COARSE-TO-FINE PSEUDO GROUND-TRUTH DIFFUSION. 20, 15 AND 10 DENOTE THE NUMBER OF TOTAL DENOISING STEPS FOR FHD. TO ACHIEVE SATISFACTORY RESULTS, THE FHD IS PERFORMED AT THE HIGHEST RESOLUTION. 5, 4 AND 3 STAND FOR THE NUMBER OF TOTAL DENOISING STEPS FROM COARSE TO FINE.

Method	Inf. (s) ↓	FLOPs (G) ↓	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
MonoDiffusion (w/ FHD, 20)	0.041	11.40	**0.103**	0.726	4.447	0.179
MonoDiffusion (w/ FHD, 15)	0.037	9.81	0.104	0.739	4.514	0.180
MonoDiffusion (w/ FHD, 10)	0.034	8.22	0.106	0.775	4.547	0.181
MonoDiffusion (w/ CFD, (5, 4, 3))	**0.028**	**6.57**	**0.103**	**0.720**	**4.412**	**0.178**
Lite-Mono-8M [24]	0.039	11.21	**0.101**	0.729	4.454	**0.178**
MonoDiffusion (w/ CFD, (5, 4, 3))	**0.028**	**6.57**	0.103	**0.720**	**4.412**	**0.178**

effectiveness of the diffusion models for self-supervised MDE (ID 2). We further integrate the contrastive depth reconstruction mechanism and achieve consistent improvements in most metrics (ID 3). Finally, we add the knowledge distillation in the training phase and obtain the best results (ID 4).

Different teacher models (Table VII). Because our training relies on the pseudo depth ground-truth derived by the teacher model for forward diffusion, we further investigate the impact of different teacher models on the MonoDiffusion. As can be seen, different teachers do not have a particularly large impact on performance. It is noteworthy that despite employing Lite-Mono or Lite-Mono-small as the teacher model, MonoDiffusion consistently achieves superior results compared to Lite-Mono. Moreover, MonoDiffusion even exceeds Lite-Mono-8M in certain metrics when the latter serves as the teacher model.

Computation efficiency (Table VIII). Diffusion-based methods are usually computationally demanding due to the traditional fixed high-resolution diffusion. We conduct a comparison between the proposed coarse-to-fine diffusion and the

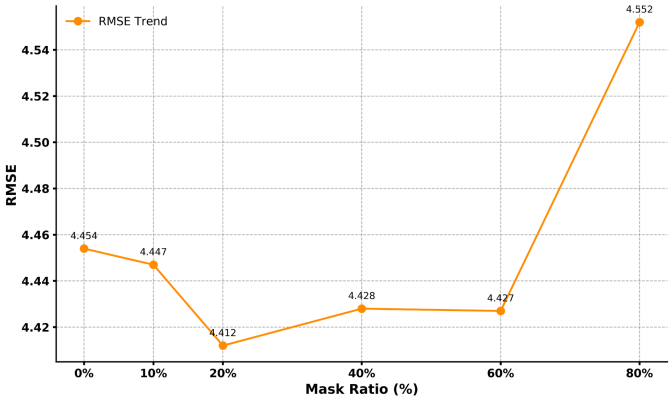

Fig. 10. **Effect of varying mask ratios in the contrastive depth reconstruction mechanism, which is utilized only during the training phase**. A 20% mask training ratio is optimal, while the model still produces acceptable results with an 80% ratio.

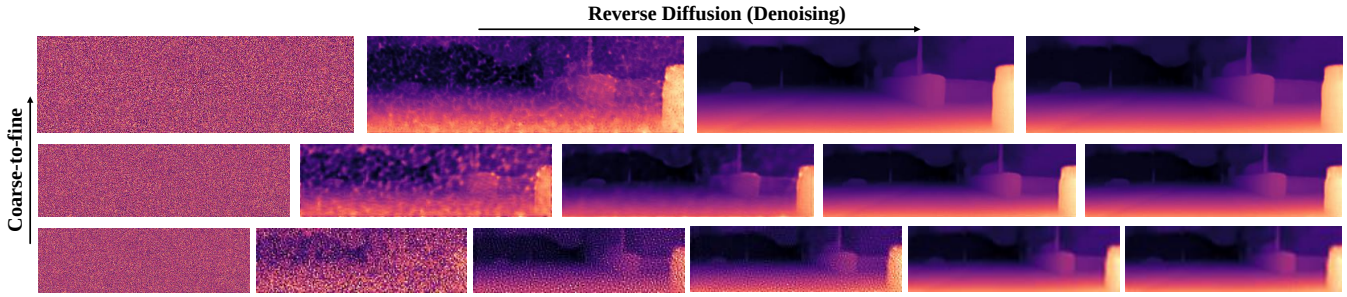

Fig. 11. **Visualization of the coarse-to-fine denoising process in MonoDiffusion.** The process involves three stages. Each stage begins with the initialization of object shapes and edges, followed by the progressive refinement of depth values and correction of distance relationships.

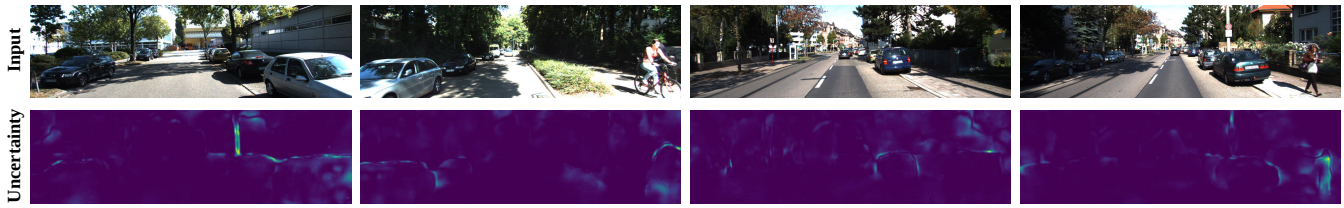

Fig. 12. **Visualization of the uncertainty maps generated by the proposed MonoDiffusion.** The uncertainty maps identify potential errors and emphasize hard areas, such as car edges and thin trunks.

fixed high-resolution diffusion in terms of their inference time and FLOPs. When the number of total denoising steps decreases, the performance of the fixed high-resolution diffusion-based models gradually declines, albeit with an increase in computation efficiency. In contrast, the model based on coarse-to-fine diffusion not only has better computation efficiency but also delivers better performance. In addition, we list the results of discriminative-based Lite-Mono-8M as a comparison. The proposed MonoDiffusion not only demonstrates strengths in inference time and FLOPs but also exceeds it in key metrics such as Sq Rel and RMSE.

Contrastive depth reconstruction. In Fig. 9, we present a qualitative depth comparison between MonoDiffusion with and without the contrastive depth reconstruction (CDR) mechanism. It can be seen that the CDR mechanism indeed alleviates the depth smearing, *e.g.*, better tree silhouettes. Besides, the object depth at distant distances can be better recovered, for instance, the sky. In the absence of CDR mechanism, the depth of sky is similar to the depth of distant trees, but in fact their depth should be highly different.

In Fig. 10, we further explore the impact of different mask ratios in the CDR mechanism. We notice that 20% is the best choice, and the model is capable of achieving feasible results even with a mask training ratio of 80%.

Denoising inference. In Fig. 11, we demonstrate an intuitive illustration showing the iterative refinement process of depth from a random noise map. From left to right, these three stages begin by initializing the shapes and boundaries of objects, then progressively refine the depth values and rectify the inter- and intra-object distance relationships. Besides, we notice that the denoising process speeds up significantly at the finest stage. We attribute this to the geometric conditions from the previous stage, which allows for satisfactory results utilizing only three denoising steps at the highest resolution.

In Fig. 12, we visualize the uncertainty maps generated by MonoDiffusion through three random noise map initializations and calculating the variance. It can be seen that the uncertainty maps exhibit potential errors and highlight difficult regions, for example, car edges and skinny trunks, enabling safety-critical applications.

VI. Conclusion

In this paper, we investigate the utility and efficacy of diffusion models for self-supervised monocular depth estimation and introduce a novel generative framework by formulating it as a conditional denoising process. In addition, we develop a coarse-to-fine pseudo ground-truth diffusion process to assist the diffusion and a contrastive depth reconstruction mechanism to strengthen the denoising ability of model. Extensive experiments on the KITTI, Make3D and DIML datasets validate the efficacy of MonoDiffusion. We believe our new initiative will encourage more sophisticated diffusion-based self-supervised depth estimation achievements.

References

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

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

[3] Oskar Natan and Jun Miura. End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. *IEEE Transactions on Intelligent Vehicles*, 2022.

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[5] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[6] Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21425–21434, 2023.

[7] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[9] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing*, 27(8):4131–4144, 2018.

[10] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5684–5693, 2019.

[11] Minsoo Song, Seokjae Lim, and Wonjun Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4381–4393, 2021.

[12] Xinchen Ye, Xin Fan, Mingliang Zhang, Rui Xu, and Wei Zhong. Unsupervised monocular depth estimation via recursive stereo distillation. *IEEE Transactions on Image Processing*, 30:4492–4504, 2021.

[13] Fangzheng Tian, Yongbin Gao, Zhijun Fang, Yuming Fang, Jia Gu, Hamido Fujita, and Jenq-Neng 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*, 32(4):1751–1766, 2021.

[14] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15560–15569, 2021.

[15] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision*, pages 36–53, 2018.

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

[17] Jia-Wang 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. In *Advances in Neural Information Processing Systems*, 2019.

[18] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017.

[19] Zhong Liu, Ran Li, Shuwei Shao, Xingming Wu, and Weihai Chen. Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7565–7577, 2023.

[20] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4756–4765, 2020.

[21] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14402–14413, 2020.

[22] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12642–12652, 2021.

[23] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020.

[24] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

[27] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[32] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. Midms: Matching interleaved diffusion models for exemplar-based image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2191–2199, 2023.

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[36] Xinxu Ge, Xin Liu, Zitong Yu, Jingang Shi, Chun Qi, Jie Li, and Heikki Kälviäinen. Difffas: Face anti-spoofing via generative diffusion models. *European Conference on Computer Vision*, 2024.

[37] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching. In *International Conference on Learning Representations*, 2023.

[38] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[39] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[40] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15977–15987, 2023.

[41] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[42] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 909–919, 2023.

[43] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

[44] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[45] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.

[46] Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023.

[47] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[48] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024.

[49] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024.

[50] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[51] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2008.

[52] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision*, pages 239–248. IEEE, 2016.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[54] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

[55] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[56] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[57] Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. *IEEE Transactions on Multimedia*, 2023.

[58] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *International Conference on Learning Representations*, 2023.

[59] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7931–7940, 2023.

[60] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[61] Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12737–12746, 2021.

[62] Michaël Ramamonjisoa, Michael Firman, Jamie Watson, Vincent Lepetit, and Daniyar Turmukhambetov. Single image depth prediction with wavelet decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11089–11098, 2021.

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

[64] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations*, 2019.

[65] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[66] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[67] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[68] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[69] Dohoon Ryu and Jong Chul Ye. Pyramidal denoising diffusion probabilistic models. *arXiv preprint arXiv:2208.01864*, 2022.

[70] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[71] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4186, 2022.

[72] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations*, 2022.

[73] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[74] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[75] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[76] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[77] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020.

[78] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[79] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[80] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16284–16294, 2023.

[81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[82] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

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

[84] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *International Conference on 3D Vision*, pages 464–473. IEEE, 2021.

[85] 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 *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 35, pages 2294–2301, 2021.

[86] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12777–12786, 2021.

[87] Jinwoo Bae, Sungho Moon, and Sunghoon Im. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 187–196, 2023.

[88] 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 *International Conference on 3D Vision*, pages 668–678. IEEE, 2022.

[89] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[90] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.

[91] C Luo, Z Yang, P Wang, Y Wang, W Xu, R Nevatia, and A Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

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

[93] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2018.

[94] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

[95] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[96] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022.

Shuwei Shao received the B.Eng. degree from Xidian University, Xi'an, Shanxi, China, in 2019. He is currently pursuing the Ph.D. degree in the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His current research interests include depth perception, image registration and 3D reconstruction.

Zhongcai Pei received the B.Eng., M.Eng. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1991, 1994, and 1997, respectively. He has been with the School of Automation Science and Electronic Engineering, Beihang University, as an Associate Professor from 2000 and as a Professor since 2006. He has published over 50 technical papers in referred journals and conference proceedings and filed more than 20 patents. His research interests include bionic CPG mechanisms and control, computer vision, parallel mechanism and robotics.

Weihai Chen received the B.Eng. degree from Zhejiang University, China, in 1982, and the M.Eng. and Ph.D. degrees from Beihang University, Beijing, China, in 1988 and 1996, respectively. He is currently a Professor with the School of Automation Science and Electrical Engineering, Beihang University. He has published more than 200 technical articles in referred journals and conference proceedings and filed more than 30 patents. His research interests include computer vision, image processing, automation, bioinspired robotics, and precision mechanisms. He is also a member of the Technical Committee on Manufacturing Automation of the IEEE Robotics and Automation Society. He was a recipient of the 2017 Best Paper Award for the IEEE Transactions on Industrial Electronics and the 2018 IET Premium Award for the *IET Intelligent Transport Systems*.

Dingchi Sun received the B.Eng. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China, in 2023. He is currently pursuing the M.Eng. degree in the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His current research interests include depth perception.

Peter C. Y. Chen received the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 1995. During 1985–1997, he worked on various projects for China International Marine Containers Ltd., China, Automation Tooling Systems, Inc., Canada, General Dynamics Corp., Electronics Division, USA, Philips Pte. Ltd., Singapore, the University of Toronto, CRS Robotics Inc., Canada, and Canadian Space Agency, Canada. From September 1997 to March 2000, he was a Research Fellow with the Singapore Institute of Manufacturing Technology, Singapore, where he developed modular reconfigurable robotic systems for manufacturing applications. He is currently with the Department of Mechanical Engineering, National University of Singapore, Singapore. His research interests include robotics and formal methods for reinforcement learning.

Zhengguo Li (F'23) received the B.Sci (Applied Mathematics). and M.Eng. (Automatic Control) from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and the Ph.D. degree (Automatic Control) from Nanyang Technological University, Singapore, in 2001.

His research interests include video processing and delivery, computational photography, switched and impulsive control, sensor fusion and physics-driven deep learning. He has co-authored one monograph, more than 200 journal/conference papers including more than 60 IEEE Transaction papers, and eleven granted USA patents, including normative technologies on scalable extension of H.264/AVC and HEVC. He has been actively involved in the development of H.264/AVC and HEVC since 2002. He had three informative proposals adopted by the H.264/AVC and three normative proposals adopted by the HEVC. Currently, he is with the Agency for Science, Technology and Research, Singapore.

Zhengguo served as a TPC Chair of IEEE IECON 2023 and 2020, a special session chair of IEEE ICASSP 2022, a Technical Brief Co-Founder of SIGGRAPH Asia, and leading Workshop Chair of IEEE ICME in 2013. He was an Associate Editor of IEEE Signal Processing Letters from 2014-2016, IEEE Transactions on Image Processing from 2016-2020, IEEE Transactions on Circuits and Systems for Video Technology from 2020-2023, APSIPA Transactions on Signal and Information Processing from 2022-2025, an editor of MDPI Sensors from 2022-2024, and a Senior Area Editor of IEEE Transactions on Image Processing from 2020-2024.