

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 that does not depend on ground-truth during the training phase has received widespread attention. Most efforts focus on designing different types of network architectures and loss functions or handling edge cases, *e.g.*, occlusion and dynamic objects. In this work, we introduce a novel self-supervised depth estimation framework, dubbed MonoDiffusion, by formulating it as an iterative denoising process. Because the depth ground-truth is unavailable in the training phase, we develop a pseudo ground-truth diffusion process to assist the diffusion in MonoDiffusion. The pseudo ground-truth diffusion gradually adds noise to the depth map generated by a pre-trained teacher model. Moreover, the teacher model allows applying a distillation loss to guide the denoised depth. Further, we develop a masked visual condition mechanism to enhance the denoising ability of model. Extensive experiments are conducted on the KITTI and Make3D datasets and the proposed MonoDiffusion outperforms prior state-of-the-art competitors. The source code will be available at <https://github.com/ShuweiShao/MonoDiffusion>.

Index Terms—Depth estimation, Self-supervised learning, Diffusion, Denoising

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]–[7] have achieved remarkable advances, where the full supervised MDE [8]–[11] shows higher accuracy due to the available depth ground-truth. Nonetheless, the 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 relies solely on monocular videos are preferred because of the difficulty in collecting stereo data, for

This work was supported by the National Natural Science Foundation of China under grant 61620106012.

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 SRO department, Institute for Infocomm Research, 1 Fusionopolis Way, Singapore. (e-mail: ezhgli@i2r.a-star.edu.sg)

*(corresponding author: Weihai Chen.)

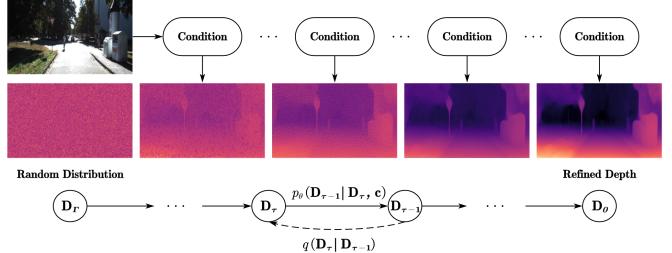


Fig. 1. Illustration of the denoising process guided by visual conditions.

example, intricate configurations and data processing. To boost the performance, following methods developed improved loss functions [15] or leveraged semantic information [16], [17] to solve the dynamic objects and occlusion. A recent work Lite-Mono [18] integrated convolutional neural network (CNN) and Transformer to design more powerful network architecture.

In this paper, we reformulate self-supervised MDE from the perspective of iterative denoising starting with a random depth distribution, as presented in Fig. 1. Diffusion models have attracted a widespread attention due to its efficacy in generative tasks [19], [20], detection [21] and segmentation [22]. More recently, Saxena *et al.* [23] and Duan *et al.* [24] applied the diffusion model to MDE in a full supervised setting. However, the lack of depth ground-truth in self-supervised MDE poses a severe challenge to the diffusion process that requires ground-truth in diffusion model.

We propose MonoDiffusion, which takes as input a random depth distribution and progressively refines it through multiple denoising steps under the guidance of visual conditions. Since the depth ground-truth is not available in the training phase, we introduce a pseudo ground-truth diffusion process to assist the diffusion in MonoDiffusion. Specifically, the pseudo ground-truth diffusion gradually appends noise to the depth generated by a pre-trained teacher model, which we refer to as the pseudo ground-truth. As a by-product, the teacher model allows us to impose a knowledge distillation loss on the denoised depth to improve the results. In order to alleviate the negative impact of depth error in pseudo ground-truth, we apply a multi-view check filter [25] to filter out erroneous depth. Furthermore, we develop a masked visual condition mechanism to enhance the denoising ability of MonoDiffusion, inspired by the success of combining diffusion model with masked image modeling [26], [27]. The difference is that our target to reconstruction is the denoised depth rather than the RGB image.

To summarize, our contributions are listed as follows:

- We propose a novel framework, dubbed MonoDiffusion,

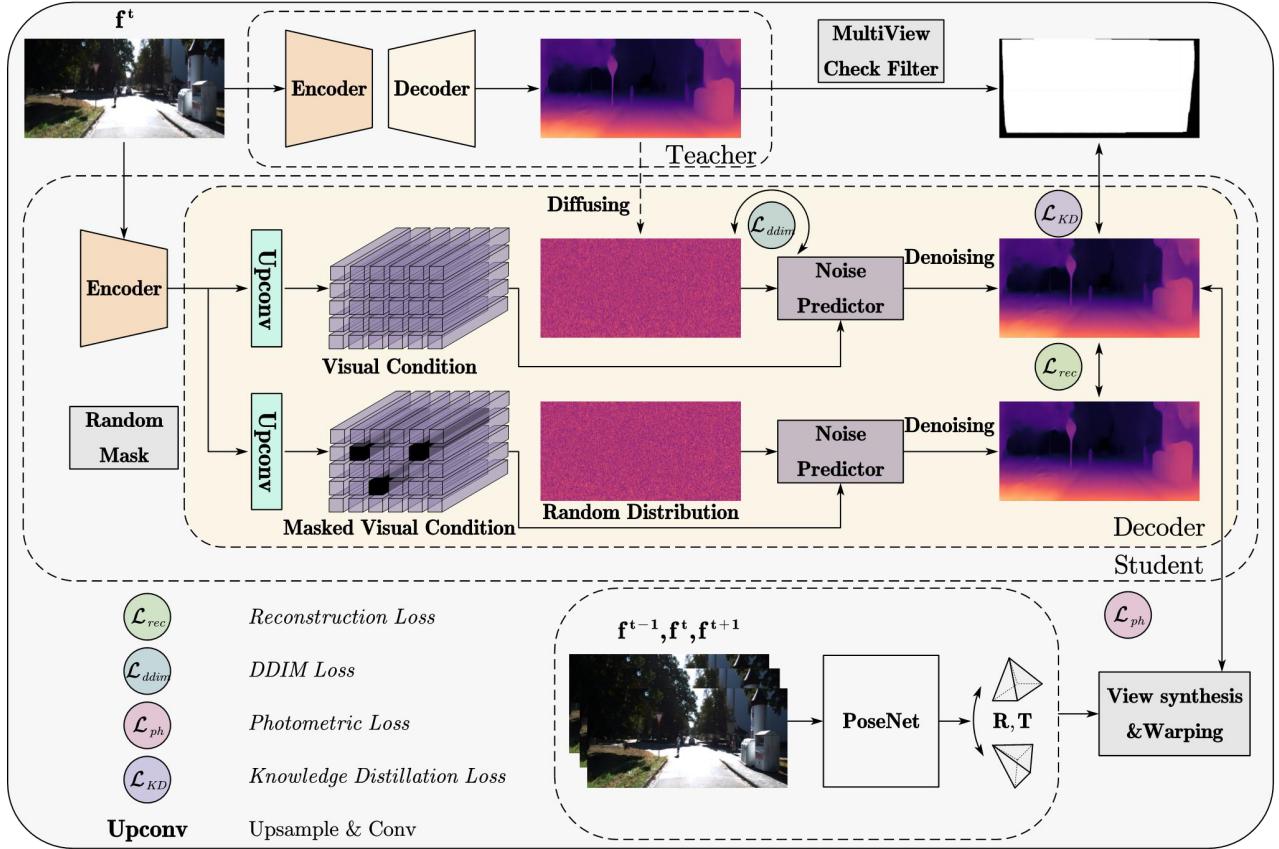


Fig. 2. **An overview of the proposed MonoDiffusion.** During the training phase, MonoDiffusion involves an additional teacher model to assist the diffusion. The teacher model is self-supervised pre-trained based on Lite-Mono [18] and will be discarded once the training is completed.

for self-supervised monocular depth estimation by regarding it as an iterative denoising process.

- We introduce a pseudo ground-truth diffusion process to assist the diffusion in MonoDiffusion and a masked visual condition mechanism to enhance its denoising ability.
- Extensive experiments show that the proposed MonoDiffusion surpasses previous state-of-the-art competitors on the KITTI [28] and Make3D [29] 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 same 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 using CNN to perform multi-scale depth estimation. Then, Laina *et al.* [30] utilized the residual CNN [31] to enable network optimization to be easier. Cao *et al.* [32] and Fu *et al.* [33] discretized the full depth range into multiple intervals and chose the optimal interval as depth estimate. Yuan *et al.* [34] developed neural window fully-connected conditional random fields (CRFs) to reduce the computation in conventional CRFs. Shao *et al.* [11] introduced the cross-distillation paradigm to integrate strengths

from CNN and Transformer. Liu *et al.* [35] enforced the first-order variational constraint to regularize depth map. However, these methods rely heavily on the quality and quantity of depth ground-truth while the proposed MonoDiffusion does not have such a requirement.

Self-supervised depth estimation does not require costly depth ground-truth and employ 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 to predict depth and 6-degree of freedom (DoF) pose, which are then utilized to perform view synthesis during training. To handle edge cases such as dynamic objects and occlusion, they proposed an explainability mask to remove these regions. Godard *et al.* [15] further introduced an auto-masking technique and a per-pixel minimum re-projection loss to better handle dynamic objects and occlusion. Bian *et al.* [36] devised a geometry consistency loss for scale-consistent depth and pose predictions. Johnston *et al.* [37] and Liu *et al.* [25] utilized the discrete disparity prediction and self-reference distillation to boost performance, respectively. Zhang *et al.* [18] developed an efficient hybrid architecture by combining CNN and Transformer. In contrast, we introduce the diffusion model and draw on its strong generative capability to produce high-quality depth predictions.

B. Diffusion Model

Diffusion models have received widespread attention due to its effectiveness in image generation [38]–[40]. Nevertheless, its potential in downstream tasks has largely been unexplored. Fortunately, Song *et al.* [41] improved the denoising process to make inference steps more affordable for these tasks. [42]–[44] extended diffusion model in image segmentation and Chen *et al.* [21] leveraged diffusion model to generate detection box proposals. There are also attempts at applying diffusion models to MDE [23], [24], but they focus on a fully supervised setting. To alleviate the performance degradation induced by diffusing on sparse ground-truth, Saxena *et al.* [23] and Duan *et al.* [24] both proposed to diffuse on the denoised output of network to assist training. Self-supervised MDE with the diffusion model is more challenging due to the lack of depth ground-truth. In addition, we find that diffusing on the denoised output as done by these two methods does not work well for self-supervised MDE.

III. METHODOLOGY

In this section, we first introduce the preliminary knowledge of diffusion model and self-supervised MDE. Then, we elaborate on the proposed MonoDiffusion, namely, depth estimation as denoising, pseudo ground-truth diffusion and masked visual condition. Finally, we demonstrate the overall architecture and loss. An overview of the whole framework is shown in Fig. 2.

A. Preliminaries

Diffusion models, for example, [38], [41], [45], belong to the category of latent variable models and widely employed in generative tasks. In practice, they are trained to denoise images blurred with Gaussian noise and reverse the diffusion process $q(\mathbf{x}_\tau | \mathbf{x}_0)$, which involves the iterative addition of noise to a desired image distribution \mathbf{x}_0 and allows acquiring the latent noisy sample \mathbf{x}_τ . Mathematically,

$$q(\mathbf{x}_\tau | \mathbf{x}_0) := \mathcal{N}(\mathbf{x}_\tau | \sqrt{\bar{\alpha}_\tau} \mathbf{x}_0, (1 - \bar{\alpha}_\tau) \mathbf{I}), \quad (1)$$

with

$$\bar{\alpha}_\tau := \prod_{n=0}^{\tau} \alpha_n = \prod_{n=0}^{\tau} (1 - \beta_n), \quad (2)$$

and

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

where τ consists of T steps, β_n stands for the noise variance schedule as DDPM [38] and \mathcal{N} stands for the Gaussian noise.

For the denoising process, a noise predictor $\epsilon_\theta(\mathbf{x}_\tau, \tau)$ learns to reverse the diffusion process and recover \mathbf{x}_0 . Each denoising step is approximated by a Gaussian distribution,

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

where $\mu_\theta(\mathbf{x}_\tau, \tau)$ is acquired by the linear combination of \mathbf{x}_τ and predicted noise of $\epsilon_\theta(\mathbf{x}_\tau, \tau)$ and σ_τ^2 denotes 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. 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

$$\mathbf{p}^{t \rightarrow s} = \mathbf{K} \mathbf{M}^{t \rightarrow s} \mathbf{D}^t(\mathbf{p}) \mathbf{K}^{-1} \mathbf{p}^t, \quad (5)$$

where $\mathbf{p}^{t \rightarrow s}$ 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, $\mathbf{M}^{t \rightarrow s}$ denotes the relative pose from target view to source view, and \mathbf{D}^t denotes the depth map of target frame. Next, we can obtain a warped frame $\mathbf{f}^{s \rightarrow t}$ via

$$\mathbf{f}^{s \rightarrow t}(\mathbf{p}) = \mathbf{f}^s \langle \mathbf{p}^{t \rightarrow s} \rangle, \quad (6)$$

where $\langle \cdot \rangle$ denotes the warping operation [46]. The appearance difference between $\mathbf{f}^{s \rightarrow t}$ and \mathbf{f}^t is used to supervise the whole framework. As in most works, *e.g.*, [15], [47], a combination of L1 loss and structural similarity (SSIM) term [48] measures the dissimilarity in appearance, written as

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

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

B. MonoDiffusion

Depth estimation as denoising. Given the target frame \mathbf{f}^t , a standard formulation for MDE is $p_\theta(\mathbf{D}^t | \mathbf{f}^t)$. By contrast, we reframe the MDE as an iterative denoising process where visual conditions guide the refinement of random depth distribution \mathbf{D}_τ into a depth estimate,

$$p_\theta(\mathbf{D}_{\tau-1} | \mathbf{D}_\tau, \mathbf{c}) := \mathcal{N}(\mathbf{D}_{\tau-1}; \mu_\theta(\mathbf{D}_\tau, \tau, \mathbf{c}), \sigma_\tau^2 \mathbf{I}), \quad (8)$$

where \mathbf{c} stands for the conditions. To speed up the inference process, we employ an improved inference process from [41], where $\sigma_\tau^2 \mathbf{I}$ is set to 0 to allow the output to be deterministic.

Pseudo ground-truth diffusion. As described above, the diffusion process in diffusion model iteratively appends noise to a desired distribution, *i.e.*, ground-truth. Unfortunately, the lack of depth ground-truth for self-supervised MDE makes it hard to perform. To address the issue, we introduce a pseudo ground-truth diffusion process. More specifically, we pre-train a teacher model based on Lite-Mono [18], a well-behaved self-supervised monocular depth estimator. Then, the diffusion is performed on the depth generated by our teacher model, which we referred to as the pseudo ground-truth $\mathbf{D}_{pseudo}(\mathbf{p})$,

$$q(\mathbf{D}_\tau | \mathbf{D}_{pseudo}) := \mathcal{N}(\mathbf{D}_\tau | \sqrt{\bar{\alpha}_\tau} \mathbf{D}_{pseudo}, (1 - \bar{\alpha}_\tau) \mathbf{I}). \quad (9)$$

The teacher model will be discarded once the training is completed. Compared with diffusion on the denoised output [23], [24], the pseudo ground-truth process is able to avoid the noisy denoised output at the early training stage from deteriorating the entire training process. As a by-product, the teacher model

allows imposing a knowledge distillation loss on the denoised depth. The knowledge distillation paradigm distills knowledge from the teacher model to a student model, which is beneficial to improve the results [11]. To mitigate the adverse impact of depth error in the pseudo ground-truth, we apply a multi-view check filter [25] to filter out erroneous depth. The knowledge distillation loss is thus defined as

$$\mathcal{L}_{KD} = \sum_{\mathbf{p}} \Phi(\mathbf{p}) \odot (\mathbf{D}^t(\mathbf{p}) - \mathbf{D}_{pseudo}(\mathbf{p})), \quad (10)$$

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. We draw on Eq. 5 to get each projected point $\mathbf{p}^{t \rightarrow s}$ in the source view and acquire a warped depth map,

$$\mathbf{D}^{s \rightarrow t}(\mathbf{p}) = \mathbf{D}^s \langle \mathbf{p}^{t \rightarrow s} \rangle, \quad (11)$$

where \mathbf{D}^s stands for the depth map of source frame. Similar to the projection procedure above, we leverage \mathbf{D}^s to perform a reprojection procedure to acquire each reprojected 2D point $\tilde{\mathbf{p}}^{s \rightarrow t}$ and depth map $\tilde{\mathbf{D}}^{s \rightarrow t}(\mathbf{p})$ in the target view. Thereafter, a reprojection error e_{repro} and a geometry error e_{geo} are defined as

$$e_{repro} = \left\| \tilde{\mathbf{p}}^{s \rightarrow t} - \mathbf{p}^t \right\|_2, \quad (12)$$

$$e_{geo} = \frac{\left| \tilde{\mathbf{D}}^{s \rightarrow t}(\mathbf{p}) - \mathbf{D}^t(\mathbf{p}) \right|}{\mathbf{D}^t(\mathbf{p})}, \quad (13)$$

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

$$\{\mathbf{p}\} = \{\mathbf{p} | e_{repro} < a\bar{e}_{repro}, e_{geo} < b\bar{e}_{geo}\}, \quad (14)$$

where \bar{e}_{repro} and \bar{e}_{geo} are the average over all pixels and a and b are set to 4 based on [25].

Masked visual condition. Inspired by the success of coupling diffusion model with masked image modeling [26], [27], we develop a masked visual condition mechanism to further enhance the denoising ability of MonoDiffusion. For multi-scale feature tokens propagated by the encoder, we generate random masks to remove part of the tokens from these feature tokens. To prevent the information leakage, these masks are shared otherwise masked information may be easily borrowed from feature tokens of different resolutions. More specifically, we generate a mask with the highest resolution, and leverage nearest neighbor interpolation to acquire a pyramid of masks. The masked feature tokens are aggregated into masked visual conditions through learnable layers composed of 3×3 convolutions and upsampling layers. Different from [26], [27], we make use of the masked visual conditions to reconstruct the denoised depth map guided by the complete visual conditions, instead of the RGB image. The reconstruction loss is defined as

$$\mathcal{L}_{rec} = \sum_{\mathbf{p}} \left| \hat{\mathbf{D}}^t(\mathbf{p}) - \mathbf{D}^t(\mathbf{p}) \right|, \quad (15)$$

where $\hat{\mathbf{D}}^t$ denotes the denoised depth map using masked visual conditions.

Layers	Channels	Input	Activation
mask 2	128	encoder 2	None
upconv 2	128	mask 2	ELU
upconv 2 _{skip}	144	↑upconv 2, encoder 1	ELU
mask 1	64	encoder 1	None
upconv 1	64	mask 1	ELU
upconv 1 _{skip}	88	↑upconv, encoder 0	ELU
mask 0	40	encoder 0	None
upconv 0	40	mask 0	ELU
upconv 0 _{skip}	24	↑upconv0	ELU
NE	1	random noise	None
TE	1	scheduler timesteps	None
NP	16	upconv 0 _{skip} , NE, TE	None

TABLE I
THE DECODER ARCHITECTURE OF STUDENT DEPTH NETWORK. THE KERNEL SIZE AND STRIDE OF CONVOLUTION ARE 3×3 AND 1. “↑” DENOTES THE 2×2 BI-LINEAR UPSAMPLING. THE SUBSCRIPT “SKIP” DENOTES THE SKIP CONNECTION. THE NE, TE AND NP DENOTE NOISE EMBEDDING, TIME EMBEDDING AND NOISE PREDICTOR, RESPECTIVELY.

C. Network Architecture and Overall Loss

Both **depth networks** adopt the prevalent encoder-decoder architecture. The teacher model is from [18], which leverages Lite-Mono-8M as the encoder. On the other hand, the student model uses two encoders from Lite-Mono family in different settings, Lite-Mono and Lite-Mono-8M. Its decoder is adapted from [18] and multi-scale feature tokens from the encoder are aggregated into visual conditions through 3×3 convolutional layers and upsampling layers. The diffusion-denoising process is performed at the full resolution. The deployed noise predictor is lightweight, only involving an embedding layer and three 3×3 convolutional layers. The detailed decoder architecture of student depth network is presented in Table I.

The **pose network** design is same to prior works [15], [18], where a pre-trained ResNet18 [31] is used as the encoder and four convolutional layers are used as the decoder. It receives a stacked pair of images and estimate a 6-DOF relative pose between them.

DDIM loss. Following [41], we apply a DDIM loss to supervise the predicted noise at each denoising step by reversing the diffusion process,

$$\mathcal{L}_{ddim} = \sum_{\mathbf{p}} \|\epsilon - \epsilon_{\theta}(\mathbf{D}_{\tau}, \tau, \mathbf{c})\|^2. \quad (16)$$

Edge-aware smoothness loss. As defined in [15], we apply an edge-aware smoothness loss to encourage the smoothness property of depth map,

$$\mathcal{L}_{es} = \sum_{\mathbf{p}} |\nabla \mathbf{D}^t(\mathbf{p})| \odot \exp^{-\|\nabla \mathbf{f}^t(\mathbf{p})\|_1}, \quad (17)$$

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

Overall loss. The overall optimization objective is summarized as

$$\mathcal{L}_{overall} = \lambda_1 \mathcal{L}_{ph} + \lambda_2 \mathcal{L}_{KD} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{ddim}, \quad (18)$$

Method	Year	Data	Depth Error (\downarrow)				Depth Accuracy (\uparrow)			Model Size (\downarrow)
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	
GeoNet [49]	2018	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975	31.6M
DDVO [50]	2018	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974	28.1M
Monodepth2-Res18 [15]	2019	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981	14.3M
Monodepth2-Res50 [15]	2019	M	0.110	0.831	4.642	0.187	0.883	0.962	0.982	32.5M
SGDepth [17]	2020	M+Se	0.113	0.835	4.693	0.191	0.879	0.961	0.981	16.3M
Johnston <i>et al.</i> [37]	2020	M	0.111	0.941	4.817	0.189	0.885	0.961	0.981	14.3M+
CADepth-Res18 [51]	2021	M	0.110	0.812	4.686	0.187	0.882	0.962	0.983	18.8M
HR-Depth [52]	2021	M	0.109	0.792	4.632	0.185	0.884	0.962	0.983	14.7M
Lite-HR-Depth [52]	2021	M	0.116	0.845	4.841	0.190	0.866	0.957	0.982	3.1M
R-MSFM3 [53]	2021	M	0.114	0.815	4.712	0.193	0.876	0.959	0.981	3.5M
R-MSFM6 [53]	2021	M	0.112	0.806	4.704	0.191	0.878	0.960	0.981	3.8M
MonoFormer [54]	2023	M	0.108	0.806	4.594	0.184	0.884	0.963	0.983	23.9M+
SRDepth-Res18 [25]	2023	M	0.111	<u>0.762</u>	4.619	0.186	0.877	0.961	0.983	23.3M
Lite-Mono [18]	2023	M	<u>0.107</u>	0.765	4.561	<u>0.183</u>	0.886	0.963	0.983	3.1M
MonoDiffusion (ours)	2023	M	0.103	0.726	4.447	0.179	0.893	0.965	0.984	3.1M
Monodepth2-Res18 [15]	2019	M \dagger	0.132	1.044	5.142	0.210	0.845	0.948	0.977	14.3M
Monodepth2-Res50 [15]	2019	M \dagger	0.131	1.023	5.064	0.206	0.849	0.951	0.979	32.5M
R-MSFM3 [53]	2021	M \dagger	0.128	0.965	5.019	0.207	0.853	0.951	0.977	3.5M
R-MSFM6 [53]	2021	M \dagger	0.126	0.944	4.981	0.204	0.857	0.952	0.978	3.8M
Lite-Mono [18]	2023	M \dagger	<u>0.121</u>	0.876	4.918	<u>0.199</u>	0.859	0.953	0.980	3.1M
MonoDiffusion (ours)	2023	M \dagger	0.119	0.843	4.868	0.196	0.858	0.955	0.981	3.1M
Monodepth2-Res18 [15]	2019	M *	0.115	0.882	4.701	0.190	0.879	0.961	0.982	14.3M
R-MSFM3 [53]	2021	M *	0.112	0.773	4.581	0.189	0.879	0.960	0.982	3.5M
R-MSFM6 [53]	2021	M *	0.108	0.748	4.470	0.185	0.889	0.963	0.982	3.8M
HR-Depth [52]	2021	M *	0.106	0.755	4.472	0.181	0.892	0.966	0.984	14.7M
SRDepth-Res18 [25]	2023	M *	0.106	<u>0.673</u>	4.379	0.180	0.886	0.965	0.984	23.3M
Lite-Mono [18]	2023	M *	0.102	0.746	4.444	0.179	0.896	0.965	0.983	3.1M
Lite-Mono-8M [18]	2023	M *	<u>0.097</u>	0.710	4.309	<u>0.174</u>	0.905	0.967	0.984	8.7M
MonoDiffusion (ours)	2023	M *	0.099	0.702	4.305	0.175	0.903	0.967	0.984	3.1M
MonoDiffusion-8M (ours)	2023	M *	0.094	0.662	4.235	0.171	0.908	0.968	0.984	8.8M
MonoViT-tiny [55]	2022	M	0.102	0.733	4.459	<u>0.177</u>	0.895	0.965	0.984	10.3M
Lite-Mono-8M [18]	2023	M	<u>0.101</u>	<u>0.729</u>	<u>4.454</u>	0.178	0.897	0.965	0.983	8.7M
MonoDiffusion-8M (ours)	2023	M	0.099	0.692	4.377	0.175	0.899	0.966	0.984	8.8M

TABLE II

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 UNDERLINE, 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 \dagger ” DENOTES BACKBONES WITHOUT PRE-TRAINING ON IMAGENET [56]. MONODIFFUSION AND MONODIFFUSION-8M USES LITE-MONO AND LITE-MONO-8M AS ENCODERS, RESPECTIVELY. ASIDE FROM THE DEPTH ERROR AND ACCURACY, WE REPORT THE MODEL SIZE.

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are empirically set to 1, 1, 0.1 and 1, respectively. Similar to [15], 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 [15] is used to deal with dynamic objects.

IV. EXPERIMENT

We conduct extensive experiments on two standard datasets, which includes KITTI [28] and Make3D [29]. In the following part, we first describe the relevant datasets, evaluation metrics and implementation details. Then, we present quantitative and qualitative comparisons to the state-of-the-art competitors. Finally, we conduct zero-shot generalization and ablation studies to perform a thorough analysis of MonoDiffusion.

A. Datasets and Evaluation Metrics

The **KITTI** dataset is composed of 61 stereo road scenes 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 [15], we adopt

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. We only use this dataset for a zero-shot generalization study by utilizing models trained on the KITTI dataset.

B. Evaluation Metrics

Similar to [18], 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 $(\max(\frac{d_t}{d}, \frac{d}{d_t}) = \delta < thr)$ 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

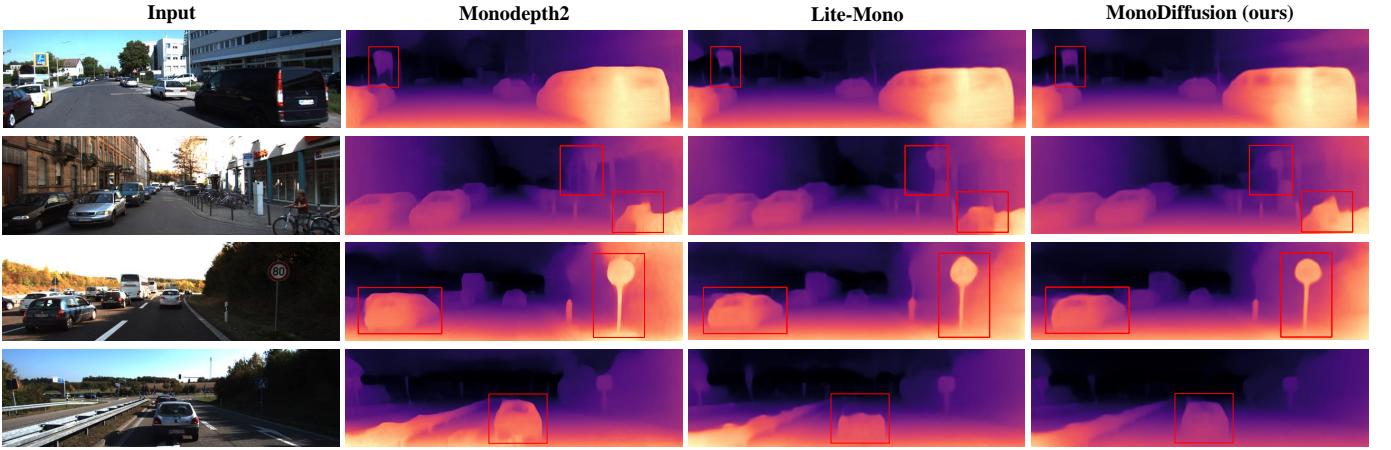


Fig. 3. Qualitative depth comparison on the KITTI dataset. The red boxes indicate the regions to emphasize.

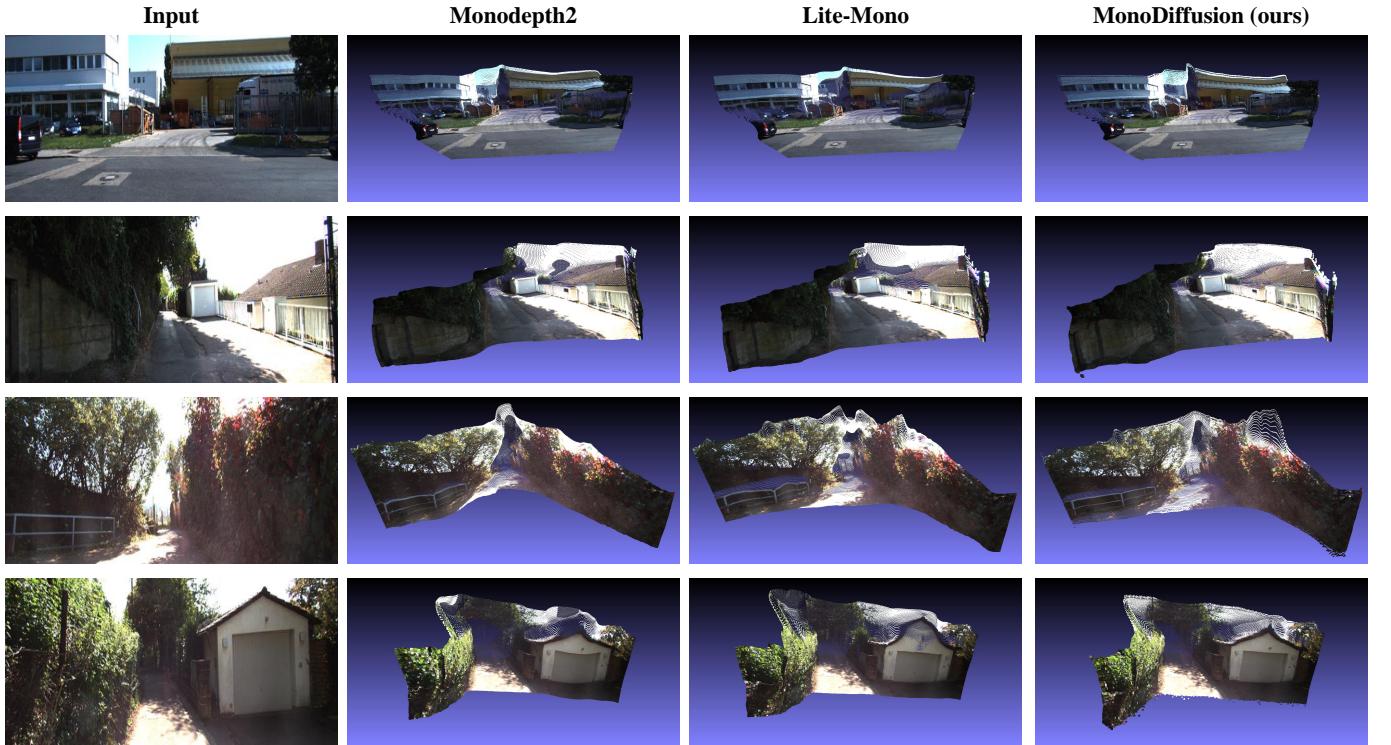


Fig. 4. Qualitative point cloud results on the KITTI dataset.

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.

C. Implementation Details

MonoDiffusion is implemented in the PyTorch library and trained on a single NVIDIA TITAN RTX. We use the AdamW optimizer [57] 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 [58]. The training process runs a total number of 35 epochs in this scenario. It is observed that pre-training on the ImageNet [56] accelerates network convergence, resulting

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 [18], [54], we leverage random horizontal flip as well as random brightness adjustment, saturation adjustment, contrast adjustment and hue jitter with a 50% chance. We utilize the improved sampling process with 1000 diffusion steps for training and 20 denoising steps for inference.

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

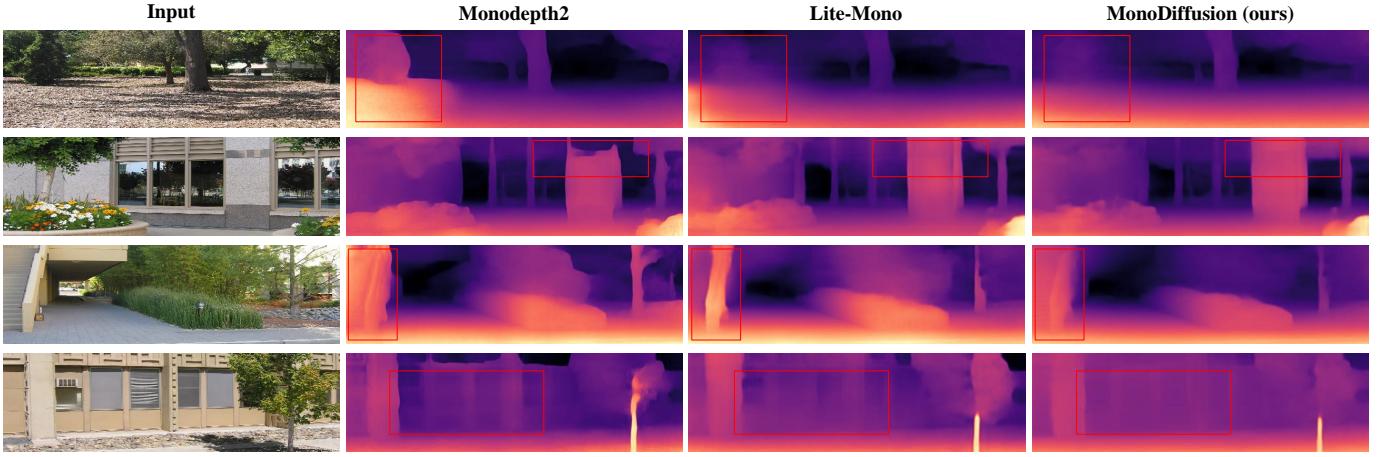


Fig. 5. Qualitative depth comparison on the Make3D dataset.

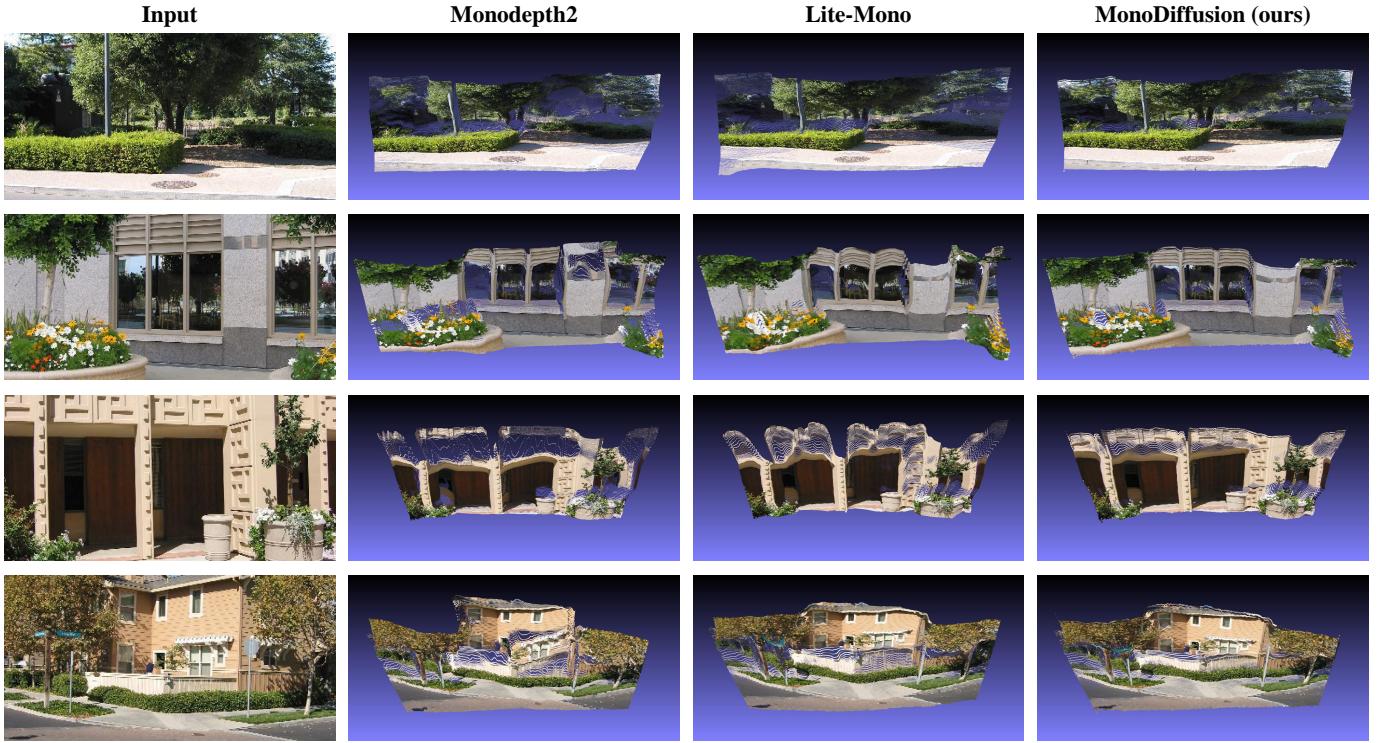


Fig. 6. Qualitative point cloud results on the Make3D dataset.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log↓
Monodepth [59]	0.544	10.940	11.760	0.193
SFMLearner [5]	0.383	5.321	10.470	0.478
DDVO [50]	0.387	4.720	8.090	0.204
Monodepth2 [15]	0.322	3.589	7.417	0.163
R-MSFM6 [53]	0.334	3.285	7.212	0.169
Lite-Mono [18]	0.305	3.060	6.981	0.158
MonoDiffusion (ours)	0.295	2.849	6.854	0.150

TABLE III

QUANTITATIVE COMPARISON ON THE MAKE3D [29] DATASET. ALL MODELS ARE TRAINED ON KITTI [28] WITH AN IMAGE RESOLUTION OF 640×192 .

in Table II. 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 Fig. 3, we show a qualitative depth comparison. MonoDiffusion is better at delineating object contours, for example, traffic sign and preserves fine-grained depth details. To further show the strengths of MonoDiffusion, we convert depth maps into point clouds and display the 3D structures in Fig 4. It can be seen that MonoDiffusion is capable of recovering the 3D world reasonably and shows less distortion than the compared methods. The outstanding performance of our MonoDiffusion evidences the strong generative capability of diffusion model.

ID	SD	PGD	\mathcal{L}_{KD}	MVC	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	RMSE log \downarrow	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
1					0.107	0.787	4.580	0.182	0.887	0.963	0.983
2	✓				0.434	4.630	11.99	0.579	0.306	0.568	0.777
3		✓			0.105	0.757	4.521	0.181	0.889	0.964	0.984
4		✓	✓		0.104	0.741	4.489	0.180	0.890	0.964	0.984
5	✓	✓	✓	✓	0.103	0.726	4.447	0.179	0.893	0.965	0.984

TABLE IV

ABALATION STUDY ON THE MONODIFFUSION. SD: SELF-DIFFUSION [24]; PGD: PSEUDO GROUND-TRUTH DIFFUSION; MVC: MASKED VISUAL CONDITION MECHANISM.

Inference step	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	RMSE log \downarrow	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
2	1.812	75.600	30.791	6.160	0.070	0.141	0.213
5	1.503	58.013	27.140	5.464	0.125	0.241	0.340
10	0.226	3.730	9.571	0.754	0.703	0.878	0.938
15	0.822	28.099	19.232	2.972	0.357	0.548	0.653
20	0.103	0.726	4.447	0.179	0.893	0.965	0.984
25	0.431	10.382	12.709	1.206	0.503	0.723	0.827

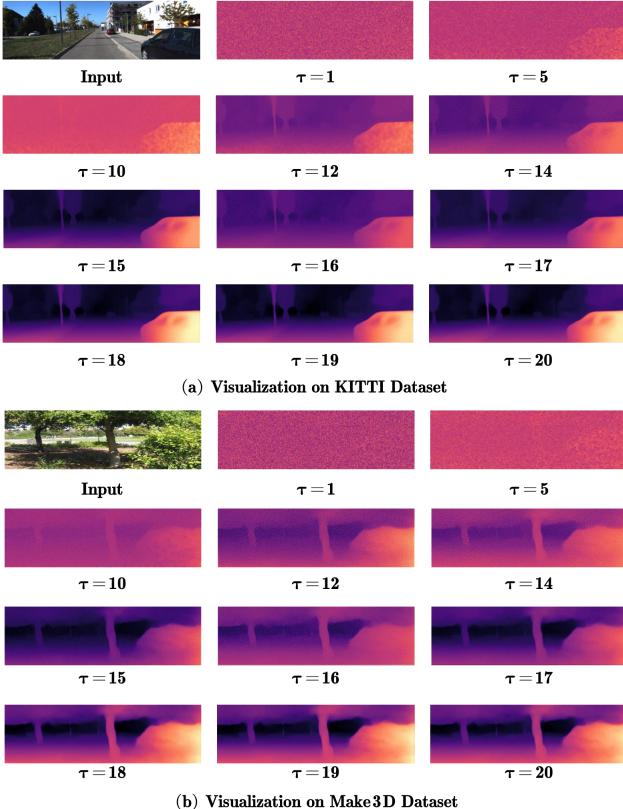
TABLE V
ABALATION STUDY ON DIFFERENT INFERENCE STEPS.

Fig. 7. Visualization of the denoising process involving 20 inference steps in total.

E. Zero-shot Generalization

In Table III, we verify the generalization ability of MonoDiffusion in a fully zero-shot setting. The models are trained on the KITTI dataset but evaluated on the Make3D dataset. The superior results of MonoDiffusion indicate that it generalizes well to unseen scenarios. In Fig. 5, we demonstrate a

qualitative depth comparison. As we can see, MonoDiffusion acquires highly-detailed depth maps with good visual quality. In Fig 6, we further show a qualitative point cloud comparison. The compared methods struggle with thin structures, *e.g.*, pole, while MonoDiffusion is able to recover these smaller details and preserves prominent geometric features of the 3D scenes at the same time.

F. Ablation Study

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

MonoDiffusion (Table IV). We leverage the Lite-Mono [18] as our baseline (ID 1). The implementation of diffusion is hard for self-supervised MDE due to the lack of depth ground-truth. To allow the diffusion, we utilize the self-diffusion [24] and the proposed pseudo ground-truth diffusion, respectively and find that the self-diffusion does not even make the model converge in the self-supervised setting (ID 2). By contrast, the pseudo ground-truth diffusion enables the model to achieve good performance (ID 3). We further appending the knowledge distillation loss in the training phase and achieves consistent improvements on almost all metrics (ID 4). Finally, we integrate the masked visual condition mechanism and acquire the best results (ID 5).

Denoising inference. We further explore the properties of different inference steps during the denoising process. It can be seen from Table V that when the number of steps increases, the performance gradually improves. However, the improvement is not always continuous. For example, the performance of step 10 is better than that of step 15. Besides, when continuing to increase the number of steps from the standard denoising step 20, the performance degrades.

In Fig. 7, we show an intuitive illustration of how the depth is iteratively refined from a random depth distribution. As can be seen, the process begins by initializing shapes and edges of

Mask ratio	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
0%	0.104	0.741	4.489	0.180	0.890	0.964	0.984
10%	0.104	0.729	4.482	0.179	0.891	0.964	0.983
20%	0.103	0.726	4.447	0.179	0.893	0.965	0.984
40%	0.103	0.741	4.463	0.179	0.892	0.965	0.984
60%	0.103	0.718	4.462	0.179	0.891	0.965	0.984
80%	0.107	0.772	4.587	0.182	0.885	0.963	0.983

TABLE VI

ABLATION STUDY ON DIFFERENT MASK RATIOS IN THE MASKED VISUAL CONDITION MECHANISM, WHICH IS ONLY UTILIZED DURING THE TRAINING PHASE.

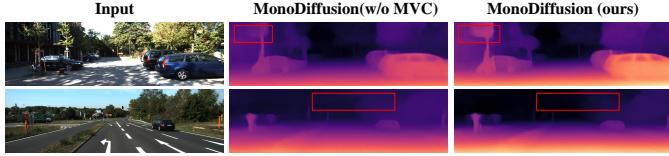


Fig. 8. Qualitative comparison between MonoDiffusion with and without masked visual condition (MVC) mechanism.

objects and then gradually refines the depth values and rectifies the inter- and intra-object distance relationships.

Masked visual condition. In Fig. 8, we display a qualitative depth comparison between MonoDiffusion with and without masked visual condition (MVC) mechanism. As can be seen, the MVC mechanism can indeed enhance the denoising ability of model, *e.g.*, better tree silhouettes. Additionally, the depth at distant distances can be better recovered, for instance, the sky. In the absence of MVC mechanism, the depth of sky is similar to the depth of distant trees, but in fact their depth should be highly different. The reason behind this may be that using the masked visual conditions to reconstruct the denoised depth enables the noise predictor to better exploit contextual information. As presented in Table VI, we further explore the impact of different mask ratios in the MVC mechanism. We find that 20% is the best choice, and the model is capable of achieving feasible results even with a mask training ratio of 80%.

V. CONCLUSION

In this work, we develop a novel self-supervised monocular depth estimation framework by reformulating it as an iterative denoising process. Besides, we introduce the pseudo ground-truth diffusion process to assist the diffusion and the masked visual condition mechanism to strengthen the denoising ability of model. Extensive experiments on the KITTI and Make3D datasets indicate the efficacy of the proposed MonoDiffusion. We hope our novel initiative will encourage more sophisticated diffusion-based 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] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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. *arXiv preprint arXiv:2302.08149*, 2023.
- [12] Xincheng 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] 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.
- [14] Aleksei Karpov and Ilya Makarov. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 711–719. IEEE, 2022.
- [15] 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.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022.
- [20] Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.
- [21] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.
- [22] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022.
- [23] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.
- [24] Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023.
- [25] Zhong Liu, Ran Li, Shuaiwei 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*, 2023.
- [26] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023.
- [27] Chen Wei, Karttikaya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*, 2023.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [39] 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.
- [40] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.
- [42] 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.
- [43] 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.
- [44] Dmitry Baranovskiy, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations*, 2022.
- [45] 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.
- [46] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] Jinwoo Bae, SungHo Moon, and SungHo 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.
- [55] 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.
- [56] 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.
- [57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2018.
- [58] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.
- [59] 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.