DiffusionDepth: Diffusion Denoising Approach for Monocular Depth Estimation

Yiquan Duan[1], Xianda Guo[2,3*], and Zheng Zhu[4]

[1] Human-centric Artificial Intelligence Centre, Australian Artificial Intelligence Institute, University of Technology Sydney, 2007 NSW [2]School of Computer Science, Wuhan University [3]Waytous [4]GigaAI
yiqun.duan-1@uts.edu.au; xianda_guo@163.com; zhengzhu@ieee.org

Abstract. Monocular depth estimation is a challenging task that predicts the pixel-wise depth from a single 2D image. Current methods typically model this problem as a regression or classification task. We propose DiffusionDepth, a new approach that reformulates monocular depth estimation as a denoising diffusion process. It learns an iterative denoising process to 'denoise' random depth distribution into a depth map with the guidance of monocular visual conditions. The process is performed in the latent space encoded by a dedicated depth encoder and decoder. Instead of diffusing ground truth (GT) depth, the model learns to reverse the process of diffusing the refined depth of itself into random depth distribution. This self-diffusion formulation overcomes the difficulty of applying generative models to sparse GT depth scenarios. The proposed approach benefits this task by refining depth estimation step by step, which is superior for generating accurate and highly detailed depth maps. Experimental results from both offline and online evaluations using the KITTI and NYU-Depth-V2 datasets indicate that the proposed method can achieve state-of-the-art performance in both indoor and outdoor settings while maintaining a reasonable inference time. The codes [1] are available online.

Keywords: Depth Estimation · Diffusion-Denoising Probalistic Models

1 Introduction

Monocular depth estimation is a fundamental vision task with numerous applications such as autonomous driving [9, 14, 68], robotics [56, 65], and augmented reality. Along with the rise of convolutional neural networks (CNNs) [13, 22, 57], numerous mainstream methods employ it as dense per-pixel regression problems, such as RAP [66], DAV [26], and BTS [30]. However, these pure regression methods suffer from severe overfitting and unsatisfactory object details. To increase the robustness, the following methods utilizing constructed additional constraints such as uncertainty (UCRDepth [47]), and piecewise planarity prior

* Corresponding Author.
[1] https://github.com/duanyiqun/DiffusionDepth

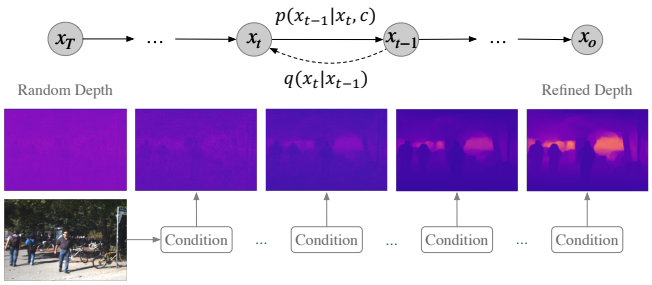


Fig. 1: Illustration of DiffusionDepth, the model refines the depth map x_t with monocular guidance c from random depth initialization x_T to the refined estimation x_o.

(P3Depth [41]). The NewCRFs [64] introduces window-separated Conditional Random Fields (CRF) to enhance local space relation with neighbor pixels. DORN [17], and Soft Ordinary [11] propose to discretize continuous depth into several intervals and reformulate the task as a classification problem on low-resolution feature maps. Follow-up methods (AdaBins [5, 27], BinsFormer [34]) merge regression results with classification prediction from bin centers. However, the discretization of depth values derived from bin centers leads to reduced visual quality, characterized by noticeable discontinuities and blurring.

We solve the depth estimation task by reformulating it as an iterative denoising process that generates the depth map from random depth distribution. The brief process is described in Fig. 1. Intuitively, the iterative refinement enables the framework to capture both coarse and fine details in the scene at different steps. Meanwhile, by denoising with extracted monocular guidance on large latent space, this framework enables accurate depth prediction in high resolution. Diffusion models have shown remarkable success in generation tasks [25, 58], or more recently, on detection [7] and segmentation [7, 8] tasks. To the best of our knowledge, this is the first work introducing the diffusion model into depth estimation.

This paper proposes DiffusionDepth, a novel framework for monocular depth estimation as described in Fig. 2. The framework takes in a random depth distribution as input and iteratively refines it through denoising steps guided by visual conditions. By performing the diffusion-denoising process in latent depth space [45], DiffusionDepth is able to achieve more accurate depth estimation with higher resolution. The depth latent is composed of a subtle encoder and decoder. The denoising process is guided by visual conditions by merging it with the denoising block through a hierarchical structure (Fig. 3). The visual backbone extracts multi-scale features from monocular visual input and aggregated it through a feature pyramid (FPN [35]). We aggregated both global and local correlations to construct a strong monocular condition.

One severe problem of adopting generative methods into depth prediction is the sparse ground truth (GT) depth problem [2], which can lead to mode collapse in normal generative training. To address this issue, DiffusionDepth introduces a self-diffusion process. During training, instead of directly diffusing on sparse GT depth values, the model gradually adds noise to refined depth latent from the current denoising output. The supervision is achieved by aligning the refined depth predictions with the sparse GT values in both depth latent space and pixel-wise depth through a sparse valid mask. With the help of random crop, jitter, and flip augmentation in training, this process lets the generative model *organize* the entire depth map instead of just regressing on known parts, which largely improves the visual quality of the depth prediction.

The proposed DiffusionDepth framework is evaluated on widely used public benchmarks KITTI [18] and NYU-Depth-V2 [44], covering both indoor and outdoor scenarios. It could reach 0.298 and 1.452 RMSE on official offline test split respectively on NYU-Depth-V2 and KITTI datasets, which exceeds state-of-the-art (SOTA) performance. To better understand the effectiveness and properties of the diffusion-based approach for 3D perception tasks, we conduct a detailed ablation study. It discusses the impact of different components and design choices on introducing the diffusion approach to 3D perception, providing valuable insights as references for related tasks such as stereo and depth completion. The contribution of this paper could be summarized in threefold.

- This work proposes a novel approach to monocular depth estimation by reformulating it as an iterative diffusion-denoising problem with visual guidance.
- Experimental results suggest DiffusionDepth achieves state-of-the-art performance on both offline and online evaluations with affordable inference costs.
- This is the first work introducing the diffusion model into depth estimation, providing extensive ablation component analyses, and valuable insights for potentially related 3D vision tasks.

2 Related Works

Monocular Depth Estimation is an important task in computer vision that aims to estimate the depth map of a scene from a single RGB image. Early approach [46] utilized Markov random field to predict depth, while more approaches [16, 17, 42] leverage deep convolutional neural networks (CNNs) to achieve drastic performance. One popular approach is to formulate monocular depth estimation as a dense per-pixel regression problem. Many methods, including RAP [66], DAV [26], and BTS [30], have achieved impressive performance using this approach. Some follow-up approaches, such as UnetDepth [21], CANet [60], and BANet [2], focus on modifying the backbone structure to enhance visual features. Recently, transformer structures have been introduced

[2] In datasets such as KITTI Depth, only a small percentage of pixels ($3.75-5\%$) have GT depth values.

in monocular depth estimation, where DPT [43], and PixelFormer [1] have shown improved performances. To increase the robustness of monocular depth estimation, some methods introduce additional constraints such as uncertainty (UCRDepth [47]) or piecewise planarity prior (P3Depth [41]). NewCRFs [64] proposes window-separated Conditional Random Fields (CRF) to enhance the local space relation with neighboring pixels. AdaBins [5] and BinsFormer [34] revisited ordinal regression networks and reformulated the task as a classification-regression task by calculating adaptive bins based on image content to estimate depth. VA-Depth [36] first introduces variational inference into refined depth prediction. We further introduce the diffusion approach to this task and leverage powerful generative capacity to generate highly refined depth prediction.

Diffusion Model for Perception Tasks Although Diffusion models have achieved great success in image generation [10,24,55], their potential for discriminative tasks remains largely unexplored. The improved diffusion process [51] has made inference times to become more affordable for perception tasks, which has accelerated the exploration. Some initial attempts have been made to adopt diffusion models for image segmentation tasks [3,4,6,8,15,19,28,59]. These segmentation tasks are processed in an image-to-image style. DiffusionDet [7] first extends the diffusion process into generating detection box proposals. We propose to use the diffusion model for denoising the input image as a conditioned depth refinement process, instead of adopting it as a normal generative head. To the best of our knowledge, this is the first work introducing the diffusion model into monocular depth estimation.

3 Methodology

3.1 Task Reformulation

Preliminaries Diffusion models [24,50,52,53] are a class of latent variable models. It is normally used for generative tasks, where neural networks are trained to denoise images blurred with Gaussian noise by learning to reverse the diffusion process. The diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ as defined in Eq. 1,

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) \coloneqq \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{1}$$

iteratively adds noise to desired image distribution $\boldsymbol{x}_0$ and gets latent noisy sample $\boldsymbol{x}_t$ for $t \in \{0, 1, ..., T\}$ steps. $\bar{\alpha}_t \coloneqq \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$ and β_s represent the noise variance schedule [24]. In the denoising process, neural network $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ is trained to reverse $\boldsymbol{x}_0$ by interactively predicting $\boldsymbol{x}_{t-1}$ as below,

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \coloneqq \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\sigma}_t^2\boldsymbol{I}), \tag{2}$$

where $\boldsymbol{\sigma}_t^2$ denotes the transition variance. Sample $\boldsymbol{x}_0$ is reconstructed from prior noise $\boldsymbol{x}_T$ an mathematical inference process [24,52] iteratively, *i.e.*, $\boldsymbol{x}_T \rightarrow \boldsymbol{x}_{T-\Delta} \rightarrow ... \rightarrow \boldsymbol{x}_0$,

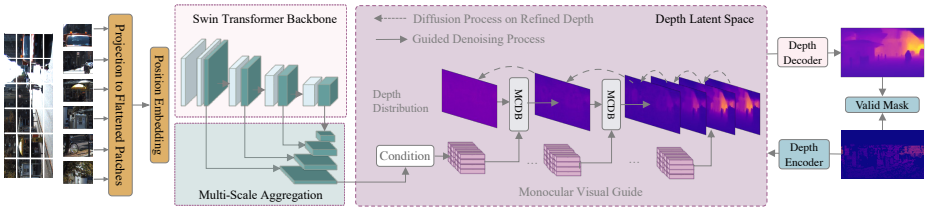


Fig. 2: Overview of DiffusionDepth. Given monocular visual input, the model employs a feature extractor and multiscale feature aggregation to construct visual guidance conditions. The Monocular Conditioned Denosing Block (MCDB) iteratively refines the depth distribution from noise initialization to refined depth prediction under the guidance of monocular visual conditions.

Denoising as Depth Refinement Given input image $\boldsymbol{c}$, the monocular depth estimation task is normally formulated as $p(\boldsymbol{x}|\boldsymbol{c})$, where $\boldsymbol{x}$ is the desired depth map. We reformulate the depth estimation as a visual-condition [3] guided denoising process which refines the depth distribution $\boldsymbol{x}_t$ iteratively as defined in Eq. 3 into the final depth map $\boldsymbol{x}_0$.

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c}) \coloneqq \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}), \boldsymbol{\sigma}_t^2 \boldsymbol{I}), \tag{3}$$

where model $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ is trained to refine depth latent $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$. To accelerate the denoising process, we utilized the improved inference process from DDIM [54], where it sets $\boldsymbol{\sigma}_t^2 \boldsymbol{I}$ as 0 to make the prediction output deterministic.

3.2 Network Architecture

We use Swin Transformer [37] shown in Fig. 2 as an example to illustrate the feature extraction. The input image is patched and projected into visual tokens with position embedding. The backbone extracts visual features at a different scale to maintain coarse and fine details of the input scene. Based on extracted multi-scale features, we employ hierarchical aggregation and heterogeneous interaction (HAHI [33]) to enhance features between scales. Feature pyramid neck [35] is applied to aggregate features into monocular visual condition. The **visual condition** is the aggregated feature map with a shape $\frac{H}{4} \times \frac{W}{4} \times c$, where H, W are respectively the height and width of the monocular image input, and c is the channel dimension the feature. The proposed DiffusionDepth model is suitable for most visual backbones which could extract multi-scale features. According to extensive experiments, other backbones such as ResNet [23], EfficientNet [57], and ViT [12] could achieve competitive performance as well.

[3] Here visual-condition denotes to extracted visual latent from backbones.

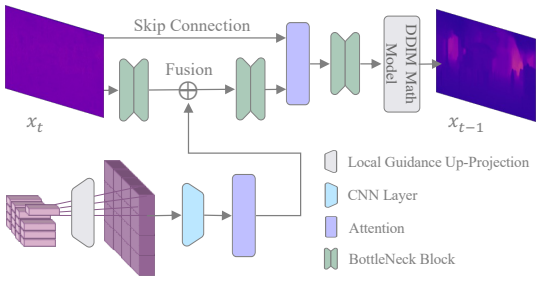


Fig. 3: Illustration of Monocular Conditioned Denoising Block. Visual condition is fused with depth latent through hierarchically.

3.3 Monocular Conditioned Denoising Block

As mentioned above (Section 3.1), we formulate the depth estimation as a denoising process $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c})$, which iteratively refines depth latent $\boldsymbol{x_t}$ and improves the prediction accuracy, guided by the visual information available in the input image. Specifically, it is achieved by neural network model $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ which takes visual condition $\boldsymbol{c}$ and current depth latent $\boldsymbol{x}_t$ and predicts the distribution $\boldsymbol{x}_{t-1}$. The monocular visual condition $\boldsymbol{c} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times c}$ is constructed through multi-scale visual feature aggregation (Section 3.2). We introduce Monocular Conditioned Denoising Block (MCDB) as shown in Fig. 3 to achieve this process.

Since the depth prediction task normally requires low inference time for practical utilization, we design the denoising head in a **light-weighted** formation. The visual condition $\boldsymbol{c}$ is actually aggregated feature map with a lower resolution which has a strong local relation to the depth latent $\boldsymbol{x}_t$ to be denoised. We first use a local projection layer to upsample the condition c into the same shape with the depth latent $\boldsymbol{x}_t \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times d}$ while maintaining the local relation between features. The projected condition is directly fused with the depth latent $\boldsymbol{x}_t$ by performing element-wise summation through a CNN block and a self-attention layer. The fused depth latent is processed by a normal BottleNeck [23] CNN layer and channel-wise attention with the residual connection. The denoising output $\boldsymbol{x}_{t-1}$ is calculated by applying DDIM [54] inference process according to prefixed diffusion schedule β, α on model outputs.

3.4 Diffusion-Denosing Process

The diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ and denoising process $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c})$ are respectively defined in Eq. 1, and Eq. 3. Trainable parameters are mainly the conditioned denoising model $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ and visual feature extractors defined above. The model is trained by minimizing the L_{ddim} loss between diffusion results and denoising prediction in Eq. 4.

$$L_{\text{ddim}} = \left\| \boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) \right\|^2 \tag{4}$$

where diffusion result $\boldsymbol{x}_{t-1}$ could be calcuated through diffusion process defined in Eq. 1 by sampling set of t. It actually supervises the depth of the latent at each step after refinement by reversing the diffusion process.

Depth Latent Space Many of the previous constraint-based or classification-based methods [5,34] are not good at generating depth maps in high resolution. We employ a similar structure with latent diffusion [45], where both diffusion processes $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ and denoising process $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c})$ are performed in encoded latent depth space. The refined depth latent $\boldsymbol{x}_0 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times d}$ with latent dimension d is transferred to depth estimation $\boldsymbol{de} \in \mathbb{R}^{H\times W\times 1}$ through a depth decoder. The depth decoder is composed of sequentially connected 1×1 convolution, 3×3 de-convolution, 3×3 convolution, and a Sigmoid [39] activation function. The depth is calculated through Eq. 5,

$$\boldsymbol{de} = 1/\text{sig}(\boldsymbol{x}_0).\text{clamp}(\eta) - 1, \tag{5}$$

where η is the max output range. We set $\eta = 1e^6$ for both indoor and outdoor scenarios. Considering the sparsity in GT depth $\hat{\boldsymbol{de}}$, we use a BottleNeck CNN block with channel dimension d and kernel size 1×1 to encode the depth GT into depth latent $\hat{\boldsymbol{x}}_0$. The decoder and encoder are trained directly in end-to-end formation by minimizing the direct pixel-wise depth loss defined in Eq. 6,

$$L_{\text{pixel}} = \sqrt{\frac{1}{T}\sum_i \boldsymbol{\delta}_i^2 + \frac{\lambda}{T^2}(\sum_i \boldsymbol{\delta}_i)^2}, \tag{6}$$

where $\boldsymbol{\delta}_i = \hat{\boldsymbol{de}} - \boldsymbol{de}$ is the pixel-wise depth error on valid pixels, λ is set to 0.85 [33] for all experiments. T is the total number of valid pixels. The supervision is also applied to both latent spaces through L2 loss between encoded GT latent $\hat{\boldsymbol{x}}_0$ and depth latent $\boldsymbol{x}_0$ through a valid mask as defined in Eq. 7,

$$L_{\text{latent}} = \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}_0}\|^2 \tag{7}$$

The DiffusionDepth is trained by combining losses through a weighted sum and minimizing the L defined in Eq. 8,

$$L = \lambda_1 L_{ddim} + \lambda_2 L_{pixel} + \lambda_3 L_{latent}. \tag{8}$$

Self-Diffusion One severe problem of adopting generative methods into depth prediction is the **sparse ground truth** (GT) depth value problem, which is prevalent in outdoor scenarios where only a fraction of pixels have GT depth values (typically around $3.75-5\%$ in datasets such as KITTI depth [18]). This sparsity can lead to mode collapse during normal generative training. To tackle this issue, DiffusionDepth introduces a self-diffusion process. Rather than directly diffusing on the encoded sparse GT depth in latent space, the model gradually adds noise to the refined depth latent $\boldsymbol{x}_0$ from the current denoising output. With the help of random crop, jitter, and flip augmentation in training,

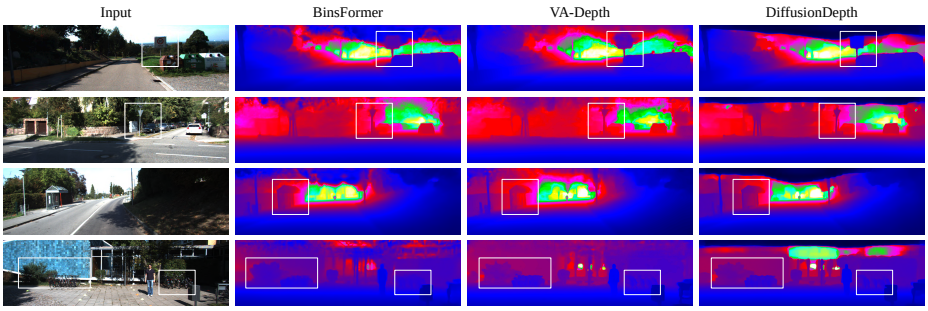


Fig. 4: Qualitative comparison of proposed DiffusionDepth on the KITTI outdoor driving scenarios against two representative methods, BinsFormer (classification-regression based) and VA-Depth (Variational Refine). We highlight the details with white boxes. The visualization is from the best online results for a fair comparison.

this process allows the model to 'organize' the entire depth map instead of just regressing on known parts, which largely improves the visual quality of the depth prediction. According to our experiments, for indoor sceneries with dense GT values, diffusion on either refined depth or GT depth is feasible.

4 Experiment

4.1 Experimental Setup

Dataset We conduct detailed experiments on outdoor and indoor scenarios to report an overall evaluation of the proposed DiffusionDepth and its properties.

KITTI dataset is captured from outdoor with driving vehicles [18] with depth range 0-100m. The image resolution is around 1216 $\times$ 352 pixels with sparse GT depth (density 3.75% to 5%). We evaluate on both Eigen split [16] with 23488 training image pairs and 697 testing images and official split [18] with 42949 training image pairs, 1000 validation images, and 500 testing images.

NYU-Depth-V2 dataset is collected from indoor scenes at a resolution of 640 $\times$ 480 pixels [38] and dense depth GT (density $>$ 95%). Following prior works, we adopt the official split and the dataset processed by Lee *et al*. [32], which contains 24231 training images and 654 testing images.

Implementation Details DiffusionDepth is implemented with the Pytorch [40] framework. We train the entire model with batch size 16 for 30 epochs iterations on a single node with 8 NVIDIA A100 40G GPUs. We utilize the AdamW optimizer [29] with (β_1, β_2, w) = (0.9, 0.999, 0.01), where w is the weight decay. The linear learning rate warm-up strategy is applied for the first 15% iterations. The cosine annealing learning rate strategy is adopted for the learning rate decay from the initial learning rate of $1e-4$ to $1e-8$. We use L1 and L2 pixel-wise depth loss at the first 50% training iterations as auxiliary subversion. For the KITTI dataset, we sequentially utilize the random crop with size 706×352, color

jitter with various lightness saturation, random scale from 1.0 to 1.5 times, and random flip for training data augmentation. For the NYU-Depth-V2 dataset, we use the same augmentation with the random crop with size 512×340.

Augmentation To prevent overfitting and improve the model's ability to refine image details, we apply various data augmentation techniques. On the KITTI dataset, we perform a sequence of random crops to 706×352, color jittering, scaling between 1.0 to 1.5 times, and horizontal flipping. For the NYU-Depth-V2 dataset, the random crop size is adjusted to 512×340. Additionally, we randomly modify brightness, contrast, saturation, and hue to mimic diverse lighting conditions, enhancing the network's robustness to color variations. Horizontal flips introduce orientation diversity, while random rotations between -5 to 5 degrees provide varied angles and perspectives.

Visual Condition DiffusionDepth is compatible with any backbone that could extract multi-scale features. Here, we respectively evaluate our model on the standard convolution-based ResNet [23] backbones and transformer-based Swin [37] backbones. We employ hierarchical aggregation and heterogeneous interaction (HAHI [33]) neck to enhance features between scales and feature pyramid neck [35] to aggregate features into monocular visual condition. The visual condition dimension is equal to the last layer of the neck. We respectively use channel dimensions $[64, 128, 256, 512]$ and $[192, 384, 768, 1536]$ for ResNet and Transformer backbones.

Diffusiong Head We use the improved sampling process [54] with 1000 diffusion steps for training and 20 inference steps for inference. The learning rate of the diffusion head is 10 times larger than the backbone parameters. The dimension d of the encoded depth latent is 16 with shape $\frac{H}{2}, \frac{W}{2}, d$, we conduct detailed ablation to illustrate different inference settings. The max depth value of the decoder is $1e6$ for all experiments.

4.2 Benchmark Comparison with SOTA Methods

Offline Evaluation on KITTI Dataset We first illustrate the efficiency of DiffusionDepth by comparing it with previous state-of-the-art (SOTA) models on KITTI offline Eigen split [16] with an evaluation range of 0-80m and report results in Tab. 1. It is observed that DiffusionDepth respectively reaches 0.050 absolute error and 2.016 RMSE on the evaluation, which exceeds the current SOTA results URCDC-Depth (RSME 2.032) and VA-Depth (RSME 2.090). On official offline split [18] in Tab. 1 with evaluation range 0-50m, our proposal reaches 0.041 absolute related error and 1.452 RMSE on the evaluation, which largely outperforms the current best URCDC-Depth (0.049 rel and 1.528 RMSE) by a large margin. This suggests that DiffusionDepth has even better performance in estimating depth with a closer depth range which is valuable for practical usage. This property is rational since the diffusion approach brings a stronger generative ability to the task.

Table 1: Evaluation metrics on the offline KITTI dataset, Eigen split [16] and official offline split [18]. The metrics of comparison metrics come from corresponding original papers. "-" indicates not applicable. The best results are highlighted in bold.

Method	Cap	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
Eigen Split [16], evaluation range 0-80m								
DORN [17]	0-80m	0.072	0.307	2.727	0.120	0.932	0.984	0.994
BTS [31]	0-80m	0.061	0.261	2.834	0.099	0.954	0.992	0.998
TransDepth [61]	0-80m	0.064	0.252	2.755	0.098	0.956	0.994	**0.999**
Adabins [5]	0-80m	0.058	0.190	2.360	0.088	0.964	0.995	**0.999**
P3Depth [41]	0-80m	0.071	0.270	2.842	0.103	0.953	0.993	0.998
DepthFormer [33]	0-80m	0.052	0.158	2.143	0.079	0.975	0.997	**0.999**
NeWCRFs [63]	0-80m	0.052	0.155	2.129	0.079	0.974	0.997	**0.999**
PixelFormer [1]	0-80m	0.051	0.149	2.081	0.077	0.976	0.997	**0.999**
BinsFormer [34]	0-80m	0.052	0.151	2.098	0.079	0.974	0.997	**0.999**
VA-Depth [36]	0-80m	**0.050**	-	2.090	0.079	**0.977**	0.997	-
URCDC-Depth [47]	0-80m	**0.050**	0.142	2.032	0.076	**0.977**	0.997	**0.999**
DiffusionDepth (ours)	0-80m	**0.050**	**0.141**	**2.016**	**0.074**	**0.977**	**0.998**	**0.999**
Official Offline Split [18], evaluation range 0-50m								
BTS [31]	0-50m	0.058	0.183	1.995	0.090	0.962	0.994	0.999
PWA [32]	0-50m	0.057	0.161	1.872	0.087	0.965	0.995	0.999
TransDepth [61]	0-50m	0.061	0.185	1.992	0.091	0.963	0.995	0.999
P3Depth [41]	0-50m	0.055	0.130	1.651	0.081	0.974	0.997	0.999
URCDC-Depth [47]	0-50m	0.049	0.108	1.528	0.072	0.981	0.998	**1.000**
VPD [67]	0-50m	0.132	-	3.262	-	0.893	0.932	0.991
DiffusionDepth (ours)	0-50m	**0.041**	**0.103**	**1.418**	**0.069**	**0.986**	**0.999**	**1.000**

Online Evaluation on KITTI Benchmark Online evaluation is conducted by submitting results to the official servers for **KITTI Online evaluation** on 500 unseen images. The results are shown in Tab.2, where the proposed model slightly underperformed compared to VA-depth and URCDC-depth. However, it's important to note that our approach only uses aggregated visual features as guidance and doesn't incorporate complicated long-range attention or constraint priors like these SOTA methods. As our diffusion head is compatible with these advanced depth feature extraction techniques, incorporating them could further improve the performance of our approach.

Table 2: Quantitative depth comparison on the official **online server of the KITTI dataset**.

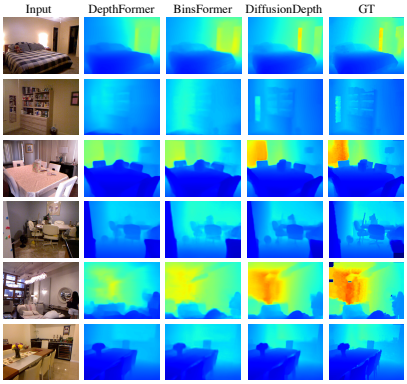
Method	SILog ↓	sqErr. ↓	absErr. ↓	iRMSE↓
BTS [31]	11.67	9.04	2.21	12.23
BANet [2]	11.61	9.38	2.29	12.23
PackNet-SAN [20]	11.54	9.12	2.35	12.38
PWA [32]	11.45	9.05	2.30	12.32
NeWCRFs [63]	10.39	8.37	1.83	11.03
PixelFormer [34]	10.28	8.16	1.82	10.84
BinsFormer [34]	10.14	8.23	1.69	10.90
P3Depth [41]	12.82	9.92	2.53	13.71
URCDC-Depth [47]	10.03	8.24	1.74	10.71
VA-Depth [36]	9.84	7.96	1.66	**10.44**
IE-Bins [36]	9.63	7.82	1.60	10.68
ND-Depth [48]	**9.62**	**7.75**	**1.59**	10.62
DiffusionDepth	9.85	8.06	1.64	10.58

Qualitative Comparison on KITTI Datset is reported in Fig. 4. Here, we show the improved visual quality brought by the diffusion-denoising process. The clarity of the objects regarding both the edges and the shape has been signifi-

Table 3: Quantitative depth comparison on the NYU-Depth-v2 dataset.

Method	Rel. ↓	RMSE ↓	$\log_{10}$ ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
VNL [62]	0.108	0.416	0.048	0.875	0.976	0.994
BTS [31]	0.113	0.407	0.049	0.871	0.977	0.995
PWA [32]	0.105	0.374	0.045	0.892	0.985	0.997
TransDepth [61]	0.106	0.365	0.045	0.900	0.983	0.996
Adabins [5]	0.103	0.364	0.044	0.903	0.984	0.997
P3Depth [41]	0.104	0.356	0.043	0.898	0.981	0.996
DepthFormer [33]	0.096	0.339	0.041	0.921	0.989	0.998
NeWCRFs [41]	0.095	0.334	0.041	0.922	**0.992**	0.998
PixelFormer [1]	0.090	0.322	0.039	0.929	0.991	0.998
BinsFormer [34]	0.094	0.330	0.040	0.925	0.989	0.997
URCDC-Depth [47]	0.088	0.316	0.038	0.933	**0.992**	0.998
VA-Depth [36]	0.086	0.304	-	0.937	**0.992**	-
VPD [67]	**0.069**	**0.254**	**0.036**	**0.964**	**0.995**	**0.999**
DiffusiongDepth	0.085	0.295	**0.036**	0.939	**0.992**	**0.999**

Fig. 5: Qualitative depth results on the NYU-Depth-v2 dataset.

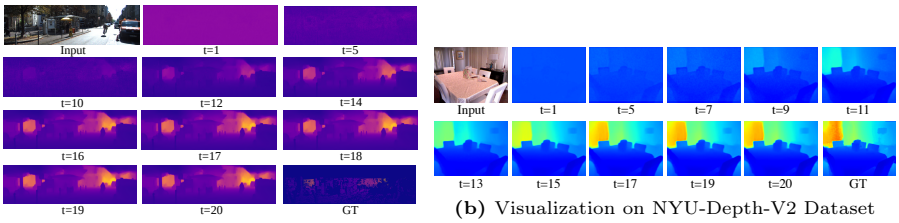


cantly improved. For example, on the first row, both BinsFormer and VA-Depth have significant blur on the signpost. Diffusion depth predicts a sharp and accurate shape for it. Classification-based methods are suffered from visible noise in the depth map. As we mentioned above, one significant advantage of introducing the diffusion-denoising approach is that we could acquire a highly-detailed depth map with good visual quality and clear shapes for practical utilization. The proposed diffusion head could also be combined with other methods, such as bins to improve the visual quality.

Evaluation on NYU-Depth-V2 Dataset We evaluate the proposed DiffusionDepth on the NYU-Depth-V2 dataset [49] to demonstrate the effectiveness of our proposal. The results are reported in Table 3. It suggests that the diffusion-denoising approach has even higher improvement than outdoor scenarios, where it respectively achieves 0.085 absolute related error and 0.295 RSME score which exceeds the previous SOTA. We think this phenomenon is rational since indoor scenarios mostly have dense depth GT values, which is naturally suitable for generative models. It is noted that for datasets with dense GTs, direct diffusion on GT value is also feasible with comparable results. To give a more direct illustration of the proposed DiffusionDepth, **we display qualitative depth comparisons in Fig. 5**.

4.3 Ablation Study

Qualitative Study of Denoising Process To give an intuitive understanding of how the denoising process refines the depth prediction step by step, we visualize the denoising process in Fig. 6. It shows that the process first initializes ($t < 10$) the shapes and edges from random depth distribution. Then the guided denoising model refines the depth values and corrects distance relations step by



(a) Visualization on KITTI Dataset

(b) Visualization on NYU-Depth-V2 Dataset

Fig. 6: Visualization of the denoising process with 20 inference steps, where t denotes the current step. It gives an intuitive illustration of how the depth estimation is refined iteratively.

step. This process is more like first recognizing the shape of the desired scenery and then considering the depth relations between these objects with visual clues. The learning process is impressive. One interesting problem is that the denoising process is even faster in more complicated outdoor scenarios (KITTI). Although the mediate results are slightly lower, the denoising steps larger than 15 could achieve competitive results on the KITTI dataset.

Denoising Inference To further reveal the properties of using different inference steps, we conduct an ablation study on different inference settings. Lower inference steps could benefit the practical usage with lower GPU memory consumption and faster inference speed. We consider two settings, 1) train with 1000 diffusion steps and 20 inference steps and change the inference step, 2) train with different inference steps. The ablation is conducted on the NYU-Depth-V2 dataset, where the variations of the metrics are reported in Tab. 4. We fix diffusion to 1000 steps throughout the training. Directly changing the inference steps will lead to a severe performance drop. This observation is different from the diffusion approach on detection boxes [7], which could change inference steps once the model is trained. We think this observation is rational since directly denoising on the highly detailed depth map is closer to a generative task, rather than denoising on anchor boxes. However, we prove the feasibility of accelerating the inference by directly training the model with the desired inference setting, which only shows a slight performance drop.

Inference Speed Although the inference speed is one shortage of diffusion-based models, as shown by Fig. 7, DiffusionDepth could reach 14 FPS and 5 FPS (Frame Per Second) respectively on ResNet Backbones and Swin Backbones with 20 inference steps, which is feasible for practical usage. With acceleration, the speed could be faster.

Diffusion As we mentioned, we employ a self-diffusion formation to add noise on refined depth latent rather than directly on the sparse depth. In this sector, we conduct a detailed ablation study on whether diffusing on GT sparse depth or

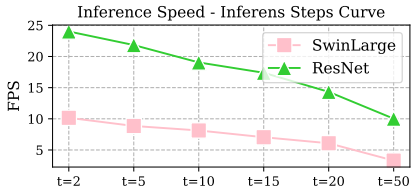


Fig. 7: Inference speed with RTX 3090 GPU on KITTI dataset.

Table 4: Ablation study on different inference settings on NYU-Depth-V2 dataset, t denotes the inference step.

Method	Rel. ↓	RMSE ↓	MAE ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
t=20	**0.086**	**0.298**	**0.166**	**0.937**	**0.992**	**0.999**
Directly change inference without training.						
t=15	0.1178	0.4552	0.3294	0.8644	0.9730	0.9928
t=10	0.1821	0.7506	0.5893	0.6475	0.9289	0.9853
t=5	0.2873	1.1750	0.9451	0.3803	0.7085	0.8825
t=2	0.3620	1.4328	1.1616	0.2808	0.5504	0.7699
Train with different inference steps.						
t=15	0.1034	0.3648	0.238	0.9022	0.9834	0.993
t=10	0.1069	0.3708	0.278	0.8815	0.9812	0.993
t=5	0.1108	0.4366	0.294	0.8345	0.9644	0.992
t=2	0.1308	0.5678	0.387	0.8016	0.9516	0.990

diffusing on the refined dense depth latent. We compare the two different ways of diffusion by comparing different diffusion methods on both KITTI (outdoor) and NYU-Depth-V2 (indoor) datasets. The results are reported in Tab. 5. It suggests that, under outdoor scenarios, sparse depth GT will lead to severe mode collapse, where diffusing on sparse GT on the KITTI dataset only reaches RMSE 12.3772 which largely falls behind self-diffusion RMSE 1.4523. For indoor scenarios, both diffusion approaches could achieve competitive results on the NYU-Depth-V2 dataset with dense GT depth values.

Depth Latent Space Analysis The ablation evaluation of different depth encoder-decoder structures with differing down-sampling rates is summarized in Table 6. While we keep the best encoder structure (Swin+HAI) and comparing the down-sampling rate on the depth latent space, it indicates that encoder-decoder pairs with both $\times 4$ and $\times 2$ down-sampling can deliver commendable performance. Notably, a depth latent space with a higher resolution marginally surpasses its lower-resolution counterparts.

Adaptability Across Different Visual Conditions This ablation study compares the performance of the proposed model given different visual encoders or visual conditions. The adaptability of our DiffusionDepth model is demonstrated in Table 6, where it exhibits proficiency with both convolutional neural

Table 5: Ablation on different diffusion methods. Both methods are evaluated on offline official splits.

Method	Rel. ↓	RMSE ↓	MAE ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
KITTI Dataset						
Refined	0.0410	1.4523	0.7364	0.986	0.999	1.000
GT	0.3480	12.3772	7.0154	0.4920	0.6894	0.8074
NYU-Depth-V2 Dataset						
Refined	0.0862	0.2983	0.1665	0.937	0.992	0.999
GT	0.0940	0.3041	0.1742	0.932	0.992	0.999

networks (CNNs) such as ResNet34 and ResNet50, and transformer-based architectures like Swin. Furthermore, preliminary tests reveal that our model can effectively integrate visual depth cues, exemplified by Bins [34], to serve as denoising guidance, thereby maintaining robust performance across different visual conditions.

Table 6: Ablation on different depth encoder-decoders and visual conditions on KITTI Dataset, official offline split (0-50m), where **DSR** denotes the down-sampling rate of the encoded depth latent.

Condition	DSR	Rel. ↓	RMSE ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
Depth Latent Space (Down-Sampling Rate)						
Swin+HAHI	**×2**	0.0410	1.4523	0.986	0.999	1.000
Swin+HAHI	×4	0.0445	1.508	0.985	0.999	1.000
Visual Conditions (Backbones)						
Res34+FPN	×2	0.0554	1.7902	0.978	0.992	0.999
Res50+FPN	×2	0.0532	1.7124	0.978	0.993	0.999
Swin+FPN	×2	0.0458	1.5569	0.985	0.998	0.999
Swin+Bins	×2	0.0468	1.5832	0.985	0.998	0.999
Swin+HAHI	×2	0.0410	1.4523	0.986	0.999	1.000

5 Conclusion

In this paper, we reformulate the monocular depth estimation problem as a diffusion-denoising approach. The iterative refinement of the depth latent helps DiffusionDepth generate accurate and highly detailed depth maps. Experimental results suggest the proposed model reaches competitive performance on both online open benchmarks and offline evaluations under both indoor and outdoor scenarios. This paper verifies the feasibility of introducing a diffusion-denoising model into 3D perception tasks. Comprehensive detailed ablation studies are provided to help understand each component of this framework.

References

1. Agarwal, A., Arora, C.: Attention attention everywhere: Monocular depth prediction with skip attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5861–5870 (2023) 4, 10, 11
2. Aich, S., Vianney, J.M.U., Islam, M.A., Liu, M.K.B.: Bidirectional attention network for monocular depth estimation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 11746–11752. IEEE (2021) 3, 10
3. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021) 4
4. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=SlxSY2UZQT 4
5. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Computer Vision and Pattern Recognition (CVPR). pp. 4009–4018 (2021) 2, 4, 7, 10, 11
6. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Denoising pretraining for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4175–4186 (2022) 4
7. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. arXiv preprint arXiv:2211.09788 (2022) 2, 4, 12
8. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. arXiv preprint arXiv:2210.06366 (2022) 2, 4
9. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. Pattern Analysis and Machine Intelligence (PAMI) (2023) 1
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems (NIPS/NeurIPS) **34**, 8780–8794 (2021) 4
11. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: Computer Vision and Pattern Recognition (CVPR). pp. 4738–4747 (2019) 2
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 5
13. Duan, Y., Feng, C.: Learning internal dense but external sparse structures of deep convolutional neural network. In: Artificial Neural Networks and Machine Learning–ICANN 2019: Deep Learning: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part II 28. pp. 247–262. Springer (2019) 1
14. Duan, Y., Guo, X., Zhu, Z., Wang, Z., Wang, Y.K., Lin, C.T.: Maskfuser: Masked fusion of joint multi-modal tokenization for end-to-end autonomous driving. arXiv preprint arXiv:2405.07573 (2024) 1
15. Duan, Y., Zhang, Q., Xu, R.: Prompting multi-modal tokens to enhance end-to-end autonomous driving imitation learning with llms. arXiv preprint arXiv:2404.04869 (2024) 4

16. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems (NIPS/NeurIPS). pp. 2366–2374 (2014) 3, 8, 9, 10
17. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2002–2011 (2018) 2, 3, 10
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 3, 7, 8, 9, 10
19. Graikos, A., Malkin, N., Jojic, N., Samaras, D.: Diffusion models as plug-and-play priors. arXiv preprint arXiv:2206.09012 (2022) 4
20. Guizilini, V., Ambrus, R., Burgard, W., Gaidon, A.: Sparse auxiliary networks for unified monocular depth prediction and completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11078–11088 (2021) 10
21. Guo, X., Li, H., Yi, S., Ren, J., Wang, X.: Learning monocular depth by distilling cross-domain stereo networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 484–500 (2018) 3
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 1
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) 5, 6, 9
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (NIPS/NeurIPS) **33**, 6840–6851 (2020) 4
25. Hoogeboom, E., Garcia Satorras, V., Vignac, C., Welling, M.: Equivariant diffusion for molecule generation in 3d. arXiv e-prints pp. arXiv–2203 (2022) 2
26. Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., Heikkilä, J.: Guiding monocular depth estimation using depth-attention volume. In: European Conference on Computer Vision (ECCV). pp. 581–597. Springer (2020) 1, 3
27. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4756–4765 (2020) 2
28. Kim, B., Oh, Y., Ye, J.C.: Diffusion adversarial representation learning for self-supervised vessel segmentation. arXiv preprint arXiv:2209.14566 (2022) 4
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
30. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019) 1, 3
31. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019) 10, 11
32. Lee, S., Lee, J., Kim, B., Yi, E., Kim, J.: Patch-wise attention network for monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1873–1881 (2021) 8, 10, 11
33. Li, Z., Chen, Z., Liu, X., Jiang, J.: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. arXiv preprint arXiv:2203.14211 (2022) 5, 7, 9, 10, 11

34. Li, Z., Wang, X., Liu, X., Jiang, J.: Binsformer: Revisiting adaptive bins for monocular depth estimation. arXiv preprint arXiv:2204.00987 (2022) 2, 4, 7, 10, 11, 14
35. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017) 2, 5, 9
36. Liu, C., Kumar, S., Gu, S., Timofte, R., Van Gool, L.: Va-depthnet: A variational approach to single image depth prediction. arXiv preprint arXiv:2302.06556 (2023) 4, 10, 11
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021) 5, 9
38. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision (ECCV) (2012) 8
39. Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S.: Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378 (2018) 7
40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems (NIPS/NeurIPS) **32** (2019) 8
41. Patil, V., Sakaridis, C., Liniger, A., Van Gool, L.: P3depth: Monocular depth estimation with a piecewise planarity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1610–1621 (2022) 2, 4, 10, 11
42. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 283–291 (2018) 3
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: International Conference on Computer Vision (ICCV) (ICCV). pp. 12179–12188 (2021) 4
44. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016) 3
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) 2, 7
46. Saxena, A., Chung, S.H., Ng, A.Y., et al.: Learning depth from single monocular images. In: Advances in neural information processing systems (NIPS/NeurIPS). vol. 18, pp. 1–8 (2005) 3
47. Shao, S., Pei, Z., Chen, W., Li, R., Liu, Z., Li, Z.: Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. arXiv preprint arXiv:2302.08149 (2023) 1, 4, 10, 11
48. Shao, S., Pei, Z., Chen, W., Wu, X., Li, Z.: Nddepth: Normal-distance assisted monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7931–7940 (2023) 10
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision. pp. 746–760. Springer (2012) 11

50. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML). pp. 2256–2265. PMLR (2015) 4
51. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 4
52. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=St1giarCHLP 4
53. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems (NIPS/NeurIPS) **32** (2019) 4
54. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems (NIPS/NeurIPS) **33**, 12438–12448 (2020) 5, 6, 9
55. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=PxTIG12RRHS 4
56. Sun, J., Zhang, Q., Duan, Y., Jiang, X., Cheng, C., Xu, R.: Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning. arXiv preprint arXiv:2309.11359 (2023) 1
57. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). pp. 6105–6114. PMLR (2019) 1, 5
58. Trippe, B.L., Yim, J., Tischer, D., Broderick, T., Baker, D., Barzilay, R., Jaakkola, T.: Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. arXiv preprint arXiv:2206.04119 (2022) 2
59. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. arXiv preprint arXiv:2112.03145 (2021) 4
60. Yan, J., Zhao, H., Bu, P., Jin, Y.: Channel-wise attention-based network for self-supervised monocular depth estimation. In: 2021 International Conference on 3D vision (3DV). pp. 464–473. IEEE (2021) 3
61. Yang, G., Tang, H., Ding, M., Sebe, N., Ricci, E.: Transformer-based attention networks for continuous pixel-wise prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 16269–16279 (October 2021) 10, 11
62. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5684–5693 (2019) 11
63. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3916–3925 (June 2022) 10
64. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: New crfs: Neural window fully-connected crfs for monocular depth estimation. CoRR **abs/2203.01502** (2022). https://doi.org/10.48550/arXiv.2203.01502, https://doi.org/10.48550/arXiv.2203.01502 2, 4
65. Zhang, Q., Cui, P., Yan, D., Sun, J., Duan, Y., Zhang, A., Xu, R.: Whole-body humanoid robot locomotion with human reference. arXiv preprint arXiv:2402.18294 (2024) 1

66. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4106–4115 (2019) 1, 3
67. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception (2023), https://arxiv.org/abs/2303.02153 10, 11
68. Zheng, W., Song, R., Guo, X., Chen, L.: Genad: Generative end-to-end autonomous driving. arXiv preprint arXiv:2402.11502 (2024) 1