Scalable Autoregressive Monocular Depth Estimation

Jinhong Wang[1] Jian Liu[2] Dongqi Tang[2] Weiqiang Wang[2] Wentong Li[1]
Danny Chen[3] Jintai Chen[4†] Jian Wu[1†]

[1]ZJU [2]Ant Group [3]University of Notre Dame [4]HKUST(Guangzhou)
† Corresponding Authors

Abstract

This paper proposes a new autoregressive model as an effective and scalable monocular depth estimator. Our idea is simple: We tackle the monocular depth estimation (MDE) task with an autoregressive prediction paradigm, based on two core designs. First, our depth autoregressive model (DAR) treats the depth map of different resolutions as a set of tokens, and conducts the `low-to-high` *resolution autoregressive objective with a patch-wise causal mask. Second, our DAR recursively discretizes the entire depth range into more compact intervals, and attains the* `coarse-to-fine` *granularity autoregressive objective in an ordinal-regression manner. By coupling these two autoregressive objectives, our DAR establishes new state-of-the-art (SOTA) on* `KITTI` *and* `NYU Depth v2` *by clear margins. Further, our scalable approach allows us to scale the model up to 2.0B and achieve the best RMSE of 1.799 on the* `KITTI` *dataset (5% improvement) compared to 1.896 by the current SOTA (Depth Anything). DAR further showcases zero-shot generalization ability on unseen datasets. These results suggest that DAR yields superior performance with an autoregressive prediction paradigm, providing a promising approach to equip modern autoregressive large models (e.g., GPT-4o) with depth estimation capabilities. Project page:* `https://depth-ar.github.io/`.

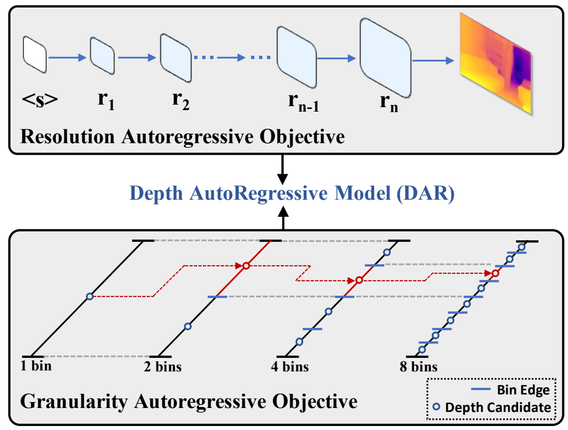

Figure 1. We exploit two “order” properties of the MDE task that can be transformed into two autoregressive objectives. **(a) Resolution autoregressive objective:** The generation of depth maps can follow a resolution order from low to high. For each step of the resolution autoregressive process, the Transformer predicts the next higher-resolution token map conditioned on all the previous ones. **(b) Granularity autoregressive objective:** The range of depth values is ordered, from 0 to specific max values. For each step of the granularity autoregressive process, we increase exponentially the number of bins (e.g., doubling the bin number), and utilize the previous predictions to predict a more refined depth with a smaller and more refined granularity. Our proposed DAR aims to perform these two autoregressive processes simultaneously.

1. Introduction

The Monocular Depth Estimation (MDE) task aims to predict per pixel depth from a single RGB image, which plays a crucial role in scene understanding and reconstruction [9, 21]. This task has broad applications, including autonomous driving [58], robotics [23], augmented reality [37], medical endoscopic surgery [36], etc. Most deep learning (DL) based methods typically followed top-down-bottom-up encoder-decoder architectures [2, 3, 11, 12, 22, 25, 27, 31, 34, 38, 47, 62, 64, 66], which extract and fuse low-level and high-level features for depth estimation.

Recently, autoregressive (AR) architectures have demonstrated strong generalization capabilities and significant scalability across a multitude of tasks: (i) Generalization: AR models achieve remarkable zero- and few-shot performance on previously unseen datasets, and exhibit substantial flexibility in adapting to diverse downstream tasks [7, 42]. (ii) Scalability: As suggested by scaling laws [17, 24], AR models allow flexible model size scaling for optimal performance across various practical applications. Recent advances in large language models (LLMs) [7, 41, 42] and multi-modal large language models (MLLMs) (e.g., GPT-4 and LLaVA [1, 35]), built upon AR architectures, achieved outstanding results in a wide range of tasks, including text-

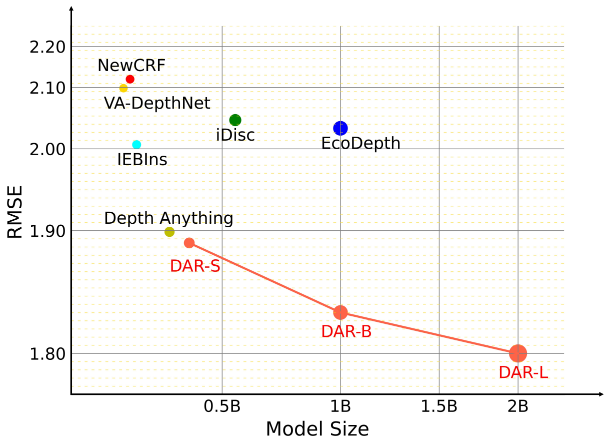

Figure 2. RMSE performances (↓) *vs.* model sizes on the `KITTI` dataset. Our DAR shows strong scalability and achieves better performance-efficiency trade-off among cutting-edge methods.

to-image generation [13, 28, 53], video generation [60, 61], object detection [10] and tracking [59], among others.

This naturally leads to an interesting question: *Can an autoregressive model be developed for the MDE task?*

However, autoregressive modeling relies on a well-organized sequential data formation, where each step's prediction is logically connected to the previous steps. While this sequential dependency may be common in other tasks, it is not intuitively align with the MDE requirements, where meaningful sequential prediction targets are less apparent.

In this paper, we introduce a simple, effective, and scalable depth autoregressive (DAR) framework for MDE. Our new approach leverages two key ordering properties of MDE, and we incorporate them as autoregressive objectives, as shown in Fig. 1. The first property is "depth map resolution": We generate depth maps at varying resolutions, ordered from low to high, and treat the sequences of depth maps of different resolutions as prediction targets [53]. This approach reframes depth map generation as a `low-to-high` resolution autoregressive objective, where each step generates a higher-resolution depth map based on previous predictions. The second property is "depth values", which inherently exist in a continuous space. Typically, known methods discretize the depth values into several intervals (or bins) to formulate an ordinal regression task [14]. By further discretizing the depth range into progressively finer intervals, we recast MDE as a `coarse-to-fine` autoregressive objective.

For the resolution autoregressive objective, we develop a depth autoregressive Transformer that predicts the next-resolution depth map based on its prefix predictions with the patch-wise causal mask. To achieve the granularity autoregressive objective, we propose a novel binning strategy called Multiway Tree Bins (MTBin), which queries the corresponding bin using prior depth predictions and then recursively refines each bin into sub-bins with error tolerance for subsequent autoregressive steps. Importantly, these bins are used to not only compute the final depth values but also embed granularity information into the latent token maps, effectively guiding the depth map generation process.

Comprehensive experiments show that our DAR achieves state-of-the-art (SOTA) performance on the `KITTI` and `NYU Depth v2` datasets. With a similar model size, DAR outperforms current SOTA (*Depth Anything*) in all the metrics, particularly by 3% in the RMSE metric on `KITTI`. DAR also exhibits scaling laws akin to those witnessed in LLMs, and its size can be easily scaled up to 2.0B, which establishes a new SOTA performance, as shown in Fig. 2. Lastly, we showcase DAR's zero-shot generalization capabilities on unseen datasets. These results further validate the generalizability and scalability of DAR.

The contributions of this work are multi-fold:

(1) To the best of our knowledge, we introduce the first autoregressive model for monocular depth estimation (MDE), called DAR — a simple, effective, and scalable framework. Our key insight lies in transforming two ordered properties in MDE, depth map resolution and granularity, into autoregressive objectives.

(2) We reformulate the low- and high-level feature fusion process in existing encoder-decoder models as a `low-to-high` resolution autoregressive objective. We introduce a new Depth autoregressive Transformer that uses the patch-wise causal mask and progressively generates depth maps at increasing resolutions, conditioned on tokens from the input RGB images.

(3) We propose a novel binning strategy called multiway tree bins (MTBin), tailored for the granularity autoregressive objective, which transforms MDE tasks into an autoregressive bin sequence prediction task. By using Bins Injection strategy, DAR effectively connects the resolution and granularity autoregressive processes.

(4) DAR establishes new state-of-the-art performance on the `KITTI` and `NYU Depth v2` datasets, and exhibits stronger zero-shot capability than `Depth Anything` [63]. A series of DAR models, ranging from 440M to 2.0B parameters, is developed based on decoder-only Transformer [42, 55]. The largest model achieves 0.205 RMSE and 0.982 δ_1 on `NYU Depth v2`, largely outperforming the existing methods.

2. Related Work

Monocular Depth Estimation (MDE). Monocular Depth Estimation has a long development of methods, ranging from traditional methods to deep learning (DL) techniques. Traditional methods relied on handcrafted features [20, 33] and used Markov Random Fields [46] to predict depth maps. However, they suffered limitations in handling complex scenes. Modern DL techniques approached the prob-

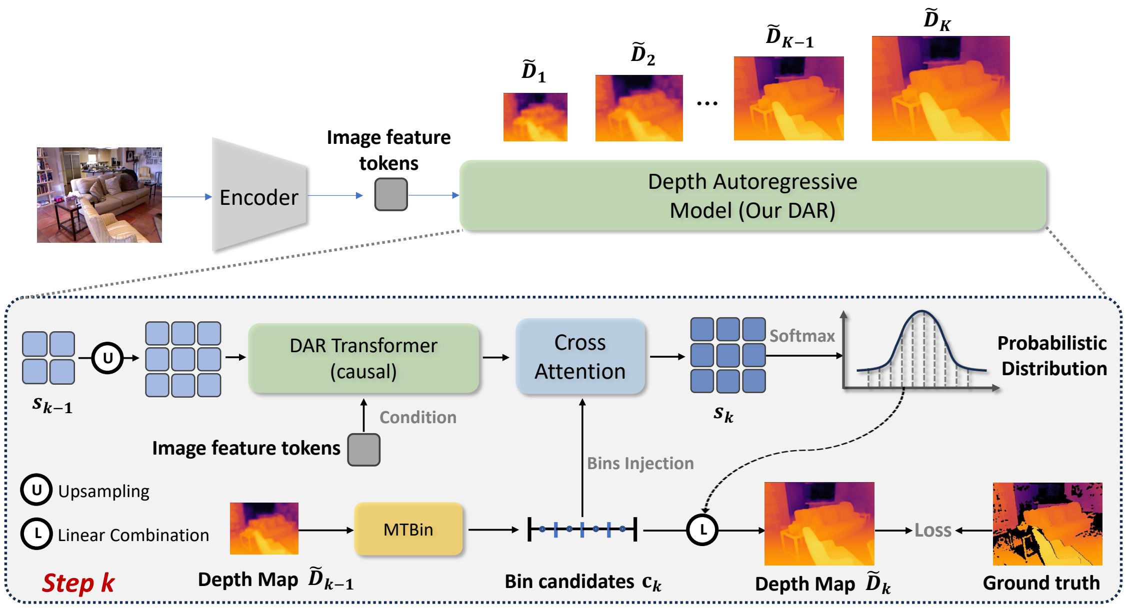

Figure 3. **An overview of DAR.** We begin with encoding the input RGB images into image tokens as the context condition. At each step, DAR Transformer with the patch-wise causal mask performs autoregressive predictions, that is, it allows the input token map (upsampled from the previous resolution token map r_{k-1}) to interact with only the prefix tokens and global image feature tokens for the next-resolution token map modeling. The output latent tokens are then sent to the ConvGRU module, which injects the prompts of new refined bin candidates $\mathbf{c}_k$ (generated by MTBin from the previous prediction $\tilde{D}_{k-1}$) for further granularity guidance and generates the next-resolution token map r_k. The new depth map $\tilde{D}_k$ is generated by a linear combination of the next-granularity bin candidates $\mathbf{c}_k$ and softmax value $\mathbf{p}_k$ of the next-resolution token map, achieving concurrently a resolution and granularity autoregressive evolution.

lem as a dense regression problem. Successive improvements mainly came from three fronts: model architecture, data-driven, and language guidance. **(1) Model architecture:** The main improvement came from the transformation of model backbones, from CNNs [12, 27, 34, 38, 64] to Transformers [2, 3, 31, 62] to the current diffusion models [11, 22, 25, 47, 66]. **(2) Data-driven methods:** The milestone data-driven method MiDas [43] proposed to train well-generalized models on large amounts of data by mixing different datasets, which achieved excellent zero-shot transfer performance. ZoeDepth [6] and Depth Anything [63] further leveraged the benefits of large-scale unlabeled data (62M) via self-supervised learning to obtain outstanding zero-shot generalization. **(3) Language-guided methods:** Benefited from the rich visual and text information of CLIP and other language-vision pre-training models, VPD and other methods [8, 39, 65, 66] used language description to facilitate depth estimation and achieved SOTA results on standard depth estimation datasets.

MDE as Ordinal Regression. Another main line of work is to treat MDE as an ordinal regression task [16, 18], which aims to predict labels on an ordinal scale. By discretizing the depth space into several bins, DORN [14] first proposed an ordinal regression network to predict discrete depth values. Since then, many studies [4, 32, 49] developed various depth binning strategies to discretize depth range and solve MDE as ordinal regression. In particular, Ord2Seq [56] treated ordinal regression as a label sequence task, and proposed an autoregressive network to predict the more refined labels progressively. Inspired by this, we transform MDE as an autoregressive prediction task from a bin perspective to achieve more granular prediction progressively.

Autoregressive Visual Generation. Many recent methods [13, 28, 45] explored the effectiveness of autoregressive models in the visual domain. VQGAN [13] proposed to use VQVAE [54] to conduct the autoregressive process in the latent space. It employed GPT-2 decoder-only Transformer to generate tokens in the raster-scan order. VQVAE-2 [45] and RQ-Transformer [28] also followed this raster-scan manner, but used extra scales or stacked codes. VAR [53] proposed text-to-image autoregressive generation via next-scale prediction, which transforms the entire image into a set of tokens and treats them as input to predict the next-scale target image. Inspired by this, we transform MDE as an autoregressive prediction task from a scale perspective to achieve larger resolution predictions progressively.

3. Method

3.1. Preliminaries

Ordinal Regression-based MDE. Some methods [4, 14] tackled MDE in an ordinal regression fashion, learning the

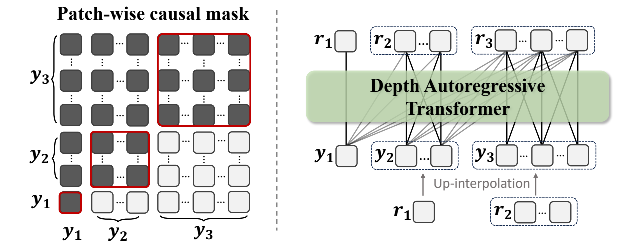

Figure 4. Illustrating the patch-wise causal mask for ensuring that the current token map can interact only with tokens from itself and the prefix token maps.

probabilistic distribution on each pixel and using a linear combination with depth candidates as the final depth prediction. Suppose we divide the depth range $[L, R]$ into N bins. Then the i-th bin's center $\mathbf{c}_i$ is $L + \frac{(i-0.5)}{N} * (R - L)$. For each pixel, the model will predict N Softmax scores $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, referring to the probabilities over the N bins. The final predicted depth value $\tilde{D}$ is calculated from the linear combination of Softmax scores and bin centers at this pixel, as:

$$\tilde{D} = \sum_{i=1}^{N} \mathbf{c}_i \mathbf{p}_i. \tag{1}$$

3.2. Overview

We propose a new Depth AutoRegressive (DAR) modeling approach to explore the potential of autoregressive models in dealing with the depth estimation task. We define the task as follows: Given an input RGB image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, where H and W are the height and width of $\mathcal{I}$ respectively, predict the depth map $\tilde{\mathcal{D}}$ of $\mathcal{I}$. Our model predicts the depth maps of different scales $\{\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_K\}$ progressively in an autoregressive process. That is, each depth map at step k is conditioned by the previous predictions, as:

$$p(\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_K) = \prod_{k=1}^{K} p_\theta(\tilde{\mathcal{D}}_k \mid \tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_{k-1}). \tag{2}$$

Our proposed autoregressive model DAR involves optimizing $p_\theta(\tilde{\mathcal{D}}_k \mid \tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_{k-1})$ over a dataset, and finally predicts depth map $\hat{\mathcal{D}} = \tilde{\mathcal{D}}_K$. Fig. 3 shows an overview of our proposed DAR. DAR involves two autoregressive objectives with ordinal properties: resolution and granularity. The former, resolution autoregressive objective, aims to predict the depth map from low resolution to high resolution; the latter, granularity autoregressive objective, aims to predict the depth map from coarse granularity to fine granularity. Specifically, DAR consists of four parts:

- **Image Encoder**: We apply an Image Encoder to extract RGB image features, which extracts image features into imaging tokens with latent representations.
- **DAR Transformer**: Our DAR Transformer can progressively predict different resolution token maps conditioned on extracted RGB imaging tokens via the patch-wise causal mask.
- **Multiway Tree Bins (MTBin)**: We develop a Multiway Tree Bins strategy to transform the depth range of each pixel into different granularity (number) bins.
- **Bins Injection**: Bins Injection utilizes the bin candidates' information to guide the refinement of the latent features of the depth token maps.

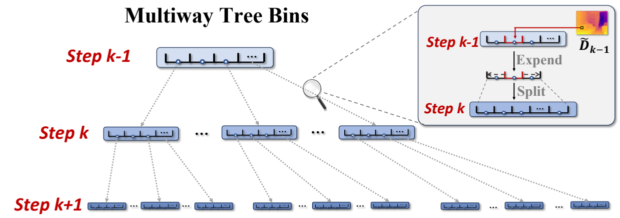

Figure 5. A schematic diagram of the multiway tree bins strategy.

3.3. Resolution Autoregressive Objective

This section elaborates on our DAR Transformer with the patch-wise causal attention mask for the resolution autoregressive objective.

DAR Transformer. Our DAR Transformer follows the vanilla architecture in [55], composed of Multi-headed Self-Attention (MSA), Layer Normalization (LN), and Multi-headed Cross-Attention (MCA) layers with residual connections, aiming to predict the logits sequence of different scale. Unlike text-to-image models that use class labels as the condition, depth estimation is mainly based on input RGB image features. Thus, we take the image features X as the condition to control depth estimation. At each step k, we first upsample the token map r_{k-1} of the previous step to the next resolution as the input token map y_{in}^k. The DAR Transformer takes y_{in}^k as input query and sends it to the MSA and MCA layers to finally produce the logits y_{out}^k, where MSA takes the previous token maps $y_{in}^{1:k}$ to compute keys and values via the patch-wise causal attention mask, and MCA takes image features X for attention calculation. We formulate the process at time step k as:

$$\begin{aligned} y_k &= \mathrm{Upsampling}(r_{k-1}), \\ y_{hidden}^k &= \mathrm{LN}(\mathrm{MSA}(y_k W_Q; y_{1:k} W_K; y_{1:k} W_V; \mathtt{Mask})), \\ y_{out}^k &= \mathrm{LN}(\mathrm{MCA}(y_{hidden}^k; X)), \end{aligned} \tag{3}$$

where W_Q, W_K, and W_V are weight matrices for computing queries, keys, and values, and $\mathtt{Mask}$ denotes the patch-wise causal attention mask tailored for the next-resolution autoregressive paradigm (to be discussed below). The logits y_{out}^k are then used to generate a latent token map r_k of step k guided by the Bins Injection module (to be discussed below). During this process, DAR can integrate previous knowledge, global image features, and bin candidate infor-

Table 1. Model sizes and architecture configurations of DAR.

Model	Model Size	Layers	Hidden Size	Heads
DAR-Small	440M	5	1024	4
DAR-Base	1B	9	2048	8
DAR-Large	2B	11	3072	12

mation to accommodate the detailed feature requirements for generating a higher-resolution depth map.

Patch-wise Causal Attention Mask. Note that to achieve "next-resolution" prediction, we adopt a patch-wise causal attention mask that treats the entire token map as a merged patch-wise token, as shown in Fig. 4. This novel mask can ensure that each token of y_k can interact only with its prefix tokens belonging to $y_{<k}$ and other tokens within y_k.

3.4. Granularity Autoregressive Objective

To achieve the autoregressive objective on granularity, we introduce two core modules: Multiway Tree Bins (MTBin) and Bins Injection.

Multiway Tree Bins (MTBin) Strategy. This strategy aims to use the previous depth map's prediction to generate smaller bins for the next step's finer-grained depth predictions. Instead of using a fixed number of bins with equal-size intervals, MTBin recursively searches for more high-quality depth by progressively reducing the search range, as illustrated in Fig. 5. Assuming that at step $k-1$, the depth range is divided into N bins in the uniform space, as:

$$[\mathbf{b}_{k-1}^1, \mathbf{b}_{k-1}^2, \dots, \mathbf{b}_{k-1}^{N+1}], \tag{4}$$

where $\mathbf{b}_{k-1}^i$ represents the left boundary of the i-th bin in step $k-1$, except $\mathbf{b}_{k-1}^{N+1}$ represents the right boundary of the N-th bin. Suppose the predicted depth $\tilde{D}_{k-1}(\mathbf{x})$ for a pixel $\mathbf{x}$ lies in the t-th target bin $[\mathbf{b}_{k-1}^t, \mathbf{b}_{k-1}^{t+1}]$, that is:

$$\mathbf{b}_{k-1}^t \leq \tilde{D}_{k-1}(\mathbf{x}) \leq \mathbf{b}_{k-1}^{t+1}. \tag{5}$$

Then MTBin will recursively divide this bin into more fine-grained sub-bins, and update the depth range. However, the ground truth may fall outside of the target bin, due to depth prediction error. Thus, to maintain the model's error tolerance, the new depth range first will be expanded to adjacent bins $[\mathbf{b}_{k-1}^{t-1}, \mathbf{b}_{k-1}^{t+2}]$, which is then split into sub-bins in the uniform space. This process can be formulated as:

$$L = \frac{\mathbf{b}_{k-1}^{\min\{t+2,N+1\}} - \mathbf{b}_{k-1}^{\max\{t-1,1\}}}{N}, \tag{6}$$

$$\mathbf{b}_k^i = \mathbf{b}_{k-1}^{t-1} + (i-1)\cdot L,\ i = 1, 2, \dots, N, \tag{7}$$

where $\min\{t+2, N+1\}$ and $\max\{t-1, 1\}$ are for avoiding out-of-right/left boundaries. Since this splitting process looks like a multiway tree, we call this concept *Multiway Tree Bins*. Each pixel's decision process is unique, ranging from coarse to fine granularity progressively. These sub-bins will serve as new bin candidates to further guide the modeling of depth features via *Bins Injection* (discussed below) and perform the linear combination with the Softmax values of the token maps predicted by the model to obtain finer-grained depth maps. Specifically, we can take the bin centers as the depth candidates at step k, formulated as:

$$\mathbf{c}_k^i = \frac{\mathbf{b}_k^i + \mathbf{b}_k^{i+1}}{2}, i = 1, 2, \dots, N, \tag{8}$$

where $\mathbf{c_i}$ represents the i-th bin center (depth candidate). When we obtain the per-pixel Softmax values of r_k, i.e., the probabilistic distribution $\mathbf{p}_k$ associated with depth candidates, we compute the final depth via a linear combination:

$$\tilde{D}_k(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{c}_k^i \cdot \mathbf{p}_k^i(\mathbf{x}), \tag{9}$$

where $\tilde{D}_k(\mathbf{x})$ and $\mathbf{p}_k^i(\mathbf{x})$ represent the predicted depth and i-th depth candidate probability of pixel $\mathbf{x}$ at autoregressive step k, respectively. Notably, at first we initialize the full depth range $[d_{min}, d_{max}]$ into N bins in the uniform space:

$$\mathbf{b}_1^i = d_{min} + (i-1)\cdot L,\ i = 1, 2, \dots, N, \tag{10}$$

where $\mathbf{b}_1^i$ represents the left boundary of the i-th bin at step 1, N is 16 by default, $[d_{min}, d_{max}]$ are [0.1, 10] and [0.1, 80] for `NYU Depth V2` and `KITTI`, respectively, and L denotes the bin width that is equal to $\frac{d_{max}-d_{min}}{N}$.

Bins Injection. Bins Injection module aims to utilize the new valid depth range and bin candidates to guide the modeling of depth features. First, we project the depth candidates $\mathbf{c}^k$ into the feature space via 3×3 convolutional layers. Then, the obtained bin features $\mathbf{f}_{bin}^k$ as the context, are used to further guide the output features of the DAR Transformer via a ConvGRU [52] module. We formulate this process as:

$$\mathbf{f}_{bin}^k = \text{Conv}_{3\times3}(\mathbf{c}^k), \tag{11}$$

$$r_k = \text{ConvGRU}(\mathbf{y}_{out}^k; \mathbf{f}_{bin}^k). \tag{12}$$

3.5. Other Details

Image Encoder. To ensure a fair comparison with existing methods, we choose ViT as the Image Encoder (which is the same as Depth Anything). The conditioned image feature tokens are obtained by aggregating the feature maps from different layers of the Image Encoder, by bringing them all to 1/8 resolution of the input image, resulting in a token map of size $1536 \times H/8 \times W/8$, where H and W are the height and width of the input RGB image.

Loss Function. Since the ground-truth depth map has missing values, we cannot resize the ground-truth to different resolutions. Thus, we upsample all the predicted depth maps to the same size of the ground-truth, and calculate and leverage the scaled Scale-Invariant Loss:

$$\mathcal{L} = \sum_{k=1}^{K} \alpha \sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{x}\in\mathbf{T}} (\mathbf{g}_k(\mathbf{x}))^2 - \frac{\beta}{|\mathbf{T}|^2}\left(\sum_{\mathbf{x}\in\mathbf{T}} \mathbf{g}_k(\mathbf{x})\right)^2}, \tag{13}$$

Table 2. Results on the indoor `NYU Depth v2` dataset. † Data-driven methods which use a pretrained encoder on a large amount of data.

Method	Encoder	Model Size	Lower is better ↓				Higher is better ↑		
			Abs Rel	RMSE	Sq Rel	$\log_{10}$	δ_1	δ_2	δ_3
Eigen et al. [12]	ResNet-101	45M	0.158	0.641	-	-	0.769	0.950	0.988
DORN [14]	ResNet-101	45M	0.115	0.509	-	0.051	0.828	0.965	0.992
BTS [30]	DenseNet-161	20M	0.110	0.392	0.066	0.047	0.885	0.978	0.994
AdaBins [4]	E-B5+mini-ViT	40M	0.103	0.364	-	0.044	0.903	0.984	0.997
P3Depth [38]	ResNet-101	45M	0.104	0.356	-	0.043	0.904	0.988	0.998
DPT [44]	VIT-L	343M	0.110	0.357		0.045	0.904	0.988	0.998
LocalBins [5]	E-B5	30M	0.099	0.357	-	0.042	0.907	0.987	0.998
NeWCRFs [64]	Swin-Large	270M	0.095	0.334	0.045	0.041	0.922	0.992	0.998
BinsFormer [32]	Swin-Large	380M	0.094	0.330	-	0.040	0.925	0.991	0.997
PixelFormer [2]	Swin-Large	365M	0.090	0.322	-	0.039	0.929	0.991	0.998
VA-DepthNet [34]	Swin-Large	262M	0.086	0.304	0.043	0.039	0.929	0.991	0.998
IEBins [49]	Swin-Large	273M	0.087	0.314	0.040	0.038	0.936	0.992	0.998
NDDepth [48]	Swin-Large	393M	0.087	0.311	0.041	0.038	0.936	0.991	0.998
DCDepth[57]	Swin-Large	259M	0.085	0.304	0.039	0.037	0.940	0.992	0.998
WorDepth [65]	Swin-Large	350M	0.088	0.317	-	0.038	0.932	0.992	0.998
VPD [66]	ViT-L	600M	0.069	0.254	0.030	0.027	0.964	0.995	0.999
EcoDepth [39]	ViT-L	954M	0.059	0.218	0.013	0.026	0.978	0.997	0.999
ZoeDepth† [6]	ViT-L	343M	0.077	0.282	-	0.033	0.951	0.994	0.999
Depth Anything† [63]	ViT-L	343M	0.063	0.235	0.020	0.026	0.975	0.997	0.999
DAR-Small	ViT-L	440M	**0.059**	**0.217**	**0.013**	**0.026**	**0.979**	**0.997**	**0.999**
DAR-Base	ViT-L	1.0B	**0.058**	**0.214**	**0.013**	**0.026**	**0.980**	**0.997**	**0.999**
DAR-Large	ViT-L	2.0B	**0.056**	**0.205**	**0.011**	**0.024**	**0.982**	**0.998**	**1.000**

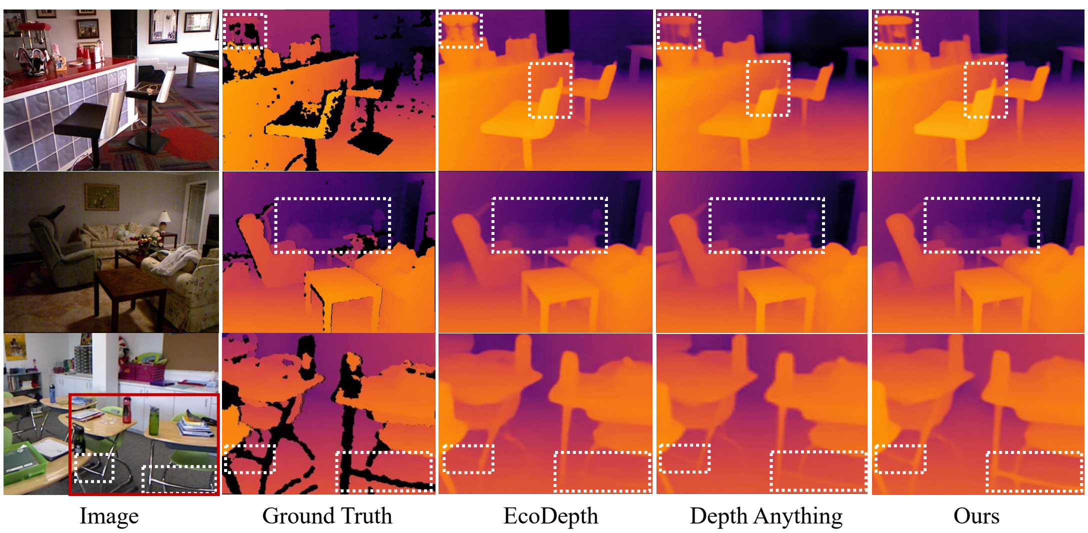

Figure 6. Qualitative results on the `NYU Depth V2` dataset. A brighter color denotes a closer distance.

where $\mathbf{g}_k(\mathbf{x}) = \log\tilde{\mathcal{D}}_k(\mathbf{x}) - \log\mathcal{D}_{gt}(\mathbf{x})$, K is the maximum number of steps and is set to 5, T represents the set of pixels having valid ground-truth values, and α and β are set to 10 and 0.85 based on [30].

4. Experiments

4.1. Datasets and Evaluation Metrics

Scaling Up Setting. Following [1, 17, 19, 24], we scale the model up with different sizes of the DAR Transformer. The configurations are shown in Table 1.

Implementation Details. Our model is implemented on the PyTorch platform. For optimization, the AdamW optimizer [26] is used with an initial learning rate of 3×10^{-5}. We first linearly increase the learning rate to 5×10^{-4}, and then linearly decrease it across the training iterations. The mini-batch size is 16. We train our model for 25 epochs for both the `KITTI` and `NYU Depth v2` datasets. In each step, the number of new bins is $N = 16$. For the DAR-Base model, each epoch takes ~30 minutes to train using 8

Table 3. Results on the outdoor KITTI dataset. † Data-driven methods which use a pretrained backbone on a large amount of data.

Method	Encoder	Model Size	Lower is better ↓				Higher is better ↑		
			Abs Rel	RMSE	Sq Rel	RMSE_{log}	δ_1	δ_2	δ_3
Eigen et al. [12]	ResNet-101	45M	0.203	6.307	1.517	0.282	0.702	0.898	0.967
DORN [14]	ResNet-101	45M	0.072	2.727	0.307	0.120	0.932	0.984	0.994
BTS [30]	DenseNet-161	20M	0.059	2.756	0.241	0.096	0.956	0.993	0.998
AdaBins [4]	E-B5+mini-ViT	40M	0.067	2.960	0.190	0.088	0.949	0.992	0.998
DPT [44]	VIT-L	343M	0.060	2.573	-	0.092	0.959	0.995	0.996
P3Depth [38]	ResNet-101	45M	0.071	2.842	0.270	0.103	0.953	0.993	0.998
NeWCRFs [64]	Swin-Large	270M	0.052	2.129	0.155	0.079	0.974	0.997	0.999
BinsFormer [32]	Swin-Large	380M	0.052	2.098	0.151	0.079	0.974	0.997	0.999
PixelFormer [2]	Swin-Large	365M	0.051	2.081	0.149	0.077	0.976	0.997	0.999
VA-DepthNet [34]	Swin-Large	262M	0.050	2.093	0.148	0.076	0.977	0.997	0.999
IEBins [49]	Swin-Large	273M	0.050	2.011	0.142	-	0.978	0.998	0.999
iDisc [40]	Swin-Large	263M	0.050	2.067	0.145	0.077	0.977	0.997	0.999
DCDepth [57]	Swin-Large	259M	0.051	2.044	0.145	0.076	0.977	0.997	0.999
WorDepth [65]	Swin-Large	350M	0.049	2.039	-	0.074	0.979	0.998	0.999
EcoDepth [39]	ViT-L	954M	0.048	2.039	0.139	0.074	0.979	0.998	1.000
ZoeDepth† [6]	ViT-L	343M	0.054	2.440	0.189	0.083	0.977	0.996	0.999
Depth Anything† [63]	ViT-L	343M	0.046	1.896	-	0.069	0.982	0.998	1.000
DAR-small	ViT-L	440M	**0.046**	**1.839**	**0.115**	**0.069**	**0.984**	**0.999**	**1.000**
DAR-Base	ViT-L	1.0B	**0.046**	**1.823**	**0.114**	**0.069**	**0.985**	**0.999**	**1.000**
DAR-Large	ViT-L	2.0B	**0.044**	**1.799**	**0.110**	**0.067**	**0.986**	**0.999**	**1.000**

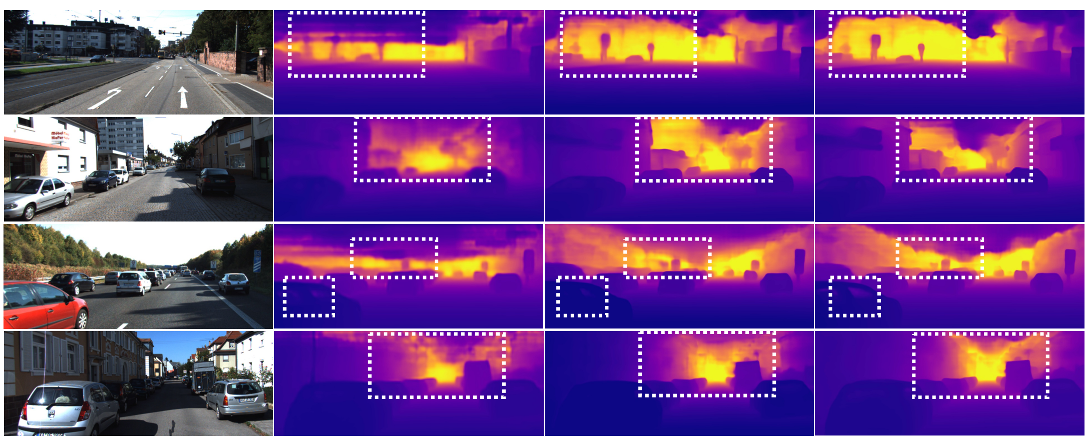

Figure 7. Qualitative results on the KITTI dataset. A darker color denotes a closer distance.

NVIDIA A100 GPUs.

Datasets. We conduct experiments on three benchmark datasets including NYU Depth V2 [50], KITTI [15], and SUN RGB-D [51]. **(a) NYU Depth V2** is a widely-used benchmark dataset that covers indoor scenes with depth values ranging from 0 to 10 meters. We follow the train-test split in [29], which uses 24,231 images for training and 654 images for testing. The ground truth depth maps were obtained using a structured light sensor with a resolution of 640×480. **(b) KITTI** is a widely-used outdoor benchmark dataset, containing images with depth values ranging from 0 to 80 meters. The official split provides 42,949 images for training, 1,000 for validation, and 500 for testing, with resolution 352×1216. **(c) SUN RGB-D**: We preprocess its images to 480 × 640 resolution for consistency. The depth values range from 0 to 10 meters. We use only the official test set (5050 images) for zero-shot evaluation.

4.2. Comparisons with Previous Methods

Quantitative Results on NYU Depth V2. We report the results in Table 2. As one can see, DAR with the same pretrained backbone and similar model size performs better than and above all the known methods. By scaling up, DAR-Large establishes a new SOTA performance in all the metrics, attaining 0.205 RMSE and 0.982 δ_1 on NYU Depth v2, largely outperforming existing methods.

Table 4. Zero-shot generalization to the SUN RGB-D dataset [51]. The models are trained on `NYU Depth V2` and tested on SUN RGB-D without fine-tuning.

Method	Lower is better ↓			Higher is better ↑		
	Abs Rel	RMSE	$\log_{10}$	δ_1	δ_2	δ_3
BinsFormer [32]	0.143	0.421	0.061	0.805	0.963	0.990
IEBins [49]	0.135	0.405	0.059	0.822	0.971	0.993
VPD [66]	0.121	0.355	0.045	0.861	0.980	0.995
Depth Anything [63]	0.119	0.346	0.043	0.864	0.981	0.995
DAR	**0.112**	**0.319**	**0.040**	**0.885**	**0.985**	**0.996**

Qualitative Results on `NYU Depth V2`. In Fig. 6, we present some qualitative comparison on the dataset. First, one can observe that our model performs better in depth estimation at the boundaries of the objects, making it more coherent and smooth (e.g., the back of the chair and the long-distant objects). This is helped by our autoregressive progressive paradigm, which maintains a coherent and smooth depth estimation when using previous predictions for the next-step, more refined prediction. Second, our DAR is much more accurate when estimating the depths of small and thin objects or long-distant visually relatively small objects, like the poles under the chair. These observations further demonstrate the superiority of our DAR.

Quantitative Results on `KITTI`. To further demonstrate the superiority of our proposed DAR in outdoor scenarios, we report results on the `KITTI` dataset in Table 3. Compared to the SOTA self-supervised model Depth Anything, our DAR-small with a similar model size achieves better performance, suggesting the superiority and potential of our proposed DAR. Compared to the SOTA supervised model Ecodepth, our DAR-small has smaller model sizes and performs better. Notably, the scaled-up version DAR-Large establishes a new SOTA performance, which outperforms the current SOTA (Depth Anything) in all the metrics (in particular, by 5% in RMSE).

Qualitative Results on `KITTI`. In Fig. 7, we present some qualitative comparison on the `KITTI` dataset. One can observe that DAR preserves fine-grained boundary details and generates more continuous depth values, further demonstrating the effectiveness of our new AR-based framework.

4.3. Zero-shot Generalization

Unlike the SOTA zero-shot transfer method (Depth Anything [63]), which requires pretraining on a large amount of data (61M) for effective performance, we show that our model generalizes well when trained only on a single `NYU Depth v2` dataset. Table 4 shows the quantitative results. As one can see, DAR achieves decent results on unseen datasets, substantiating the generalization ability of DAR.

4.4. Ablation Study

We conduct a comprehensive ablation study to examine the effectiveness of the two sub-AR objectives with their respective components and scale up. All the experiments presented in this section are conducted on `NYU-Depth-V2`. We choose the backbone of Depth Anything as the baseline model. The comparison results are reported in Table 5. Observe that each sub-AR objective and its components improve the baseline performance, supporting our hypothesis that the AR model is an effective monocular depth estimator. Furthermore, by scaling up the model size to 2.0B, DAR achieves the best performance in all the metrics, demonstrating the strong scalability of DAR. We also conduct ablation study on the bin number N, and show qualitative results of each step of DAR in the Supplemental Materials.

Table 5. Ablation study of DAR. “BI”: The Bins Injection module.

Method	Param.	Abs Rel ↓	RMSE ↓	d1 ↑
Baseline + Transformer	420M	0.063	0.229	0.976
Baseline + MTBins + BI	363M	0.061	0.220	0.978
Baseline + DAR	440M	0.059	0.217	0.979
Baseline + DAR + Scale Up	2.0B	0.056	0.205	0.982

4.5. Limitations

We need to acknowledge the following limitations. First, we apply a multi-step progressive paradigm to predict depths, which is smoother and more continuous, but consequently, it may blur the boundaries and reduce sharpness. Second, since we use the autoregressive Transformer, the model parameter number of DAR is relatively high, especially when scaling the model size up. But, we believe that further improvements in the complexity-accuracy trade-off of DAR can be achieved through large model distillation or lightweight AR foundation model design techniques.

5. Conclusions

In this paper, we presented a novel depth autoregressive (DAR) modeling approach to introduce scalability and generalization to an AR-based framework for monocular depth estimation (MDE). Our key idea is to transform the MDE task into two parallel autoregressive objectives, resolution and granularity, inspired by the insight of ordered properties of MDE in these two aspects. Further, we proposed the DAR Transformer and Multiway Tree Bins strategy to achieve these two autoregressive objectives, and connect them by the proposed Bin Injection module. We demonstrated the effectiveness of our approach on several benchmark datasets, and showed that it outperforms SOTA methods by a significant margin. Our proposed DAR also offers a promising way to integrate autoregressive depth estimation into existing autoregressive-based foundation models.

Acknowledgments

This research was partially supported by National Natural Science Foundation of China under grants No. 62176231, No. 12326612 and Zhejiang Key R&D Program of China under grant No. 2023C03053, No. 2024SSYS0026, and the Opening Foundation of the State Key Laboratory of Transvascular Implantation Devices, grant NO. SKLTID2024003.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 6

[2] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023. 1, 3, 6, 7

[3] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11746–11752. IEEE, 2021. 1, 3

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3, 6, 7

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. LocalBins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 6

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3, 6, 7

[7] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[8] Agneet Chatterjee, Tejas Gokhale, Chitta Baral, and Yezhou Yang. On the robustness of language guidance for low-level vision tasks: Findings from depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2794–2803, 2024. 3

[9] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019. 1

[10] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2Seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 2

[11] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 1, 3

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NIPS*, 27, 2014. 1, 3, 6, 7

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2, 3

[14] 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. 2, 3, 6, 7

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 7

[16] Frank E Harrell. Regression modeling strategies. *R Package Version*, pages 6–2, 2012. 3

[17] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 1, 6

[18] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. 1999. 3

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 6

[20] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75:151–172, 2007. 2

[21] 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. 1

[22] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023. 1, 3

[23] Weibin Jia, Wenjie Zhao, Zhihuan Song, and Zhengguo Li. Object servoing of differential-drive service robots using switched control. *Journal of Control and Decision*, 10(3): 314–325, 2023. 1

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 6

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 3

[26] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] 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 (3DV)*, pages 239–248. IEEE, 2016. 1, 3

[28] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 2, 3

[29] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019. 7

[30] 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. 6, 7

[31] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023. 1, 3

[32] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024. 3, 6, 7, 8

[33] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. SIFT Flow: Dense correspondence across different scenes. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 28–42. Springer, 2008. 2

[34] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. VA-DepthNet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023. 1, 3, 6, 7

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[36] Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D Hager, Austin Reiter, Russell H Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5):1438–1447, 2019. 1

[37] Seyda Oney, Nils Rodrigues, Michael Becher, Thomas Ertl, Guido Reina, Michael Sedlmair, and Daniel Weiskopf. Evaluation of gaze depth estimation from eye tracking in augmented reality. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020. 1

[38] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3Depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022. 1, 3, 6, 7

[39] Suraj Patni, Aradhye Agarwal, and Chetan Arora. ECoDepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28285–28295, 2024. 3, 6, 7

[40] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. iDisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487, 2023. 7

[41] Alec Radford. Improving language understanding by generative pre-training. 2018. 1

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 1, 2

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 3

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 6, 7

[45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[46] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Learning 3-D scene structure from a single still image. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[47] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 1, 3

[48] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. NDDepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023. 6

[49] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. IEBins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6, 7, 8

[50] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 7

[51] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark

suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 7, 8

[52] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[53] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 2, 3

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4

[56] Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, and Jian Wu. Ord2Seq: Regarding ordinal regression as label sequence prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5875, 2023. 3

[57] Kun Wang, Zhiqiang Yan, Junkai Fan, Wanlu Zhu, Xiang Li, Jun Li, and Jian Yang. DCDepth: Progressive monocular depth estimation in discrete Cosine domain. *arXiv preprint arXiv:2410.14980*, 2024. 6, 7

[58] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[59] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9697–9706, 2023. 2

[60] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜWA: Visual synthesis pretraining for neural visual world creation. In *European Conference on Computer Vision*, pages 720–736. Springer, 2022. 2

[61] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and Transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[62] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 1, 3

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 3, 6, 7, 8

[64] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. NeW CRFs: Neural window fully-connected CRFs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022. 1, 3, 6, 7

[65] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. WorDepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024. 3, 6, 7

[66] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 1, 3, 6, 8