REPDEPTH: REDESIGNING LIGHTWEIGHT MONOCULAR DEPTH ESTIMATION BEYOND NAIVE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular Depth Estimation (MDE) plays a vital role in enabling intelligent systems to perceive 3D structures from a single image. However, while high-performing frameworks rely on computationally heavy Vision Transformer (ViT) encoders, the performance of existing lightweight models still lag behind. To bridge this gap, we introduce a lightweight model named Representation-Preserving Depth (RePDepth), demonstrating a viable pathway to adapt the Large Foundation Model (LFM) architecture for lightweight architectures with minimal performance degradation. RePDepth is built upon the encoder-DPT decoder architecture, leveraging a ViT-CNN hybrid encoder named MobileViT to simultaneously achieve global receptive fields and high inference efficiency. Crucially, revealing that conventional Multi-scale Skip Connections (MSC) are sub-optimal in multiple lightweight settings, we adopt Deep-shifted Skip Connections (DSC) to route features with larger receptive fields. We further design a lightweight metric depth head, HieR module, based on a bin-centric approach with coarse-to-fine refinement. In-domain evaluations demonstrate that RePDepth outperforms existing lightweight MDE methods, achieving competitive accuracy with a marginal 3% performance drop on the KITTI dataset compared to LFMs, while significantly reduce the model size by around 200 times (with minimum of 1.85M parameters). Finally, extensive experiments provide deeper insights into the widely observed indoor-outdoor performance gap, validate our model's robust zero-shot metric depth estimation across diverse datasets, and highlight its edge deployment capabilities. *The code will be available later.*

1 INTRODUCTION

Monocular Depth Estimation (MDE) provides spatial understanding for intelligent systems using minimal sensors, benefiting applications in autonomous driving, VR/AR, robotics, etc. Recently, large foundation MDE models have significantly improved the performance of monocular depth estimation. However, these performance gains come at the cost of substantially increased model size and computational complexity, which hinders the practical deployment of MDE in real-world scenarios. In contrast, customized lightweight MDE models have weak encoders that suffer from limited feature representation capabilities, resulting in a substantial in-domain performance gap compared with Large Foundation Model (LFM).

In the domain of MDE, Eigen (Eigen et al., 2014) pioneered a deep learning-based MDE model based on Convolutional Neural Network (CNN). Then, a series of developments (Godard et al., 2017; Cao et al., 2017; Fu et al., 2018; Godard et al., 2019; Bhat et al., 2022; Li et al., 2024) have been published to further improve the overall performance of MDE. DPT (Ranftl et al., 2021) published an early work to utilize a Vision Transformer (ViT) (Dosovitskiy, 2020) as the encoder for feature extraction. To date, foundation MDE models (Bhat et al., 2023; Yang et al., 2024a;b; Lin et al., 2025) primarily rely on ViT-based encoders paired with DPT decoders to achieve state-of-the-art performance. However, this design not only increases parameters but also reduces inference speed. To enable practical deployment, related research explores lightweight monocular depth estimation that balances efficiency and accuracy. Early pioneering work, such as FastDepth (Wofk et al., 2019), significantly reduced model complexity by utilizing MobileNetV1 (Howard et al., 2017) combined

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with network pruning. Subsequent efforts have continued to explore efficient backbone alternatives, with models leveraging DDRNet-23-slim (Rudolph et al., 2022) or, more recently, coupling MobileNetV2 with a Vision Mamba (VMamba) decoder in LMDepth (Long & Zhou, 2025). Taken together, the encoder plays a critical role in determining the performance of monocular depth estimation (MDE) models (Bhat et al., 2023; Yang et al., 2024a;b). However, existing lightweight MDE methods commonly rely on manually designed compact encoders whose limited representation capacity constrains depth estimation accuracy. For instance, TST (Lee et al., 2023) attempted to integrate ViT modules into lightweight MDE, employing multi-scale skip connections with Multi-Head Attention to share global tokens with the decoder. This manually designed encoder prevents features from deeper encoder layers from benefiting from the global representations provided by the ViT blocks. Moreover, current research also predominantly employ Multi-scale Skip Connections (MSC), assuming that shallow encoder features provide essential geometric details. In lightweight encoders, however, shallow features are less robust than those learned by large-scale ViTs and may propagate local noise rather than meaningful geometric context. Finally, the metric depth prediction head also needs to be specifically designed to preserve accurate depth representations under limited parameter and computational constraints.

To address these issues, we introduce a lightweight Representation-Preserving Depth (RePDepth) model built on an encoder-DPT decoder architecture. The design of RePDepth is motivated by our observation that direct pruning of foundation models based on encoder-DPT decoder framework leads to significant performance degradation, proving naive compression to be inviable for lightweight MDE. We instead seek a backbone that is inherently lightweight while retaining global context modeling. Accordingly, we adopt MobileViT (Mehta & Rastegari, 2021), a compact CNN–ViT hybrid that achieves a more favorable accuracy–efficiency trade-off than pruned DINOv2 (Oquab et al., 2023) and MobileNet variants (Sandler et al., 2018; Howard et al., 2019). To fully exploit the global representations learned by MobileViT, we further introduce Deep-Shifted Skip Connections (DSC), which propagate features from deeper encoder stages to the decoder. We further find that DSC consistently outperforms conventional multi-scale skip connections across multiple encoders in multiple existing models. Moreover, we introduce Hierarchy Regression (HieR), a lightweight metric depth head that combines exponentially spaced bin anchors with adaptive adjustment and residual-based fine-grained refinement. Finally, RePDepth achieves a lightweight size of 1.85M and 5.86M, a strong in-domain accuracy, robust zero-shot generalization, and low-latency edge deployment, while providing key insights into how dataset camera view bias drives the indoor-outdoor performance gap. Our contributions can be summarized as follows:

- We demonstrate that adapting the Encoder-DPT decoder framework yields a viable path for lightweight MDE superior performance when paired with thoughtfully tailored compact components. Besides, naively compressing large vision encoders fails to deliver optimal performance in our case.
- We present the rationale behind our encoder selection and accordingly select MobileViT, a lightweight hybrid CNN–ViT backbone that achieves most of the parameter reduction while largely preserving depth estimation accuracy.
- We reveal that shallow features in multi-scale skip connections of lightweight encoders are not an optimal design. To solve this, we adopt deep-shifted skip connections to preserve global context.
- We introduce a lightweight metric depth representation head, Hierarchy Regression (HieR) module, enabling fine-grained depth decoding while maintaining minimal computational overhead.
- Extensive experiments confirm that RePDepth achieves SOTA in-domain accuracy yielding δ_1 of 94.2%/95.5% on KITTI with only 1.85M/5.86M, only $\sim 3\%$ performance loss compared with LFMs (Yang et al., 2024a;b), and real-time inference on edge devices, accompanied by in-depth analysis of the indoor-outdoor performance gap.

2 Related Works

FastDepth (Wofk et al., 2019) introduced an encoder-decoder model optimized for mobile deployment, combining depthwise separable convolutions with real-time inference capability. In this work,

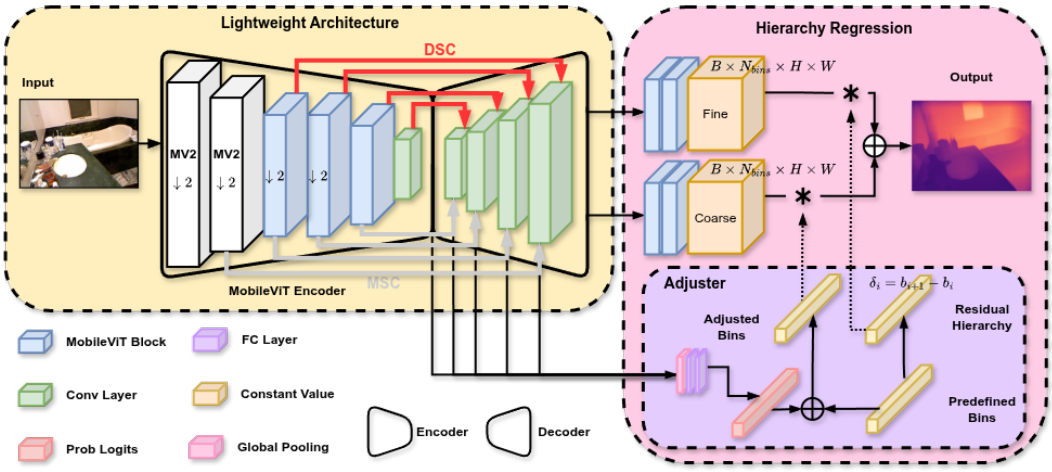

Figure 1: Overview of the proposed model. The framework comprises encoder-DPT decoder architecture (yellow shaded area)—pairing a MobileViT encoder with a pruned DPT (Ranftl et al., 2021) decoder—and a Hierarchy Regression (HieR) module (pink shaded area).

they also applied network pruning to remove redundant computational demand and improve inference speed. Tu et al. proposed TuMDE (Tu et al., 2020) using MobileNetv2 and a more simplified decoder than FastDepth. In their study, a reinforcement learning strategy was adopted to automatically prune redundant channels. GuideDepth (Long & Zhou, 2025) further improved the performance of lightweight MDE by leveraging guided feature fusion and distillation to preserve structural details under limited computation. LMDepth (Long & Zhou, 2025) leveraged MobileNetv2 as their encoder and established a vision-mamba-based decoder. More recently, LMDepth (Long & Zhou, 2025) achieved competitive results by combining MobileNetv2 with VMamba decoder, demonstrating the potential of advanced depth regressors. TST (Lee et al., 2023), adopted a Token-Sharing Transformer (TST). The TST module is a skip connection between encoder and decoder, incorporating a multi-head attention and a feed-forward network. However, the multi-head attention features produced in the TST module are not shared with other layers within the encoder. A closer look at the evolution of MDE reveals that the progress of lightweight approaches is closely tied to advancements in encoder design. Despite these advances, most lightweight models still face challenges in maintaining high accuracy across large-scale datasets. Our work follows this line of research by designing a lightweight and high-performing MDE framework that integrates efficient feature extraction, hierarchical regression, and loss-aware depth refinement.

3 Methodology

3.1 Light-weight Architecture

Figure 1 illustrates the overall architecture of RePDepth. Following the encoder-DPT decoder architecture adopted by recent foundation MDE models (Bhat et al., 2023; Yang et al., 2024a;b), RePDepth is designed from the perspective of representation preservation rather than straightforward model compression. Further, we employ a lightweight MobileViT (Mehta & Rastegari, 2021) encoder to maintain efficient global representation learning, a pruned DPT decoder for computationally efficient dense prediction, and Deep-shifted Skip Connections to enable deeper-layer features with stronger global context. Finally, a lightweight metric depth head, Hierarchy Regression (HieR) module, to produce accurate metric depth estimation with minimal computational overhead.

3.1.1 Lightweight Encoder

Monocular depth estimation critically requires a global receptive field to capture scene geometry. While traditional CNNs lack long-range dependencies due to local operations, pure ViT encoders incur heavy computational overhead and latency. A ViT-CNN hybrid encoder provides the ideal trade-off, achieving a lightweight footprint while preserving strong representation capacity.

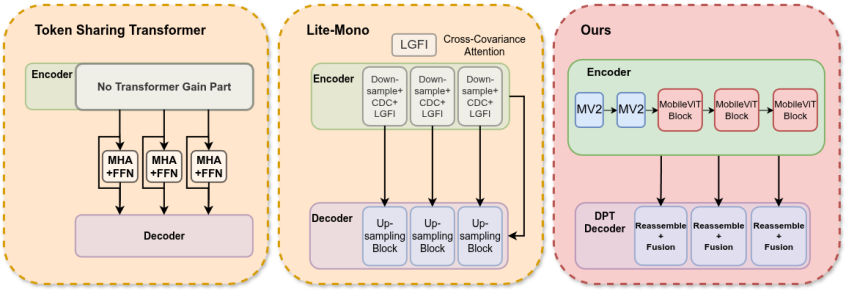

Figure 2: A representative architecture comparison of similar architectures.

As illustrated in the right panel of Figure 2, compared to existing approaches, MobileViT (Mehta & Rastegari, 2021) provides a global receptive field directly at the encoder level by integrating standard ViT modules, making it inherently better suited for foundation model paradigms. Transformer modules were frequently integrated into UNet skip connections to accelerate inference speed, with TST (Lee et al., 2023) serving as representative examples, as shown in the left panel of Figure 2. However, incorporating Transformer modules exclusively into skip connections prevents the deeper encoder layers from benefiting from global receptive field gains. Compared to task-specific custom architectures like Lite-mono (Zhang et al., 2023) (the middle panel of Figure 2) and ZipDepth (Tosi et al., 2026), MobileViT (Mehta & Rastegari, 2021) is inherently compatible with foundation model frameworks. Furthermore, MobileViT retains complete Vision-Transformer-based global reasoning rather than relying on attention approximations. Compared to hand-crafted lightweight backbones, MobileViT achieves superior performance across standard vision benchmarks, delivering more robust semantic representations. Besides, MobileViT is more compatible with transferring architectural knowledge from large-scale pretraining.

3.1.2 Pruned DPT Decoder

While our architecture preserves the original DPT decoder, it is pruned to achieve a lighter design (approximately 0.85M). In particular, the channel dimension is reduced to 64, resulting in a substantial reduction in model parameters and a notable improvement in computational speed. Furthermore, the channel dimensions of the multi-scale features from the skip connections are reduced to one-eighth, effectively compressing the feature representation and lowering the parameter count.

3.1.3 Deep-shifted Skip Connection

Conventional encoder-decoder architectures typically aggregate multi-scale encoder features through skip connections to preserve spatial details. However, features from different stages exhibit distinct contextual information, and directly aggregating them may introduce feature misalignment and contextual discrepancies (Huang et al., 2021b;a). This issue can be more pronounced in lightweight encoders, whose limited depth and capacity provide less opportunity to contextualize local representations. Thus, densely incorporating shallow features may not necessarily benefit lightweight MDE models.

To address this issue, we introduce *Deep-shifted Skip Connection*, which shifts skip connections toward deeper encoder stages. Instead of densely propagating features from early layers, we selectively connect the decoder to deeper, more contextually enriched representations. This design reduces reliance on less-contextualized shallow features while retaining effective hierarchical information for depth prediction. Since it only modifies the feature connection topology without introducing additional modules or parameters, it is particularly suitable for resource-constrained MDE.

3.2 Hierarchy Regression Module

We introduce a lightweight metric depth head, Hierarchy Regression (HieR) module, to achieve smaller size and better deploying performance. The HieR module is built upon three key components: Predefined Logarithmic Bin Anchors, a Simplified Adjuster, and a Coarse-to-Fine Depth Representation. Since all pixels share a common depth range, independently generating a complete set of depth bins for every pixel introduces unnecessary computational overhead and parameter in-

crease. We therefore construct a shared set of logarithmically spaced anchors, which efficiently captures the multiplicative variation of depth across the predefined range. To retain input adaptivity, the Simplified Adjuster applies bounded, data-dependent corrections to these anchors with minimal additional computation. Based on the resulting coarse prediction, the Coarse-to-Fine Depth Representation further estimates a fine-grained residual within the corresponding depth interval, improving depth precision through additive residual refinement.

3.2.1 Predefined Logarithmic Bin Anchors

We generate the bins based on a prior of an exponential distribution of depth, and subsequently refine them at the data level using an adjuster module. Concretely, coarse bins are constant values and generated according to the maximum and the minimum value of a dataset, following an exponential distribution. First, the min-max depth interval is converted into logarithmic space $d_{max}^{ln} = ln(d_{max}), d_{min}^{ln} = ln(d_{min})$. The interval $[d_{min}^{ln}, d_{max}^{ln}]$ is then split equally as follows:

$$b_i^{\text{ln}} = d_{min}^{ln} + i \cdot \frac{d_{max}^{ln} - d_{min}^{ln}}{n-1}, \quad i = 0, 1, \ldots, n-1 \tag{1}$$

Where n denotes the number of coarse bins. Eventually, the definition of logarithmic bin anchors B is:

$$B = [exp(b_0^{\text{log}}), exp(b_1^{\text{log}}), ..., exp(b_{n-1}^{\text{log}})] \tag{2}$$

3.2.2 Simplified Adjuster

A light-weight adjuster module is designed to enable learnable data-driven strategies. The adjuster slightly changes the predefined bin anchors according to the m input multi-scale feature iteratively. The adaptive average pooling flattens four multi-scale features into a 1D feature f_i, where $i \in \{1, ..., m\}$. Then, a two-layer MLP block with RELU activation is used to learn the bias. The output of the MLP block is normalized and then scaled with a factor λ. In practice, we use $\lambda = 0.05$ as (Li et al., 2024). The predefined bins are iteratively refined with adjuster:

$$\hat{B} = [\hat{b}_0, \hat{b}_1, ..., \hat{b}_{n-1}], \hat{b}_i = \frac{\sum_{i=1}^{m}(b_i + b_i \cdot \lambda \cdot tanh(f_i))}{m} \tag{3}$$

3.2.3 Coarse-to-Fine Depth Representation

The proposed Hierarchy Regression (HieR) incorporates an additional residual component, enabling fine-grained refinement of the final bin prediction. The overall definition of HieR is as follows:

$$d^{xy} = \sum_{i=1}^{m} p_i^{xy} \cdot B_i + \sum_{i=1}^{m} p^{xy\prime}{}_i \cdot R_i \tag{4}$$

Here, $\sum_{i=1}^{m} p^{xy\prime}{}_i \cdot R_i$ represents the bins of the residual hierarchy, serving to refine the prediction of the last bin width. p^{xy} denotes the bin probability of pixel at location (x, y). Specifically, the residual hierarchy decomposes the predefined bins into each defined by its local width δ_i:

$$\delta_i = b_{i+1} - b_i, i \in [0, 1, ..., n-1] \tag{5}$$

Since this process yields a residual hierarchy with one fewer dimension than the predefined bins, we reuse the width of the last bin to fill the missing dimension and preserve structural consistency. Therefore, the final Residual hierarchy R is defined as:

$$R = [\delta_0, \delta_1, ..., \delta_i, \delta_i], i \in [0, 1, ..., n-1] \tag{6}$$

4 Experiment

4.1 Implementation Details

We implement RePDepth in Pytorch and train on a single NVIDIA V100. During the training process, we adopt the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}. The model is trained for 20 epochs with a batch size of 32 on NYUD v2 dataset. On KITTI dataset, we train 15 epochs with a batch size of 16. Both datasets are Eigen split. We adopt the widely recognized Eigen split (Eigen et al., 2014) for data partitioning across both datasets.

Table 1: In-domain comparison with state-of-the-art lightweight MDE models on KITTI dataset.

Method	Resolution	Arch	Params	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL↓	Sq-rel↓	RMSE↓
Low Resolution									
FastDepth (Wofk et al., 2019)	192×640	CNN	3.9M	0.778	0.942	0.987	0.142	0.591	–
Lite-Mono-8M (Zhang et al., 2023)	192×640	ViT+CNN	8.7M	0.886	0.963	0.983	0.107	0.765	4.56
DaNet (Sheng et al., 2022)	192×620	ViT+CNN	8.2M	0.885	0.972	0.990	0.110	0.550	6.20
LMDepth (Long & Zhou, 2025)	192×620	Mamba+CNN	2.9M	0.908	0.983	0.996	0.089	0.441	3.63
RePDepth-XXS	256×576	ViT+CNN	**1.85M**	**0.933**	**0.988**	**0.997**	**0.076**	**0.33**	**3.00**
RePDepth-S	256×576	ViT+CNN	5.8M	**0.948**	**0.991**	**0.998**	**0.068**	**0.257**	**2.76**
High Resolution									
FastDepth (Tu et al., 2020)	384×1280	CNN	3.9M	0.825	0.961	0.991	0.136	0.714	4.33
Lite-Mono-8M (Zhang et al., 2023)	320×1024	ViT+CNN	8.7M	0.905	0.967	0.984	0.097	0.710	4.30
DaNet (Sheng et al., 2022)	384×1280	ViT+CNN	8.2M	0.920	0.981	0.995	0.086	0.411	3.52
BoRe-Depth (Liu et al., 2025)	-	ViT+CNN	8.7M	0.889	0.967	0.986	0.103	-	4.323
LMDepth (Long & Zhou, 2025)	384×1280	Mamba+CNN	2.9M	0.926	0.988	0.997	0.079	0.352	3.28
RePDepth-XXS	384×1152	ViT+CNN	**1.85M**	**0.942**	**0.991**	**0.998**	**0.070**	**0.280**	**2.76**
RePDepth-S	384×1152	ViT+CNN	5.86M	**0.955**	**0.992**	**0.998**	**0.065**	**0.234**	**2.53**

Table 2: In-domain comparison with state-of-the-art lightweight MDE models on NYUD v2 dataset.

Method	Resolution	Arch	Params	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL↓	log10↓	RMS↓
Low Resolution									
FastDepth (Wofk et al., 2019)	240 × 320	CNN	3.9M	0.771	0.939	0.976	0.135	-	0.580
DaNet (Sheng et al., 2022)	228 × 304	ViT+CNN	8.2M	0.829	0.963	0.991	0.140	0.059	0.487
LMDepth (Long & Zhou, 2025)	240 × 320	Mamba+CNN	2.9M	0.833	0.968	0.993	0.134	0.055	0.471
RePDepth-XXS	256 × 384	ViT+CNN	1.85M	0.812	0.953	0.987	0.148	0.061	0.490
RePDepth-S	256 × 384	ViT+CNN	5.86M	**0.846**	**0.970**	0.993	**0.128**	**0.054**	**0.439**
High Resolution									
FastDepth (Wofk et al., 2019)	480 × 640	CNN	3.9M	0.778	0.942	0.987	0.142	-	0.591
DaNet (Sheng et al., 2022)	480 × 640	ViT+CNN	8.2M	0.829	0.963	0.991	0.140	0.059	0.487
BoRe-Depth (Liu et al., 2025)	-	ViT+CNN	8.7M	0.843	0.954	0.983	0.124	-	0.675
LMDepth (Long & Zhou, 2025)	480 × 640	Mamba+CNN	2.9M	0.854	**0.977**	0.995	0.123	0.052	0.438
RePDepth-XXS	448 × 576	ViT+CNN	**1.85M**	0.827	0.963	0.991	0.137	0.058	0.468
RePDepth-S	448 × 576	ViT+CNN	5.8M	**0.858**	0.974	0.994	**0.121**	**0.051**	**0.420**

4.2 In-domain Comparisons

To ensure a standardized comparison with previous methods, we follow the evaluation protocol of LMDepth (Long & Zhou, 2025) to perform in-domain comparisons on KITTI (Geiger et al., 2013) and NYUD v2 (Nathan Silberman & Fergus, 2012), representing outdoor and indoor scenes, respectively. Table 1 presents a comparison between RePDepth and the existing methods on KITTI dataset with multiple metrics. We introduce two variants based on the backbone scale: RePDepth-S (MobileViT-S) and RePDepth-XXS (MobileViT-XXS). Our experiments are grouped into high- and low-resolution settings. The high-resolution setup closely mirrors the native dataset resolutions (640 $\times$ 480 for NYUD v2 and 384 $\times$ 1152 for KITTI), whereas the low-resolution variant operates at half scale. Notably, to accommodate the varying structural constraints of different network architectures, we select the resolution closest to, yet strictly not exceeding, the target input sizes to ensure a fair and competitive evaluation.

According to constraints on the input image size of MobileViT (Mehta & Rastegari, 2021), we used 256×576 for small input size, and 384×1152 for large input size. RePDepth-S outperforms the best existing method, LMDepth (Long & Zhou, 2025), by 2.9%, and only 3% lower than existing LFMs (Bhat et al., 2023; Yang et al., 2024a;b). Despite having only 1.85M parameters, our smallest model (RePDepth-XXS) achieves a δ_1 accuracy of 94.2%, outperforming the previous best result around 1.6%. Table 2 shows a comparison of our method with publicly available approaches on NYUD v2 (Nathan Silberman & Fergus, 2012). RePDepth-S outperforms the existing methods, even when the input image size is slightly smaller.

4.3 Ablation Studies

In our ablation studies, we aim to answer the following questions to validate our design choices: 1. Is our method compared well with naively compressing the foundation model. 2. How does the MobileViT backbone perform compared to other mainstream lightweight architectures? 3. Can deeper connections achieve better feature representations than conventional multi-scale fusions? 4. What is the contribution of the proposed HieR module to the overall performance?

Table 3: The ablation study on the impact of different decoders on NYUD v2 dataset with a fixed input resolution of 448×576. Chn Number denotes the channel number of decoder.

Encoder	Chn Number	Total Size	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL ↓	log10 ↓	RMS ↓
MobileVit XXS (Mehta & Rastegari, 2021)	64	1.85M	0.827	0.962	0.989	0.138	0.0587	0.469
Mobilenet V2 (Sandler et al., 2018)	48	2.76M	0.812	0.968	0.993	0.136	0.0594	0.470
Mobilenet V2	64	3.49M	0.840	0.970	0.994	0.127	0.0551	0.445
Mobilenet V3 (Howard et al., 2019)	48	3.24M	0.825	0.966	0.992	0.136	0.0579	0.458
Mobilenet V3	64	4.00M	0.831	0.969	0.994	0.135	0.0568	0.454
MobileViT S	48	5.61M	0.842	0.969	0.992	0.129	0.0548	0.443
MobileViT S	64	5.86M	**0.858**	**0.972**	**0.993**	**0.121**	**0.0515**	**0.421**

4.3.1 Comparison Among Encoders and Different Decoder Settings.

Table 3 compares lightweight encoders (MobileNets (Sandler et al., 2018; Howard et al., 2019) and MobileViT (Mehta & Rastegari, 2021)) within our encoder-DPT decoder framework. For a fair evaluation, we fix the DPT decoder at 64 channels with a 1/8 skip-connection compression ratio across all models, yielding varying total sizes. Input resolutions are set to 448×576 for MobileViT and 480×640 for MobileNets. Results show that MobileViT-S achieves the best metrics, while the ultra-lightweight MobileViT-XXS outperforms several significantly larger counterparts. Although increasing decoder channels predictably improves accuracy at the expense of model size, subsequent parameter-matched experiments confirm that these performance gains stem from architectural superiority rather than mere capacity differences.

Table 4: Performance comparison of feature connection strategies across various lightweight encoders and existing methods in the literature. 'DSC' denotes Deep-shifted Skip Connection, whereas 'MSC' represents conventional Multi-scale Skip Connection.

Encoder/Model	Con Type	Size	Layers	$\delta_1 \uparrow$	Abs Rel↓
Indoor(NYU)					
MobileNet v2	MSC	5.36M	[1 , 3 , 11, 18]	0.803	0.145
MobileNet v2	DSC	5.74M	[11, 15, 17, 18]	**0.812**	**0.141**
MobileNet v3	MSC	5.24M	[3 , 6 , 12, 16]	0.840	0.130
MobileNet v3	DSC	6.83M	[10, 12, 15, 16]	**0.844**	**0.128**
MobileViT S	MSC	5.73M	[4 , 6 , 8 , 11]	0.825	0.137
MobileViT S	DSC	5.86M	[8 , 9 , 10, 11]	**0.858 (3.3%↑)**	**0.121 (11.6%↓)**
GuideDepth (Rudolph et al., 2022)	MSC	5.8M	[0 , 0 , 0, 0]	0.840	0.128
GuideDepth	DSC	6.0M	[11, 18, 20, 23]	0.847	0.126
Outdoor(KITTI)					
MobileNet v2	MSC	5.36M	[1 , 3 , 11, 18]	0.948	0.067
MobileNet v2	DSC	5.74M	[11, 15, 17, 18]	**0.959**	**0.066**
MobileNet v3	MSC	5.24M	[3 , 6 , 12, 16]	0.948	0.068
MobileNet v3	DSC	6.83M	[10, 12, 15, 16]	0.947	**0.066**
MobileViT S	MSC	5.73M	[4 , 6 , 8 , 11]	0.944	0.071
MobileViT S	DSC	5.86M	[8 , 9 , 10, 11]	**0.955 (1.1%↑)**	**0.065 (8.4%↓)**
Lite-Mono (Zhang et al., 2023)	MSC	8.76M	[8, 13, 24]	0.904	0.097
Lite-Mono	DSC	9.15M	[13, 23, 24]	**0.911 (0.6%↑)**	**0.093(4.1%↓)**

4.3.2 Deep Connections VS Multi-scale Feature

We investigate the impact of the skip connections (between encoder and decoder) on model performance and size within our model. Since skip connections at different locations vary in channel dimensions, they further introduce discrepancies in overall model size. To address this, we adjust the channel compression ratios (1/2, 1/4, 1/8) following the skip connections to keep the parameter scales closely aligned. As shown in Table 4, the deep skip connections generally outperform multi-scale skip connections across different encoders across different architectures (encoder-DPT decoders (Yang et al., 2024a;b), Lite-Mono(Zhang et al., 2023), and GuideDepth (Rudolph et al., 2022)). This discrepancy is more evident in MobileNetV2 for indoor scenes and in MobileViT across both indoor and outdoor scenes. Early MobileViT stages (1–2) consist of local CNN convolutions lacking global attention, meaning MSC inevitably injects high-frequency noise into the decoder. In contrast, deeper stages (3–4) incorporate ViT-based global context, allowing the removal of shallow connections to deliver cleaner, semantically purified guidance.

4.3.3 HieR Module

We further verified the effectiveness of the HieR module, shown in Table 5. All hyperparameters were unchanged. The model with an appropriately configured HieR module reaches a superior performance. δ_1 can be marginally improved without finer bins. However, this comes at the cost of worse performance across all other metrics. The results also show that the design with 64 coarse bins with 16 hierarchy residual achieved highest performance. Moreover, the HieR module contains fewer parameters than the corresponding module in LocalBins (Bhat et al., 2022) (approximately

Table 5: The results of the ablation study on each proposed components on NYUD v2. The input size was set to 448×576.

Method	Bin (coarse/fine)	**Params**	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	**REL** $\downarrow$	**log10** $\downarrow$	**RMS** $\downarrow$
HieR (- fine bins)	64/-	5.86M	**0.8601**	0.969	0.993	0.1227	0.0521	0.4238
HieR	64/16	5.86M	0.8578	**0.972**	**0.993**	**0.1214**	**0.0514**	**0.4208**
HieR	32/8	5.86M	0.8503	0.970	0.993	0.1251	0.0533	0.4330
HieR	128/32	5.88M	0.8524	0.973	0.993	0.1235	0.0530	0.4281
HieR	64/64	5.86M	0.8541	0.971	0.993	0.1258	0.0533	0.4363
ZoeDepth head(Bhat et al., 2023; Yang et al., 2024a)	64/-	5.98M	0.8462	0.971	0.993	0.1275	0.0542	0.4348

Table 6: Zero-shot metric depth comparison on multiple datasets.

Indoor Datasets							
Method	Size	SUN RGB-D (Song et al., 2015)		iBims-1 (Koch et al., 2019)		DIODE Indoor (Vasiljevic et al., 2019)	
		AbsRel ($\downarrow$)	δ_1 ($\uparrow$)	AbsRel	δ_1	AbsRel	δ_1
RePDepth-XXS	1.85M	0.793	0.218	0.269	0.583	**0.382**	**0.400**
RePDepth-S	5.86M	0.815	0.18	**0.242**	**0.622**	0.397	0.388
GuideDepth (Rudolph et al., 2022)	5.8M	**0.391**	**0.350**	0.766	0.274	0.580	0.294
DANet (Sheng et al., 2022)	8.2M	0.651	0.133	0.457	0.474	0.390	0.291
DAv1 (Yang et al., 2024a)	355M	0.500	0.660	0.150	0.714	-	-
ZoeDepth (Bhat et al., 2023)	345M	0.520	0.545	0.169	0.656	-	-
Outdoor Datasets							
Method	Size	VKITTI v2 (Cabon et al., 2020)		DIODE Outdoor (Vasiljevic et al., 2019)			
		AbsRel ($\downarrow$)	δ_1 ($\uparrow$)	AbsRel	δ_1		
RePDepth-XXS	1.85M	0.172	0.694	0.679	0.194		
RePDepth-S	5.86M	**0.160**	0.737	0.714	0.204		
GuideDepth (Rudolph et al., 2022)	5.8M	0.360	0.377	0.827	0.186		
Lite-mono-8m (Zhang et al., 2023)	8.7M	0.162	**0.761**	**0.577**	**0.671**		
DAv1 (Yang et al., 2024a)	355M	0.085	0.913	0.794	0.288		
ZoeDepth (Bhat et al., 2023)	345M	0.106	0.844	0.814	0.237		

0.04M vs. 0.16M). Our design shows superior results compared to the ZoeDepth head (also used in Depth Anything v1/v2) counterpart.

4.3.4 Beyond Naive Compression

To evaluate whether naive compression can substitute a dedicated lightweight design, we baseline against a pruned Depth Anything V1 (Yang et al., 2024a). By restricting both decoders to 64 channels, we align their parameter budgets (ours: 5.86M vs. pruned DAv1: 5.78M). **The results in Table 8 correspond to relative improvements of 27.3%/33.2% in δ_1 and relative reductions of 63.3%/51.6% in Abs Rel.** Comparing Table 8 with the last row of Table 5, our MobileViT-DPT framework achieves 84.6% δ_1 using the same DAv1 (ZoeDepth) head and decoder, significantly outperforming naively compressed DAv1. By eliminating head dependencies, using the same decoder, this result confirms the intrinsic superiority of our proposed architecture.

4.4 Indoor & Outdoor Performance Gap

We observe a constant performance gap between indoor and outdoor scenarios. As shown in Tables 1 and 2, existing methods (including ours) generally achieve superior quantitative performance in outdoor scenarios compared to indoor environments. We attribute this performance discrepancy primarily to the domain-inherent characteristics of the respective datasets. Accordingly, we focus our empirical validation of this dataset-driven phenomenon specifically within our proposed framework. To verify this hypothesis, we manipulated the Field of View (FoV) through cropping to observe our model's behavior under context-shifting crops, as shown in Figure 3. Furthermore, we applied aspect ratio stretching to the input images to investigate whether the model possesses robustness against geometric shape distortions.

In Table 7, the results show that our model experienced a significant performance loss when the upper part of the inputs is stretched to the original size. Swapping the upper and lower parts of the input also leads to a performance degradation. This is consistent with Dijk's (Dijk & Croon, 2019) conclusion that certain MDE models rely on the vertical image position of obstacles but not on their apparent size. Our experiments further examine whether this observation holds in indoor environments (NYUD v2), revealing a similar directional trend, even with less severe degradation. However, it is worth noting that the performance degradation on NYUD v2 is less pronounced than on KITTI. We hypothesize that this difference stems from the greater diversity in camera poses present in NYUD V2. Nevertheless, such pose variations are insufficient for the model to withstand more severe FOV truncation.

Figure 3: The data manipulation pipeline used for Indoor & Outdoor Performance Gap investigation.

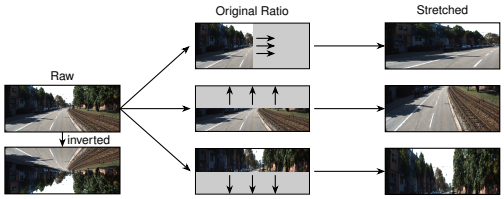

Table 7: Diagnostic analysis of spatial context **on indoor/outdoor dataset**.

KITTI				NYUD v2			
Region	Strategy	$\delta_1 \uparrow$	$AbsRel \downarrow$	Region	Strategy	$\delta_1 \uparrow$	$AbsRel \downarrow$
Left	Origin	0.919	0.083	Left	Origin	0.790	0.151
Left	Stretch	0.810	0.119	Left	Stretch	0.361	0.304
Upper	Origin	0.425	0.311	Upper	Origin	0.577	0.231
Lower	Origin	0.312	0.313	Lower	Origin	0.550	0.297
Upper	Stretch	0.020	0.691	Upper	Stretch	0.111	0.418
Lower	Stretch	0.683	0.159	Lower	Stretch	0.302	0.297
Swap	Up/Down	0.110	0.375	Swap	Up/Down	0.491	0.232

Table 8: Comparison between naive compression and the proposed lightweight design.

Architecture	**Depth Head**	$\delta_1 \uparrow$ **(KITTI/NYU)**	**Abs Rel** $\downarrow$ **(KITTI/NYU)**
Pruned DAv1	HieR	0.750 / 0.644	0.177 / 0.250
RePDepth	HieR	**0.955 / 0.858**	**0.065 / 0.121**

Table 9: Efficiency and computational complexity comparison on NVIDIA Jetson Orin Nano. All models are tested at an input resolution of 384×384 under 25w power mode.

Model	Framework	Accuracy	VRAM	FPS
Lite-Mono	TensorRT	FP16	-	58
RT-MonoDepth	TensorRT	FP16	-	63
ZipDepth	Pytorch	FP16	-	66
RePDepth-S	Pytorch	FP32	880MB	22
RePDepth-S	TensorRT	FP16	273MB	66
RePDepth-XXS	Pytorch	FP32	434MB	36
RePDepth-XXS	TensorRT	FP16	269MB	96
Module	**Framework**	**Size**	**Delay**	**Speed up**
ZoeDepth head	Pytorch	0.16M	12.77ms	1×
HieR module	Pytorch	0.04M	3.72ms	3.43×

4.5 Zero-Shot Experiments

We investigate the zero-shot metric depth generalization of the proposed model on unseen indoor/outdoor datasets, as shown in Table 6. iBims-1 (Koch et al., 2019), SUN RGB-D (Song et al., 2015), and DIODE indoor (Vasiljevic et al., 2019) datasets were used for zero-shot testing of indoor scenes. Similarly, the outdoor dataset is trained and fine-tuned on the KITTI dataset. (Geiger et al., 2013). VKITTI2 (Cabon et al., 2020), and DIODE outdoor (Vasiljevic et al., 2019) are selected as the zero-shot test protocol. All models are trained on NYUD v2 and KITTI, and directly tested on unseen datasets without any fine-tuning. Generally, RePDepth performs slightly lower than larger models. However, RePDepth consistently retains a considerable portion of the generalizability observed in large models across both indoor and outdoor settings. The results indicate that carefully designed lightweight networks have stronger generalizability, instead of larger model capacity.

4.6 Edge Deployment

We investigated the deployment performance of our model and other counterparts, shown in Table 9. The inference speed of RePDepth-XXS achieves 96FPS under TensorRT and FP16 settings. The lightest VRAM consumption of our models remains as little as 269 MB. Furthermore, the lower part of Table 9 shows proposed lightweight HieR module is 3.43 times faster than ZoeDepth head (Bhat et al., 2023; Yang et al., 2024a;b). These results demonstrate that our model achieves competitive real-time inference speeds and high memory efficiency compared to CNN-based approaches, which naturally hold an edge in inference speed.

5 Conclusions

In this study, we propose a lightweight monocular depth estimation model, RePDepth, following an encoder-DPT decoder architecture which utilizes a mobile ViT encoder, deep-shifted skip connections (DSC) and a novel Hierarchy Regression (HieR) module to better preserve global representation in lightweight MDE. Extensive evaluations demonstrate that RePDepth outperforms existing lightweight MDE models across indoor and outdoor benchmarks, while enabling real-time edge deployment and robust zero-shot generalization. Beyond raw performance, our work highlights the necessity of encoder-level ViT paradigms, refined skip-connection topologies, and effective depth representations, while offering critical insights into how dataset camera view bias affects cross-scene robustness. Moving forward, future research will prioritize advancing the zero-shot generalization of metric depth across unconstrained, real-world scenes, while exploring next-generation representation-preserving paradigms to further close the performance gap between compact models and heavy foundation architectures.

LLM / AI USAGE DISCLOSURE

In this work, we used generative AI tools for refining text readability and recommending relative references. We have not used generative AI tools for tasks with required disclosure, such as synthetic data generation, theoretical modeling, or method implementation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ACKNOWLEDGMENTS

This section will be added in the camera-ready version.

REFERENCES

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pp. 480–496. Springer, 2022.

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.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

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

Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2183–2191, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

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*, pp. 2002–2011, 2018.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 270–279, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 844–853. IEEE, 2021a.

Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):550–557, 2021b.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. iBims-v1, 2019.

Dong-Jae Lee, Jae Young Lee, Hyounguk Shon, Eojindl Yi, Yeong-Hun Park, Sung-Sik Cho, and Junmo Kim. Lightweight monocular depth estimation via token-sharing transformer. *arXiv preprint arXiv:2306.05682*, 2023.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Chang Liu, Juan Li, Sheng Zhang, Jie Li, and Xu Zhang. Bore-depth: Self-supervised monocular depth estimation with boundary refinement for embedded systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2455–2462. IEEE, 2025.

Jiahuan Long and Xin Zhou. Lmdepth: Lightweight mamba-based monocular depth estimation for real-world deployment. *arXiv preprint arXiv:2505.00980*, 2025.

Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arxiv 2021. *arXiv preprint arXiv:2110.02178*, 2021.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Michael Rudolph, Youssef Dawoud, Ronja Güldenring, Lazaros Nalpantidis, and Vasileios Belagiannis. Lightweight monocular depth estimation through guided decoding. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2344–2350. IEEE, 2022.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Fei Sheng, Feng Xue, Yicong Chang, Wenteng Liang, and Anlong Ming. Monocular depth distribution alignment with low computation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6548–6555. IEEE, 2022.

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*, pp. 567–576, 2015.

Fabio Tosi, Luca Bartolomei, Matteo Poggi, and Stefano Mattoccia. Zipdepth: Bringing lightweight zero-shot monocular depth anywhere, on any device. *arXiv preprint arXiv:2607.08771*, 2026.

Xiaohan Tu, Cheng Xu, Siping Liu, Renfa Li, Guoqi Xie, Jing Huang, and Laurence Tianruo Yang. Efficient monocular depth estimation for edge devices in internet of things. *IEEE Transactions on Industrial Informatics*, 17(4):2821–2832, 2020.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. URL `http://arxiv.org/abs/1908.00463`.

Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.

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*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18537–18546, 2023.

A Appendix

A.1 Accuracy vs. Efficiency Trade-offs

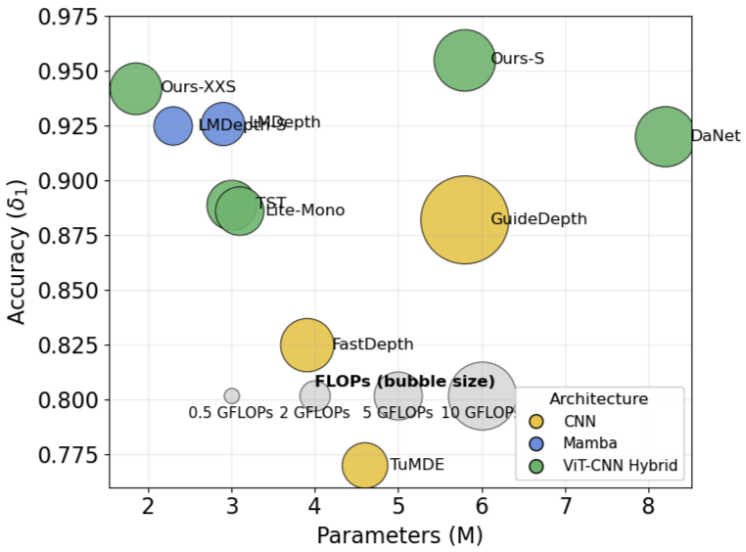

Figure 4: Comparison of performance, parameter size, and computational cost. The results are based on the outdoor dataset KITTI (Geiger et al., 2013) with input size of 384×1152.

Figure 4 shows the model size, computational, and performance comparisons of existing lightweight MDE models. RePDepth-S and RePDepth-XXS achieve the highest δ_1 with relatively small model size at a similar computational budget. Notably, with a mere 1.86M parameters, our model achieves 94.2% δ_1, surpassing the previous SOTA by over 2%. Remarkably, this performance trails large foundation models (e.g., Depth Anything V1 and V2 (Yang et al., 2024a;b)) by only $\sim 4\%$. RePDepth-S achieves a δ_1 accuracy of 95.5%, trailing large foundation models (e.g., Depth Anything V1 and V2) by only $\sim 4\%$.

A.2 Qualitative Comparison

The qualitative results are presented in Figure. 5. As observed, our model achieves superior detail reconstruction across both indoor and outdoor scenes.

Concretely, in indoor environments, fine structures are restored with significantly higher fidelity to the ground truth. As shown in the zoom in view of NYUD V2 results, our model demonstrates exceptional performance in capturing fine-grained structural details, faithfully preserving complex

Table 10: Diagnostic analysis of spatial context **on the KITTI dataset**. The deeper shades of gray represent the magnitude of the negative performance impact.

RePDepth-S						Depth Anything V1					
Region	Strategy	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$AbsRel \downarrow$	Region	Strategy	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$AbsRel \downarrow$
Left	Origin	0.919	0.987	0.997	0.083	Left	Origin	0.978	0.997	1.0	0.053
Left	Stretch	0.810	0.943	0.983	0.119	Left	Stretch	0.969	0.997	1.0	0.066
Upper	Origin	0.425	0.716	0.884	0.311	Upper	Origin	0.507	0.854	0.979	0.221
Lower	Origin	0.312	0.955	0.995	0.313	Lower	Origin	0.356	0.995	1.000	0.279
Upper	**Stretch**	**0.020**	**0.059**	**0.124**	**0.691**	**Upper**	**Stretch**	**0.040**	**0.146**	**0.397**	**0.496**
Lower	Stretch	0.683	0.960	0.996	0.159	Lower	Stretch	0.926	0.999	1.000	0.118

Table 11: Diagnostic analysis of spatial context **on the NYUD v2dataset**. The deeper shades of gray represent the magnitude of the negative performance impact. The data shows the stretched inputs exhibit a worse performance.

RePDepth-S						Depth Anything V1					
Region	Strategy	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$AbsRel \downarrow$	Region	Strategy	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$AbsRel \downarrow$
Left	Origin	0.790	0.969	0.988	0.151	Left	Origin	0.965	0.997	0.999	0.070
Left	Stretch	0.361	0.676	0.872	0.304	Left	Stretch	0.885	0.990	0.998	0.100
Upper	Origin	0.577	0.876	0.970	0.231	Upper	Origin	0.572	0.929	0.988	0.191
Lower	Origin	0.550	0.848	0.957	0.297	Lower	Origin	0.870	0.992	0.999	0.115
Upper	**Stretch**	**0.111**	**0.329**	**0.642**	**0.418**	**Upper**	**Stretch**	**0.080**	**0.490**	**0.943**	**0.353**
Lower	**Stretch**	**0.302**	**0.665**	**0.903**	**0.297**	**Lower**	**Stretch**	**0.268**	**0.893**	**0.997**	**0.251**

geometries such as door frames, bedside lamps, bookcase frameworks, and individual books on shelves. Notably, our model yields superior reconstruction quality on thinner structures, ranging from bedside lamps to bookcase frames and shelved books.

In outdoor settings, our method demonstrates better boundary recovery for distant objects as well as challenging small structures. Furthermore, our model exhibits superior detail preservation across diverse scenarios, yielding sharp depth predictions for near-range railings, distant traffic cones, trees, streetlights poles, and traffic signal poles. Crucially, our model successfully captures the subtle depth structure of the traffic cones (including the trees in the background), whereas all competing approaches fail to beware the existence of this object.

A.3 Extension on Indoor-Outdoor Performance Gap Experiments

To verify whether the spatial cropping sensitivity observed in Section 4.4 is specific to our architecture, we extend the partial cropping evaluation (e.g., retaining only the left, top, or bottom half of the input image) to a representative foundation model, Depth Anything V1 (Yang et al., 2024a). The results are shown in Table 11 and 10. Overall, the results reveal that both lightweight and large foundation models suffer performance degradation under half-cropping evaluations. However, the extent of this drop varies substantially across cropping configurations, exhibiting a remarkably consistent trend across model scales. Most notably, across both indoor and outdoor scenes, retaining the upper half of the input image and stretching it back to original resolution (*Upper, Stretch*) incurs the most severe performance collapse. While outdoor models are primarily governed by a global vertical gradient, indoor models exhibit a significantly higher sensitivity to the input aspect ratio. This demonstrates that the indoor-outdoor performance gap primarily stems from dataset-inherent biases, rather than the model architecture or its capacity.

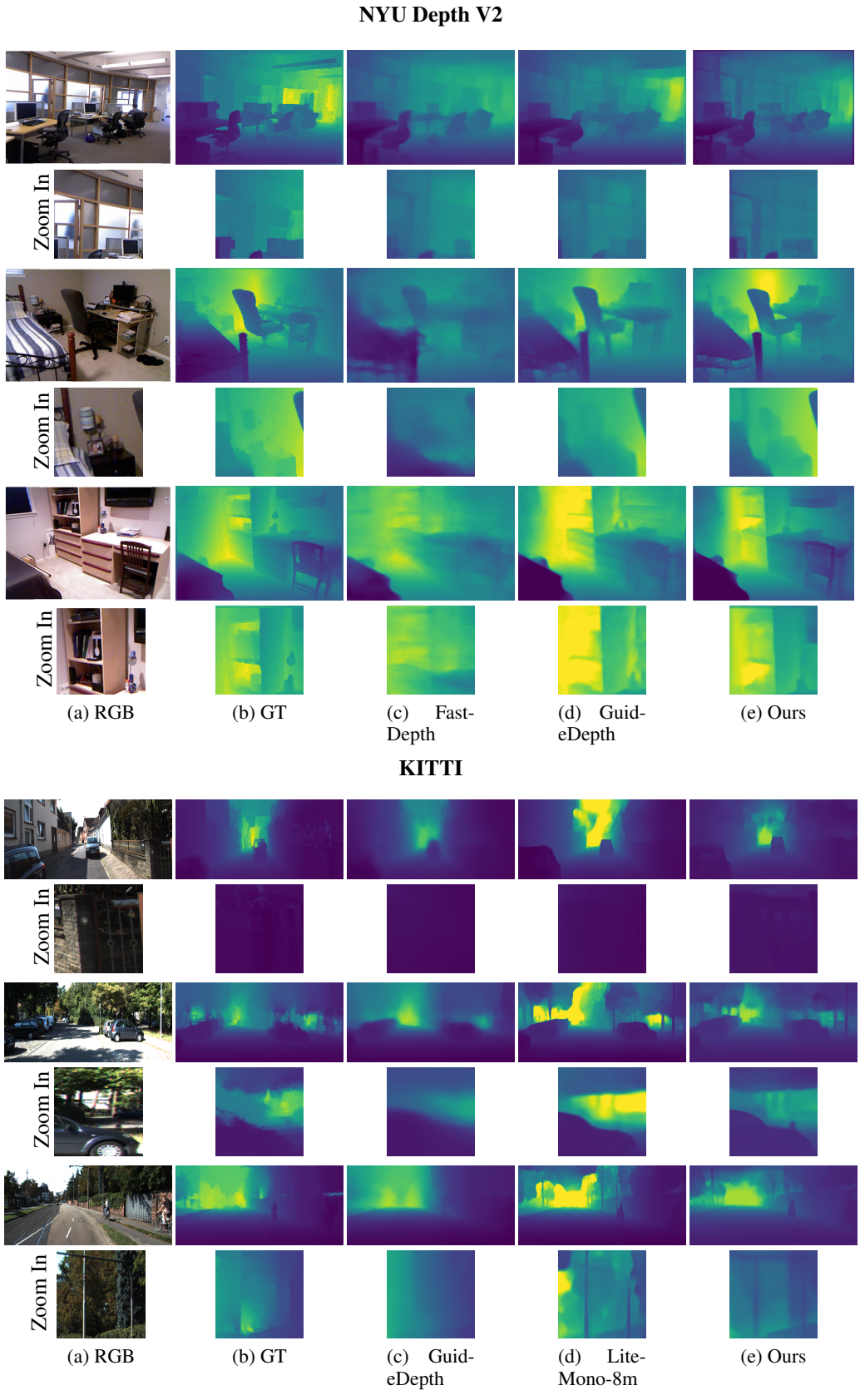

Figure 5: Qualitative comparison of depth estimation results on NYUD v2 and KITTI scenarios. Overall, Our method demonstrates superior detail preservation.