Last-Layer-Centric Feature Recombination: Unleashing 3D Geometric Knowledge in DINOv3 for Monocular Depth Estimation

Gongshu Wang, Zhirui Wang, and Kan Yang*

Aerospace Information Research Institute Chinese Academy of Sciences, Beijing, China
wanggs@aircas.ac.cn, yangkan@aircas.ac.cn

Abstract. Monocular depth estimation (MDE) is a fundamental yet inherently ill-posed task. Recent vision foundation models (VFMs), particularly DINO-based transformers, have significantly improved accuracy and generalization for dense prediction. Prior works generally follow a unified paradigm: sampling a fixed set of intermediate transformer layers at uniform intervals to build multi-scale features. This common practice implicitly assumes that geometric information is uniformly distributed across layers, which may underutilize the structural 3D cues encoded in VFMs. In this study, we present a systematic layer-wise analysis of DINOv3, revealing that 3D information is distributed non-uniformly: deeper layers exhibit stronger depth predictability and better capture inter-sample geometric variation. Motivated by this, we introduce a Last-Layer-Centric Feature Recombination (LFR) module to enhance geometric expressiveness. LFR treats the final layer as a geometric anchor and adaptively selects complementary intermediate layers according to a minimal-similarity criterion. Selected features are fused with the last-layer representation via compact linear adapters. Extensive experiments show that LFR module consistently improves MDE accuracy and achieves state-of-the-art performance. Our analysis sheds light on how geometric knowledge is organized within VFMs and offers an efficient strategy for unlocking their potential in dense 3D tasks. Code is released at https://github.com/book-book24/LFR.

Keywords: Monocular depth estimation · 3D understanding · Foundation model

1 Introduction

Monocular Depth Estimation (MDE) aims to predict pixel-wise scene depth from a single RGB image. It is a critical task in computer vision with diverse applications, including autonomous driving [41], virtual reality [12], embodied navigation [49] [13], and 3D reconstruction [42] [43]. However, MDE is inherently ill-posed, as a 2D image can correspond to infinitely many possible 3D scenes.

* Corresponding Author

Early approaches sought to mitigate this ambiguity by incorporating handcrafted geometric priors [10] [2] [47] [3] to guide model learning. Despite some success, these priors are often limited in their ability to generalize across diverse real-world scenarios [26].

The emergence of large-scale vision foundation models(VFMs) has created new opportunities for MDE. Pre-trained on massive datasets with proxy objectives, these models capture generic representations that can be effectively transferred to downstream tasks. Among them, DINO-style [6] [22] [35] vision transformers (ViTs) [8] have shown excellent performance for both semantic and dense prediction tasks, particularly those involving 3D understanding. Notably, recent state-of-the-art (SOTA) models such as Depth Anything [45] [46] and VGGT [38] build upon DINO-family, leveraging its representations through fine-tuning on large-scale 3D datasets. While these approaches focus on data driven and training objectives, their architectural design largely follows a common paradigm: features from a fixed set of intermediate ViT layers are fed into a DPT [30] decoder to produce multi-scale representations for final prediction.

Although this paradigm has proven effective, it may not fully exploit the rich geometric knowledge embedded in VFMs. Due to the weak inductive bias of ViTs [23], feature distributions vary significantly across layers and training objectives. For instance, attention maps in CLIP [28] differ markedly from those in DINO [22]. Thus, understanding the internal distribution of 3D knowledge within DINO is crucial for designing more effective transfer strategies.

In this work, we first perform a thorough layer-wise statistical analysis of DINOv3-L [35] features(Figure 1). We examine inter-sample representational distances, linear depth predictability, and representational similarity to ground-truth depth. Our analysis reveals that deeper layers encode richer and more explicit 3D information, and are better at capturing variations in 3D geometry across samples. In addition, prior works have shown that last-layer features already contain sufficient information to describe multi-scale representations [17], and that fusing last-layer with intermediate features benefits dense prediction [48].Together, these insights suggests that deeper features should play a more central role in MDE.

Motivated by these findings, we propose a simple yet effective last-layer-centric feature recombination (LFR) module (Figure 2). This plug-and-play module sits between the ViT backbone and the DPT decoder. It regards the last-layer features as the dominant geometric anchor and adaptively selects a small set of intermediate layers that are maximally complementary to this anchor. Concretely, we introduce a minimal-similarity-based selector that picks layers whose features have the lowest similarity to the last-layer features, maximizing knowledge complementarity. A lightweight adapter—composed of simple linear layers—then fuses each selected auxiliary feature with the last-layer features to produce recomposed representations. The final depth prediction is obtained as a weighted sum of per-level predictions. Our experiments on standard MDE benchmarks show that LFR consistently improves DINOv3-based MDE and attains

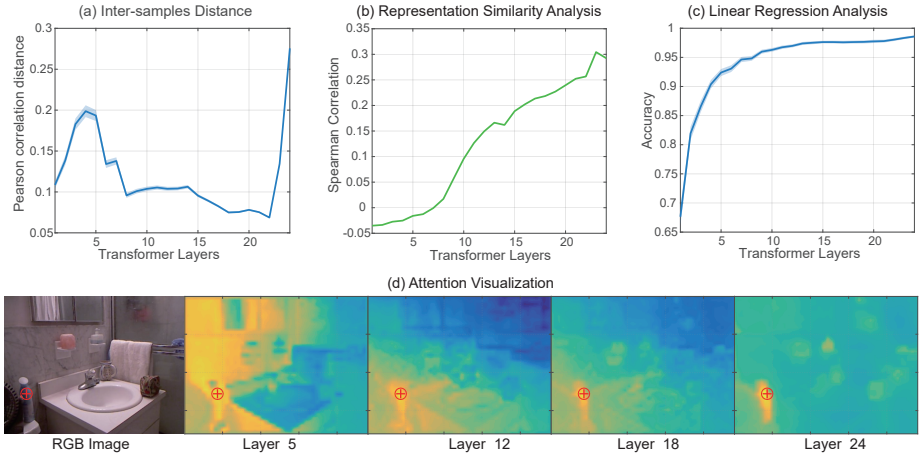

Fig. 1: Layer-wise feature analysis of DINOv3-L. For each transformer layer, we compute (a) inter-sample representational distance (mean pairwise Pearson correlation distance), (b) representational similarity to ground-truth depth (Spearman correlation between RDM and DDM), (c) depth predictability via linear regression. Shaded regions in (a) and (c) indicate 95% confidence intervals. (d) visualizes cosine similarity between the anchor token and all other tokens (brighter indicates higher similarity).

new SOTA accuracy with minimal overhead.We summarize our contributions as follows:

(1) We present the first systematic layer-wise analysis of DINOv3 for 3D geometry, demonstrating that deeper layers contain denser and more explicit geometric information that is highly predictive for depth.

(2) We propose LFR, a simple and effective last-layer-centric recombination module that adaptively selects and fuses complementary intermediate features to amplify the geometric signal for MDE.

(3) We validate LFR on multiple benchmarks, showing consistent gains over strong baselines and achieving SOTA results with efficient computation and strong zero-shot generalization.

2 Related Work

2.1 Monocular Depth Estimation

MDE aims to predict per-pixel depth from a single RGB image and has been advanced substantially by deep learning. The literature can be grouped into three main directions:

(1) Geometric Priors. Early methods sought to reduce scale ambiguity by incorporating explicit geometric constraints derived from human knowledge of the visual world. For instance, piecewise planarity priors decompose scenes into planar segments [50] [18], while normal consistency enforce alignment between

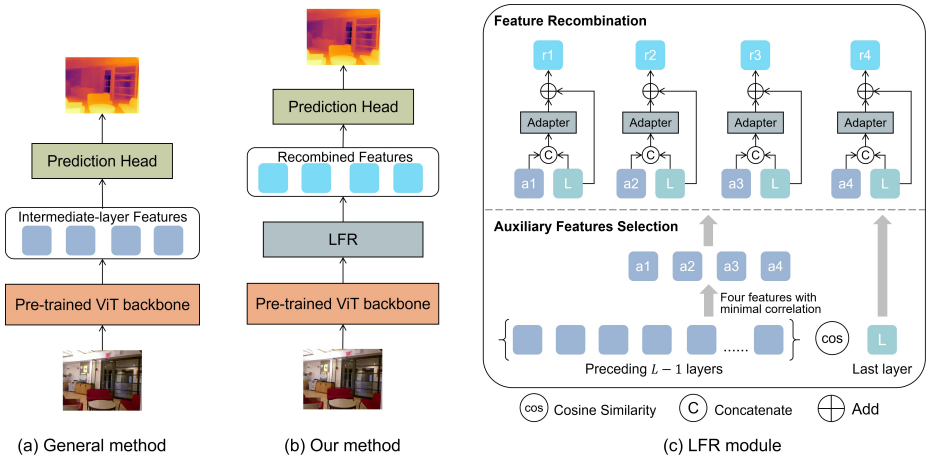

Fig. 2: Overview of the proposed method. (a) Generic dense prediction pipeline based on a ViT backbone. (b) Our improved architecture with the Last-layer-centric Feature Recombination (LFR) module inserted between the backbone and the prediction head. (c) Detailed illustration of the LFR module: the last-layer features serve as the dominant representation; four complementary intermediate layers are selected via a minimal-similarity criterion; lightweight linear adapters fuse each selected auxiliary feature with the last-layer features to produce recomposed features for the decoder.

normals derived from predicted and ground-truth depth [27] [20]. Although effective in constrained settings, these handcrafted priors often fail to generalize across diverse, unconstrained environments and may inadvertently restrict model expressivity.

(2) Generative Models. Motivated by the success of diffusion models in image synthesis, several approaches formulate MDE as a conditional generation task. DDP [14] adopts a noise-to-depth paradigm, conditioning the diffusion process on the input image. VPD [53] leverages denoising UNets as backbones to extract multi-scale features. ECoDepth [25] employs a conditional diffusion architecture with the semantic context. DAR [39] draws inspiration from VAR [37] and recasts MDE as a scale-wise autoregressive prediction task. These methods can achieve strong performance but typically require expensive iterative inference and large computational costs.

(3) Large-Scale Data-Driven Methods. The introduction of MiDaS [29] marked a paradigm shift by demonstrating that training on diverse, mixed datasets can yield strong zero-shot generalization. Building on this, ZoeDepth [4] and Depth Anything [45] [46] further scale up data—leveraging 62M unlabeled images via self-supervised learning—to achieve remarkable zero-shot capabilities. However, such approaches demand extensive data collection and training resources.

2.2 Vision Foundation Models

The emergence of VFMs pre-trained on massive datasets has reshaped the dense prediction tasks. Prominent examples include CLIP [28], the SAM family [15] [31] [5], and the DINO family [6] [22] [35]. Among these, DINO—a purely vision-based VFM built upon the ViT [8] architecture—has demonstrated exceptional performance on both semantic tasks and dense prediction tasks.

Notably, a growing work in 3D understanding leverages the dense features of DINO models. For instance, MoGe [40], a SOTA open-domain monocular geometry model; VGGT [38], which infers key 3D attributes from single or multiple views; and the Depth Anything series [45] [46], which achieves SOTA zero-shot MDE performance—all benefit substantially from DINO pre-trained weights. With the recent release of DINOv3 [35]—a 7-billion-parameter model trained on approximately 17 billion images—yield notably improved geometric encodings. Remarkably, simply transferring DINOv3 to relative depth estimation already surpasses the Depth Anything series. Anydepth pushes this further by attaching a lightweight prediction head to DINOv3, achieving superior results in relative depth estimation [32]. These findings underscore the rich 3D geometric knowledge encoded in DINOv3 and highlight its untapped potential.

Our work departs from prior practice by explicitly studying how 3D knowledge is arranged across DINOv3 layers and by proposing a principled, last-layer-centric recombination scheme that leverages this structure for MDE.

3 Method

3.1 Preliminaries: DINOv3 Architecture

DINOv3 adopts a standard ViT architecture. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, a patch embedding layer $P(\bullet)$ first partitions the image into non-overlapping patches and linearly projects each patch into a sequence of tokens: $T_0 = P(I) \in \mathbb{R}^{N\times C}$, where $N = \frac{H}{16} \times \frac{W}{16}$ denotes the number of tokens (i.e., the spatial resolution is reduced by a factor of 16) and C is the embedding dimension. This token sequence is then processed by a stack of L transformer encoder layers $\{B_l\}_{l=1}^{L}$. Each layer consists of multi-head self-attention and feed-forward networks, progressively integrating global contextual information and refining high-level semantic representations.

For dense prediction tasks based on ViT, a common practice is to equidistantly sample features from four intermediate layers (e.g., layers 3, 6, 9, and 12 in ViT-B) to construct multi-scale representations that mimic the hierarchical feature pyramids of convolutional neural networks [30]. Our analysis questions whether uniform sampling is optimal for transferring geometric knowledge.

3.2 Feature Analysis of DINOv3

To understand how 3D geometric knowledge is distributed across layers, we conduct a systematic statistical analysis of DINOv3-L ($L = 24,\ C = 1024$).

Table 1: Comparison on NYU Depth v2. † denotes generative models that require multiple forward passes. * denotes data-driven methods that require large-scale pre-training on relative depth. Bold indicates the best result; underline indicates the second best.

Method	Size	AbsRel↓	RMSE↓	log10↓	SqRel↓	δ_1↑	δ_2↑	δ_3↑
Eigenetal. [9]	45M	0.158	0.641	-	-	0.769	0.95	0.988
DORN [10]	45M	0.115	0.509	0.051	-	0.828	0.965	0.992
BTS [16]	20M	0.11	0.392	0.047	0.066	0.885	0.978	0.994
AdaBins [2]	40M	0.103	0.364	0.044	-	0.903	0.984	0.997
DPT [30]	343M	0.11	0.357	0.045	-	0.904	0.988	0.998
P3Depth [24]	43M	0.104	0.356	0.043	-	0.898	0.981	0.996
NeWCRFs [51]	270M	0.095	0.334	0.041	0.045	0.922	0.992	0.998
SwinV2-L [19]	197M	0.112	0.381	0.051	-	0.886	0.984	0.997
Localbins [3]	74M	0.099	0.357	0.042	-	0.907	0.987	0.998
PixelFormer [1]	365M	0.09	0.322	0.039	0.043	0.929	0.991	0.998
MIM [44]	197M	0.083	0.287	0.035	-	0.949	0.994	0.999
VPD [53] †	600M	0.069	0.254	0.030	0.027	0.964	0.995	0.999
EcoDepth [25] †	954M	0.059	0.218	0.026	0.013	0.978	0.997	0.999
DAR-B [39] †	1.0B	0.058	0.214	0.026	<u>0.013</u>	0.980	0.997	0.999
DAR-L [39] †	2.0B	**0.056**	**0.205**	**0.024**	**0.011**	0.982	0.998	**1.000**
ZoeDepth [4] *	343M	0.075	0.270	0.032	0.030	0.955	0.995	0.999
DAv1 [45] *	343M	**0.056**	<u>0.206</u>	**0.024**	-	**0.984**	**0.998**	**1.000**
DAv2 [46] *	343M	**0.056**	<u>0.206</u>	**0.024**	-	**0.984**	**0.998**	**1.000**
DINOv3-B [35]	120M	0.072	0.252	0.031	0.026	0.964	0.997	0.999
DINOv3-B+LFR	127M	0.070	0.251	0.030	0.025	0.965	0.997	0.999
DINOv3-L [35]	342M	0.060	0.212	0.026	0.018	0.981	0.998	**1.000**
DINOv3-L+LFR	350M	<u>0.057</u>	<u>0.206</u>	<u>0.025</u>	0.017	**0.984**	**0.998**	**1.000**

Specifically, we randomly sample 100 RGB images from the NYU Depth V2 [34], feed them through DINOv3-L, and extract image tokens from all transformer layers, denoted as$\left\{F_l \in \mathbb{R}^{N\times C}\right\}_{l=1}^{L}$. We then perform the following analyses:

Representational Dissimilarity Across Samples. For each layer, we compute the mean token representation per sample and calculate the pairwise Pearson correlation distance between samples, yielding a representational dissimilarity matrix $\left\{RDM_l \in \mathbb{R}^{S\times S}\right\}_{l=1}^{L}$, where S is the number of samples. The average value of RDM_l quantifies the inter-sample representational distance at layer l.

Representational Similarity to Depth Geometry. We first construct a depth distance matrix DDM $\in \mathbb{R}^{S\times S}$ by flattening each ground-truth depth map into a 1D vector and computing pairwise Pearson correlation distances. We then compute the Spearman correlation between RDM_l and DDM for each layer , measuring how well the representations capture variations in 3D geometry across samples.

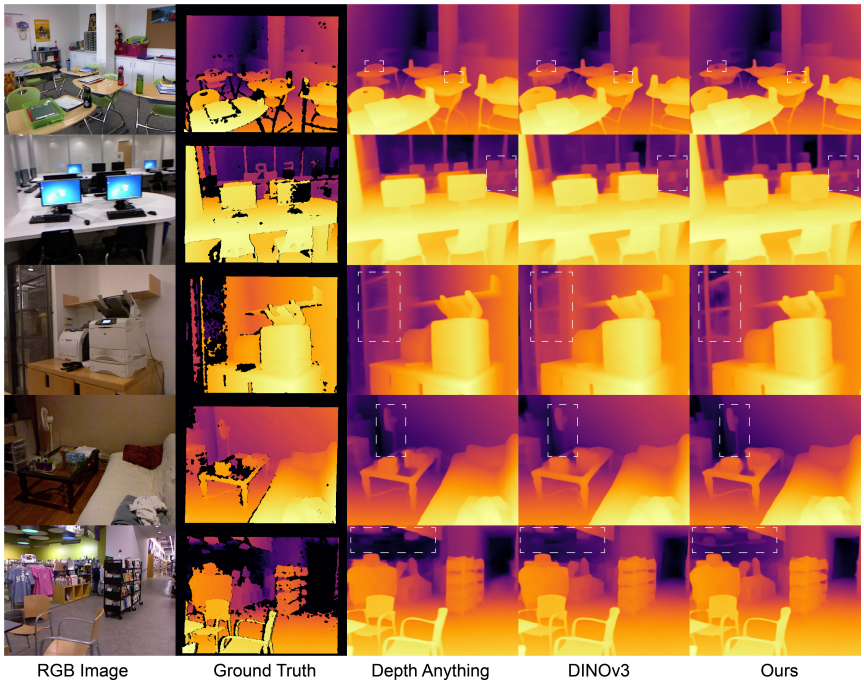

Fig. 3: Qualitative depth predictions on NYU Depth v2. Brighter pixels indicate closer distances.

Depth Predictability via Linear Regression. Each depth map is bilinearly downsampled to match the spatial resolution of the token grid and flattened into a depth vector. For each layer in each sample, we independently train a simple linear regression model that uses all channels of the image tokens as predictors and the corresponding depth vector as targets. The coefficient of determination (R^2) between predicted and ground-truth depth serves as a layer-wise depth-predictability metric.

As illustrated in Figure 1, our analysis reveals two key trends. First, features in the middle-to-deep layers tend to converge toward a shared representational space across samples, while the final layers shift toward more individualized representations. Second, depth predictability and the capacity to explain inter-sample variations in 3D geometry consistently increase with layer depth. Together, these trends indicate that deeper layers in DINOv3 encode richer and more explicit 3D information—suggesting they should be emphasized when transferring to MDE.

3.3 LFR module

Motivated by the above analysis, we propose a lightweight, plug-and-play feature recombination module inserted between the ViT backbone and the prediction

Table 2: Comparison on KITTI. Notation follows Table 1.

Method	Size	AbsRel↓	SqRel↓	RMSElog↓	RMSE↓	δ_1↑	δ_2↑	δ_3↑
Eigenetal [9]	45M	0.203	1.517	0.282	6.307	0.702	0.898	0.967
DORN [10]	45M	0.072	0.307	0.120	2.727	0.932	0.984	0.994
BTS [16]	20M	0.059	0.241	0.096	2.756	0.956	0.993	0.998
AdaBins [2]	40M	0.067	0.190	0.088	2.960	0.949	0.992	0.998
DPT [30]	343M	0.060	-	0.092	2.573	0.959	0.995	0.996
P3Depth [24]	45M	0.071	0.270	0.103	2.842	0.953	0.993	0.998
NeWCRFs [51]	270M	0.052	0.155	0.079	2.129	0.974	0.997	0.999
PixelFormer [1]	365M	0.051	0.149	0.077	2.081	0.976	0.997	0.999
IEBins [33]	273M	0.050	0.142	0.075	2.011	0.978	0.998	0.999
MIM [44]	197M	0.050	0.139	0.075	1.966	0.977	0.998	1.000
DDP [14] †	207M	0.050	0.148	0.076	2.072	0.975	0.997	0.999
EcoDepth [25] †	954M	0.048	0.139	0.074	2.039	0.979	0.998	0.999
DAR-B [39] †	1.0B	0.046	0.114	0.069	1.832	0.985	**0.999**	**1.000**
DAR-L [39] †	2.0B	<u>0.044</u>	0.110	<u>0.067</u>	1.799	0.986	**0.999**	**1.000**
ZoeDepth [4] *	343M	0.054	0.189	0.083	2.440	0.970	0.996	0.999
DA v1 [45] *	343M	0.046	-	0.069	1.896	0.982	0.998	**1.000**
DA v2 [46] *	343M	0.045	-	<u>0.067</u>	1.861	0.983	0.998	**1.000**
DINOv3-B [35]	120M	0.057	0.145	0.077	1.998	0.979	0.998	**1.000**
DINOv3-B+LFR	127M	0.049	0.128	0.072	1.931	0.981	0.998	**1.000**
DINOv3-L [35]	342M	0.046	<u>0.107</u>	<u>0.067</u>	<u>1.794</u>	<u>0.986</u>	**0.999**	**1.000**
DINOv3-L+LFR	350M	**0.043**	**0.099**	**0.063**	**1.711**	**0.988**	**0.999**	**1.000**

head. Without modifying the backbone, LFR centers on the final-layer features and adaptively selects a small number of complementary intermediate layers to fuse.

Minimal-Similarity-Based Feature Selection. Let F_L and Cls_L denote the image tokens and class token of the final layer, which serve as the dominant features. We first compute the average cosine similarity between Cls_L and the image tokens of each preceding layer. The K layers with the smallest similarity scores are selected as auxiliary layers ($K = 4$ in our study), and their image tokens and class tokens are denoted as $\{F_{ak},\ Cls_{ak}\}_{k=1}^{K}$. This minimal-similarity criterion ensures that the selected auxiliary features are semantically most complementary to the last-layer features, thereby maximizing information gain and minimizing redundancy. We also experiment with alternative selection strategies (e.g., maximal similarity, node degree, predicted-value-based selection), but find that minimal-similarity selection yields the best performance (Table 6).

Feature Recombination. For each selected auxiliary layer a_k, we recombine its features with the dominant last-layer features using a lightweight adapter. Specifically, the auxiliary and dominant features are concatenated along the channel dimension to preserve their respective semantic information. The concatenated features are then passed through a compact adapter A^k, which consists of a linear down-projection layer, a non-linear activation function, and a linear

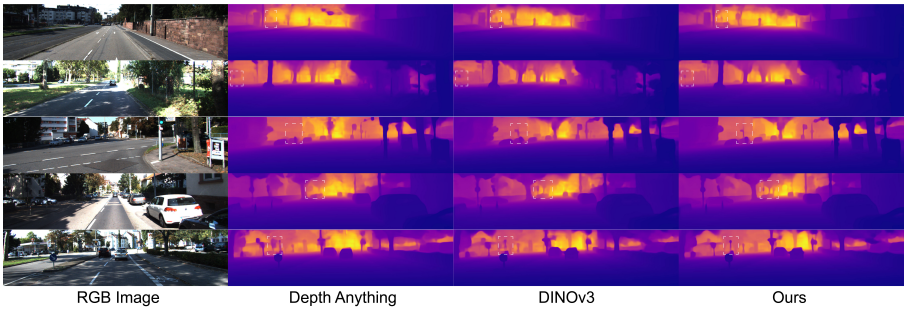

Fig. 4: Qualitative depth predictions on KITTI. Brighter pixels indicate farther distances.

up-projection layer, forming a bottleneck structure that facilitates feature fusion. A learnable scalar gate g^k, modulated by a tanh function to constrain its range to $[-1, 1]$, controls the contribution of the fused features. The final recomposed features are obtained via a residual connection:

$$F_{rk} = Tan\big(g^k\big) \bullet A^k\left([F_{ak}; F_L]\right) + F_L \tag{1}$$

$$Cls_{rk} = Tan\big(g^k\big) \bullet A^k\left([Cls_{ak}; Cls_L]\right) + Cls_L \tag{2}$$

The recomposed features $\{F_{rk}\}_{k=1}^{K}$ and $\{Cls_{rk}\}_{k=1}^{K}$ are subsequently used for depth regression.

Prediction Head. We adopt a DPT-style decoder [30] for depth regression but modify the aggregation strategy to exploit the recomposed features. In the original DPT, multi-scale features are assembled into a feature pyramid, and only the final layer's features are used for prediction. In contrast, we perform depth regression independently for each recomposed feature level. Specifically, after constructing feature pyramids from $\{F_{rk}\}_{k=1}^{K}$ and $\{Cls_{rk}\}_{k=1}^{K}$, we obtain per-level depth predictions. The final depth map is computed as a weighted sum of these per-level predictions, where the weights are derived from the recomposed class tokens:

$$w^k = Liner\left(Cls_{rk}\right), \; {w^k}_{k=1}^{K} = \mathrm{softmax}({w^k}_{k=1}^{K}) \tag{3}$$

and the final prediction is $\widetilde{D} = \sum_{k=1}^{K} w^k \cdot \widetilde{D}_k$, with $\widetilde{D}_k$ denoting the depth prediction from level k. This aggregation strategy allows each recomposed feature, which already encodes rich 3D knowledge centered on the last layer, to contribute according to its confidence, enhancing prediction robustness.

Loss Functions. We employ two complementary loss functions to supervise training. The first is the scale-invariant log loss [9], commonly used in MDE:

$$\mathcal{L}_{slog} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} g_n^2 - \frac{\lambda}{N^2}\left(\sum_{n=1}^{N} g_n\right)} \tag{4}$$

Table 3: Zero-shot generalization on SUN RGB-D. Models trained on NYU Depth v2 are directly evaluated on SUN RGB-D without fine-tuning.

Method	AbsRel↓	RMSE↓	log10↓	δ_1↑
BTS [16]	0.143	0.421	0.061	0.805
AdaBins [2]	0.159	0.476	-	0.768
NeWCRFs [51]	0.15	0.429	-	0.799
VPD [53]	0.121	0.355	0.045	0.861
EcoDepth [25]	0.112	0.319	-	0.885
DA v1 [45]	0.119	0.346	0.043	0.864
DAR-L [39]	0.112	0.319	0.040	0.885
DINOv3-L [35]	0.101	0.312	0.046	0.895
DINOv3-L+LFR	**0.099**	**0.300**	**0.044**	**0.907**

where $g_n = \log \widetilde{d}_n - \log d_n$, d_n and $\widetilde{d}_n$ are ground-truth and predicted depth values, N is the number of valid pixels, and λ is a balancing parameter (set to 0.85 following prior work). The second is a hierarchical normalization loss [52] that integrates multi-scale depth normalization to balance global coherence and local detail. The depth map is partitioned into patches at multiple scales $M \in 1, 4, 8$. For each patch u at scale M, both predicted and ground-truth depth are normalized within the patch. The per-pixel loss is then averaged over all patches containing that pixel:

$$\mathcal{L}_{HN}^{n} = \frac{1}{|U_n|} \sum_{u \in U_n} \left| \mathcal{N}_u \left(d_n \right) - \mathcal{N}_u \left(\widetilde{d}_n \right) \right| \tag{5}$$

where U_n denotes the set of patches containing pixel n, and $\mathcal{N}_u$ denotes normalization within patch u. The final hierarchical loss is averaged over all valid pixels:

$$\mathcal{L}_{HN} = \frac{1}{N} \sum_{n=1}^{N} L_{HN}^{n}. \tag{6}$$

This loss balances global coherence and local detail, and empirically helps elicit explicit 3D structures from the recomposed features.The final training objective is $\mathcal{L} = \mathcal{L}_{\text{slog}} + \mathcal{L}_{\text{HN}}$

4 Experiments

4.1 Datasets

We evaluate the proposed method on two standard MDE benchmarks:

NYU-Depth v2 [34]. An indoor dataset with 464 scenes; we follow standard practice and use 24,231 training images and 654 test images at 640×480 resolution with depth range 0-10m. **KITTI** [11]. An outdoor driving dataset with stereo images and Velodyne LiDAR scans. We use images resized to 1241×376

Table 4: Zero-shot generalization on Argoverse. Models trained on KITTI are directly evaluated on Argoverse without fine-tuning.

Method	AbsRel↓	RMSE↓	SIlog↓	δ_1↑
BTS [16]	0.307	15.98	0.518	0.307
AdaBins [2]	0.383	17.07	0.523	0.383
P3Depth [24]	0.277	17.97	0.441	0.277
NeWCRF [51]	0.311	15.75	0.468	0.311
iDisc [26]	0.560	12.18	0.334	0.560
DINOv3-L [35]	0.219	11.13	0.317	0.677
DINOv3-L+LFR	**0.204**	**10.830**	**0.309**	**0.730**

Table 5: Ablation studies. Starting from the baseline (DINOv3 + DPT), we progressively add each component of our method.

Last layer	Recombine	HN loss	Multi-level	AbsRel↓	SqRel↓	RMSE↓	SIlog↓	δ_1↑
×	×	×	×	0.046	0.107	1.794	0.062	0.986
✓	×	×	×	0.045	0.105	1.743	0.061	0.987
✓	✓	×	×	0.043	0.099	1.721	0.060	0.987
✓	✓	✓	×	0.043	0.099	1.719	0.059	0.987
✓	✓	✓	✓	0.043	0.099	1.711	0.059	0.988

and adopt the Eigen split [9] (23,158 training and 652 test samples), depth range 0-80m.

For zero-shot evaluation, we employ two additional datasets: **SUN RGB-D** [36] for indoor generalization, depth range 0-10m (resized to 640×480, official test set of 5,050 samples).**Argoverse** [7] for outdoor transfer, depth range 0-150m (images resized to 1241×376; follow evaluation protocol in [26] to select 476 test samples).

4.2 Implementation Details

Our implementation uses PyTorch and training is performed on four NVIDIA A100 GPUs. We train for 30 epochs using AdamW [21] ($\beta_1 = 0.9, \beta_2 = 0.999$) with batch size 8. Learning rates are 10^{-5} for the backbone and 10^{-4} for adapters and the prediction head. We apply a linear warmup over the first 10% of iterations followed by cosine annealing; the backbone is frozen after 20 epochs. Data augmentation includes random rotation, scaling, and photometric jittering (brightness, gamma, saturation, hue). Training one epoch takes 7 minutes; a single inference on a 640×480 image costs approximately 680 GFLOPs.

4.3 Comparison on NYU-Depth v2

Table 1 reports quantitative results on the NYU-Depth v2 dataset. A vanilla DINOv3 [35] backbone with a standard DPT decoder already attains competi-

Table 6: Comparison of different auxiliary layer selection strategies. 'Scores': layers are selected based on a score predicted from their class tokens. 'Node degree': layers with the smallest average similarity between class and image tokens are selected. 'Maximal-similarity': layers with the highest average similarity to the last-layer class token are selected. 'Minimal-similarity': our used strategy selecting layers with the lowest similarity to the last-layer class token.

Method	AbsRel↓	SqRel↓	RMSElog↓	RMSE↓	SIlog↓	δ_1↑
Scores	0.044	0.101	0.064	1.738	0.060	0.987
Node degree	0.044	0.100	0.064	1.721	0.060	0.987
Maximal-similarity	0.044	0.100	0.064	1.723	0.060	0.987
Minimal-similarity	0.043	0.099	0.063	1.711	0.059	0.988

tive performance; augmenting it with LFR yields consistent improvements across all standard metrics. Compared to prior SOTA methods—which often rely on multi-step generative frameworks (diffusion/autoregressive) or extensive relative-depth pretraining—our approach achieves superior or comparable accuracy while avoiding expensive iterative inference or massive pretraining. Qualitative examples (Figure 3) show improved boundary fidelity and preservation of fine geometric structures. Failure case analysis see supplementary materials.

4.4 Comparison on KITTI

Table 2 presents results on the KITTI dataset. Inserting our LFR module into DINOv3 yields substantial improvements over the baseline and outperforms all previous methods, including those based on generative modeling or large-scale pre-training. The gains are particularly pronounced for distant small objects (Figure 4), indicating enhanced geometric understanding.

4.5 Zero-Shot Generalization

To assess cross-dataset generalization, we directly transfer models trained on NYU-Depth v2 to SUN RGB-D (Table 3) and models trained on KITTI to Argoverse (Table 4), without any fine-tuning. Our method consistently outperforms prior approaches, demonstrating superior zero-shot transferability. This confirms that our feature recombination strategy effectively extracts and reinforces generic 3D geometric priors embedded in the VFMs, enabling robust adaptation to unseen environments.

4.6 Ablation Studies

Table 5 ablates the key components of our method. Starting from a strong baseline (DINOv3 + DPT), we evaluate the following variants:

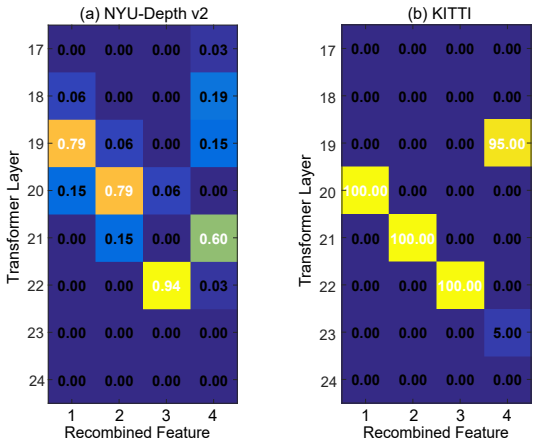

Fig. 5: Distribution of selected auxiliary layers. Statistics are computed over 100 random samples per dataset. Numbers indicate the proportion of times each layer is selected; each column sums to 1.

- Last-layer only: Replicating the last-layer features four times and feeding them into DPT already improves performance over the uniform-sampling baseline, supporting our finding that deeper layers encode rich multi-scale information sufficient for dense prediction.
- + Recombination module: Adding our feature recombination module brings a significant further gain, validating that the selected intermediate layers provide complementary information.
- + HN loss: Incorporating the hierarchical normalization loss yields additional improvement, indicating that multi-scale normalization helps elicit explicit 3D knowledge.

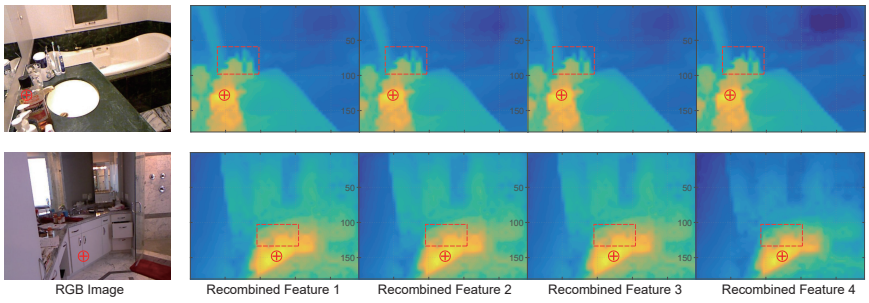

Fig. 6: Attention maps of recomposed features. For each recomposed feature level, we visualize the cosine similarity between the anchor token and all other tokens. Brighter regions indicate higher similarity.

- + Multi-level weighted prediction: Finally, using weighted aggregation of per-level predictions further boosts accuracy, demonstrating that each recomposed feature contributes useful cues.

The results obtained when setting the number of auxiliary layers to 3 ($K = 3$) see supplementary materials.

4.7 Analysis of Layer Selection

Table 6 compares different strategies for selecting auxiliary layers. All strategies outperform the last-layer-only baseline, confirming that intermediate layers provide complementary information. Among them, minimal-similarity selection achieves the best results, likely because it maximizes semantic complementarity and reduces redundancy.

Figure 5 visualizes the distribution of selected layer indices (over 100 random samples per dataset). The selected auxiliary layers predominantly lie in the middle-to-deep range (layers 17–22), which, as shown in Figure 1, encode substantial 3D knowledge but exhibit lower inter-sample variation. This suggests that injecting shared representations from these layers into the more individualized final-layer space improves prediction stability. Interestingly, the selected layers are more diverse for indoor scenes (NYU Depth v2 [34]) than for outdoor scenes (KITTI [11]), possibly reflecting the higher complexity and variability of indoor environments. Distribution of layers selected by different strategies see supplementary materials.

4.8 Attention Visualization

Figure 6 visualizes attention maps derived from the recomposed features. All maps exhibit similar spatial patterns, accurately highlighting regions corresponding to query points, with differences only in fine details. This indicates that each recomposed feature retains the strong localization capability of the last layer while offering subtle variations in visual bias, which collectively enhance robustness across diverse scenes. Global attention maps see supplementary materials.

5 Conclusion

In this work, we focused on effectively transferring the powerful DINOv3 to MDE. Recognizing that DINOv3 inherently possesses strong 3D understanding capabilities, we first conducted a comprehensive layer-wise feature analysis, revealing that deeper layers encode richer and more critical 3D geometric knowledge. Motivated by this insight, we proposed a simple yet effective method that recombines the last-layer features with carefully selected intermediate-layer features, thereby further unlocking DINOv3's inherent strengths for MDE. Our LFR module is lightweight, plug-and-play, and requires no modification to the backbone. Extensive experiments on multiple benchmarks demonstrate that our approach consistently outperforms SOTAs, achieving superior accuracy and strong

zero-shot generalization. Beyond providing an advanced tool for MDE, our work offers theoretical guidance for transferring VFMs to downstream tasks by revealing how 3D knowledge is distributed across layers. In future work, we plan to extend this method to other 3D understanding tasks and dense prediction problems, further exploring the potential of LFR.

References

1. Agarwal, A., Arora, C.: Attention attention everywhere: Monocular depth prediction with skip attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5861–5870 (2023)
2. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
3. Bhat, S.F., Alhashim, I., Wonka, P.: Localbins: Improving depth estimation by learning local distributions. In: European Conference on Computer Vision. pp. 480–496. Springer (2022)
4. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8748–8757 (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
12. Gerig, N., Mayo, J., Baur, K., Wittmann, F., Riener, R., Wolf, P.: Missing depth cues in virtual reality limit performance and quality of three dimensional reaching movements. PLoS one **13**(1), e0189275 (2018)
13. Huang, T., Zhang, Z., Tang, H.: 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. arXiv preprint arXiv:2507.23478 (2025)
14. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21741–21752 (2023)

15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
16. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
17. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European conference on computer vision. pp. 280–296. Springer (2022)
18. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet: Piece-wise planar reconstruction from a single rgb image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2579–2588 (2018)
19. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
20. Long, X., Lin, C., Liu, L., Li, W., Theobalt, C., Yang, R., Wang, W.: Adaptive surface normal constraint for depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12849–12858 (2021)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)
23. Park, N., Kim, S.: How do vision transformers work? In: 10th International Conference on Learning Representations, ICLR 2022 (2022)
24. Patil, V., Sakaridis, C., Liniger, A., Van Gool, L.: P3depth: Monocular depth estimation with a piecewise planarity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1610–1621 (2022)
25. Patni, S., Agarwal, A., Arora, C.: Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28285–28295 (2024)
26. Piccinelli, L., Sakaridis, C., Yu, F.: idisc: Internal discretization for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21477–21487 (2023)
27. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: 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)
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)

31. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
32. Ren, Z., Zhang, Z., Li, W., Liu, Q., Tang, H.: Anydepth: Depth estimation made easy. arXiv e-prints p. arXiv: 2601.02760 (2026)
33. Shao, S., Pei, Z., Wu, X., Liu, Z., Chen, W., Li, Z.: Iebins: Iterative elastic bins for monocular depth estimation. Advances in Neural Information Processing Systems **36**, 53025–53037 (2023)
34. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: 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. pp. 746–760. Springer (2012)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Song, S., Lichtenberg, S.P., Xiao, J.: 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)
37. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
39. Wang, J., Liu, J., Tang, D., Wang, W., Li, W., Chen, D., Chen, J., Wu, J.: Scalable autoregressive monocular depth estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6262–6272 (2025)
40. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
41. Wang, T., Pang, J., Lin, D.: Monocular 3d object detection with depth from motion. In: European Conference on Computer Vision. pp. 386–403. Springer (2022)
42. Wang, W., Chen, D.Y., Zhang, Z., Shi, D., Liu, A., Zhuang, B.: Zpressor: Bottleneck-aware compression for scalable feed-forward 3dgs. arXiv preprint arXiv:2505.23734 (2025)
43. Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Zhu, Z., Chen, D.Y., Zhuang, B.: Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297 (2025)
44. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14475–14485 (2023)
45. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: 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 (2024)
46. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Advances in Neural Information Processing Systems. vol. 37, pp. 21875–21911 (2024)
47. Yang, X., Ma, Z., Ji, Z., Ren, Z.: Gedepth: Ground embedding for monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12719–12727 (2023)

48. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29968–29978 (2025)
49. Ye, A., Zhang, Z., Wang, B., Wang, X., Zhang, D., Zhu, Z.: Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623 (2025)
50. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piece-wise planar 3d reconstruction via associative embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1029–1037 (2019)
51. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3916–3925 (2022)
52. Zhang, C., Yin, W., Wang, B., Yu, G., Fu, B., Shen, C.: Hierarchical normalization for robust monocular depth estimation. Advances in Neural Information Processing Systems **35**, 14128–14139 (2022)
53. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5729–5739 (2023)