Hybrid-grained Feature Aggregation with Coarse-to-fine Language Guidance for Self-supervised Monocular Depth Estimation

Wenyao Zhang[123*] Hongsi Liu[234*] Bohan Li[123*] Jiawei He[5] Zekun Qi[6]
Yunnan Wang[123] Shengyang Zhao[23] Xinqiang Yu[5] Wenjun Zeng[23] Xin Jin[23‡]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]Ningbo Key Laboratory of Spatial Intelligence and Digital Derivative, Ningbo, China
[4]University of Science and Technology of China [5]CASIA [6]Tsinghua University

Abstract

*Current self-supervised monocular depth estimation (MDE) approaches encounter performance limitations due to insufficient semantic-spatial knowledge extraction. To address this challenge, we propose Hybrid-depth, a novel framework that systematically integrates foundation models (*e.g.*, CLIP and DINO) to extract visual priors and acquire sufficient contextual information for MDE. Our approach introduces a coarse-to-fine progressive learning framework: 1) Firstly, we aggregate multi-grained features from CLIP (global semantics) and DINO (local spatial details) under contrastive language guidance. A proxy task comparing close-distant image patches is designed to enforce depth-aware feature alignment using text prompts; 2) Next, building on the coarse features, we integrate camera pose information and pixel-wise language alignment to refine depth predictions. This module seamlessly integrates with existing self-supervised MDE pipelines (*e.g.*, Monodepth2, ManyDepth) as a plug-and-play depth encoder, enhancing continuous depth estimation. By aggregating CLIP's semantic context and DINO's spatial details through language guidance, our method effectively addresses feature granularity mismatches. Extensive experiments on the KITTI benchmark demonstrate that our method significantly outperforms SOTA methods across all metrics, which also indeed benefits downstream tasks like BEV perception. Code is available at* https://github.com/Zhangwenyao1/Hybrid-depth.

1. Introduction

Monocular depth estimation (MDE) is a challenging task involving the accurate prediction of per-pixel depth values within a scene from a single image. It plays a critical role in advancing 3D vision, especially in applications such as autonomous driving [59, 60, 76, 107, 123], robotics [87, 114, 131, 132, 134, 143], and 3D reconstruction [18]. Recent advancements in deep learning methodologies have enabled supervised MDE algorithms [3, 4, 82, 116] to achieve notable improvements in depth estimation accuracy. Nonetheless, the reliance on large-scale datasets with per-pixel depth annotations for supervised training remains a significant bottleneck, limiting their practical applicability in realistic scenarios.

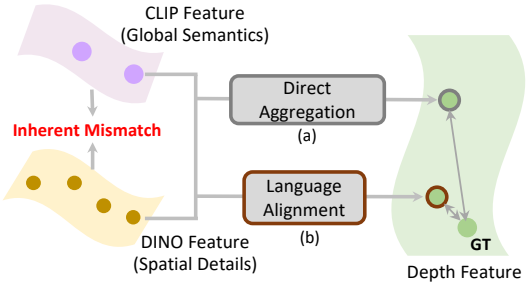

Figure 1. CLIP and DINO exhibit complementary strengths: CLIP excels in capturing global semantic context, while DINO specializes in local spatial detail extraction. However, their fusion is hindered by inherent feature-level mismatches. Direct aggregation strategies like channel concatenation (Fig. 1(a)) result in suboptimal depth representations due to misaligned semantic and spatial features. In contrast, our approach (Fig. 1(b)) employs depth-aware language prompts as a granularity calibrator to align cross-level features into a unified depth hierarchy, ensuring semantic coherence and spatial precision.

To address these limitations, numerous studies [5, 9, 27, 142] have explored self-supervised MDE by exploiting 3D geometric consistency in monocular video sequences. For instance, existing methods typically employ separate network branches to predict depth and camera pose [27, 142], optimizing the model via image reconstruction and smoothness losses. However, this paradigm suffers from con-

*Equal contribution. ‡Corresponding author.

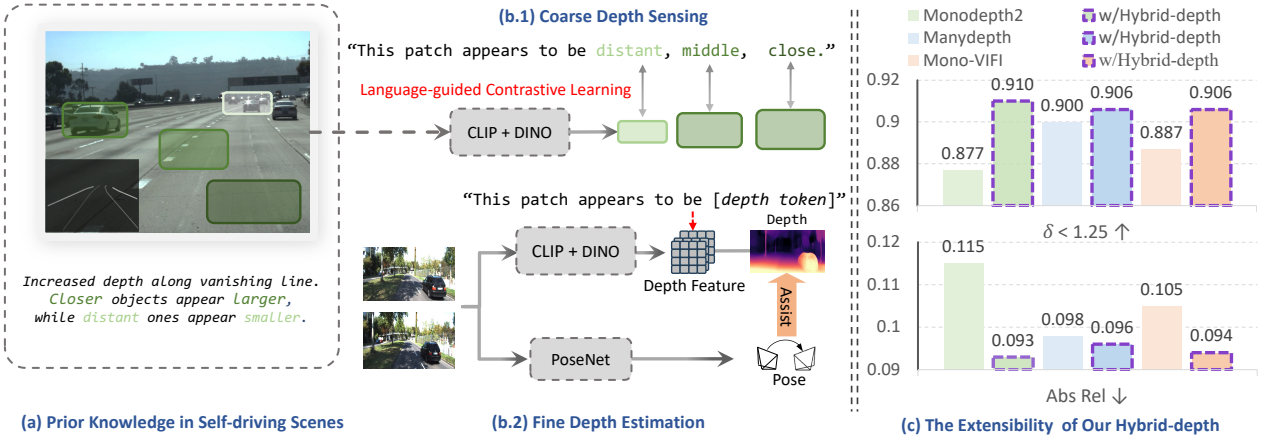

Figure 2. The detailed pipeline of our proposed method. We first aggregate different-grained features from CLIP and DINO for coarse depth sensing under contrastive language guidance (Fig. 2(b.1)), incorporating prior geometric knowledge (Fig. 2(a)). These models are then optimized with the help of the auxiliary camera pose from PoseNet like existing methods to learn a fine depth estimation (Fig. 2(b.2)). It also can be extended to improve the capture of continuous depth variations. By equipping our method, existing self-supervised MDE methods (*e.g.*, Monodepth2 [27], ManyDepth [112], and Mono-VIFI [70]) achieve significant performance improvements (Fig. 2(c)).

strained data diversity and limited model capacity, resulting in performance gaps compared to supervised approaches.

The remarkable success of foundation models such as CLIP [88] and DINO [8, 77], pre-trained on massive datasets and subsequently fine-tuned for downstream tasks, has revolutionized a range of vision applications including zero-/few-shot classification [51, 68, 141], detection [30, 138], and segmentation [15, 72, 91]. This raises the question: "Can large-scale pre-training on coarse data, followed by task-specific fine-tuning, advance self-supervised MDE ?" Recent works [21, 53, 101] highlight the complementary strengths of CLIP and DINO: CLIP captures global semantic context (*e.g.*, object categories, scene layouts) through text-image alignment, while DINO excels at local spatial detail extraction (*e.g.*, edges, local geometry) via self-supervised patch-level learning. This synergy suggests a promising pathway for improving self-supervised MDE. However, as shown in Fig. 1, naive feature fusion (*e.g.*, direct concatenation) often produces suboptimal results due to inherent feature-level mismatches. Specifically, CLIP's global semantic representations lack spatial precision for depth estimation, whereas DINO's local features fail to encode contextual depth hierarchies [46, 100, 127]. Consequently, harmonizing these complementary capabilities remains a pivotal challenge for advancing self-supervised MDE.

To address these issues, we propose Hybrid-depth, a novel framework that systematically integrates foundation models (*e.g.*, CLIP and DINO) to extract visual priors and contextual information for MDE. Our approach introduces a coarse-to-fine learning scheme that progressively endows CLIP and DINO with depth estimation capabilities while fully exploiting their hybrid-grained features through language alignment. As illustrated in Fig. 2, Hybrid-depth consists of two phases: 1). **Coarse depth sensing**. This stage aims to harness geometric priors from autonomous driving scenes (*e.g.*, depth gradients along vanishing lines) to design a proxy task with close-distant patch comparison. Specifically, we extract multi-grained features from CLIP and DINO along lane markings in images and align these feature patches with depth-related textual prompts (*e.g.*, "This patch appears to be [*close/far/very far*]") to formulate two novel contrastive learning losses. Unlike previous 2D CLIP-based methods [1, 41, 130] that employ foundation models solely as depth encoders, this stage equips CLIP and DINO with coarse depth reasoning ability through structured semantic guidance. 2). **Fine depth estimation**. This stage refines Hybrid-depth to achieve precise depth estimation. Different from previous methods, Hybrid-depth not only combines camera pose information but also conducts pixel-wise alignment with learnable depth prompts for hybrid-grained features. Due to the modular independence from typical self-supervised MDE pipelines, Hybrid-depth can be seamlessly integrated as an enhanced depth encoder to boost performance in existing frameworks. Additionally, the coarse-to-fine language guidance harmonizes multi-grained features by acting as a granularity calibrator (Fig. 1), balancing the model's focus between high-level semantic richness (from CLIP) and low-level spatial precision (from DINO). This structured multi-scale feature handling maximizes the complementary strengths of both models. The key contributions of our work are summarized as follows:

- To the best of our knowledge, our framework is the first work that leverages foundation models (CLIP and DINO) for self-supervised MDE. Our approach introduces a coarse-to-fine learning framework that progressively transfers 2D priors from foundation models to enhance 3D geometric perception.
- By synergistically aggregating CLIP's semantic context

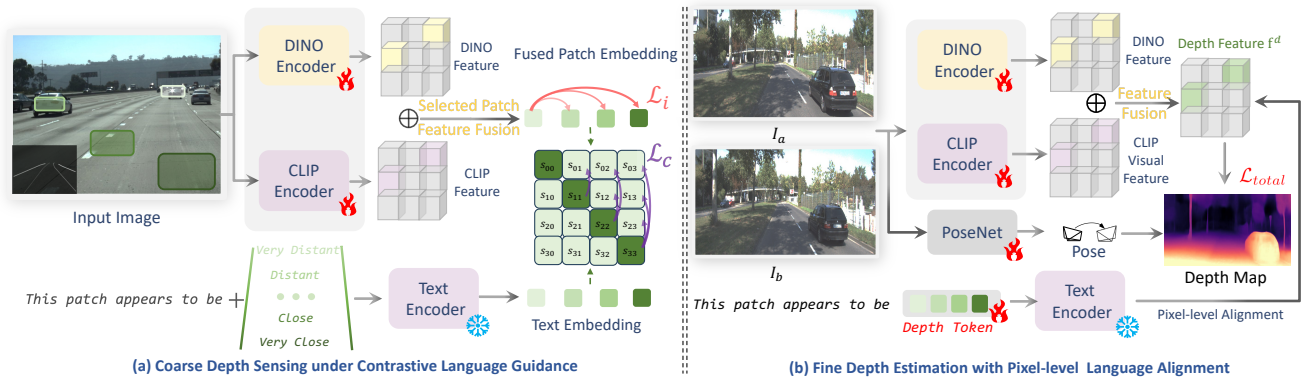

Figure 3. Left: For the coarse depth sensing stage, we first aggregate the CLIP and DINO features and then design two contrastive learning strategies to endow them with coarse depth sensing capabilities by leveraging geometric priors from self-driving scenes. Right: During the fine depth estimation phrase, different from previous methods, Hybrid-depth not only combines camera pose information from PoseNet but also conducts pixel-wise alignment with learnable depth prompts for hybrid-grained features to learn a fine depth estimation ability.

and DINO's spatial details through language guidance, we defectively resolve feature granularity mismatches.

- Our method functions as an efficient depth encoder compatible with existing self-supervised MDE pipelines, achieving significant performance improvements.

2. Related Works

2.1. Monocular Depth Estimation

2.1.1. Supervised Monocular Depth Estimation

Early researchers [39, 69, 93] conduct monocular depth estimation based on traditional computer vision techniques and hand-crafted features. However, these methods encounter limitations due to their reliance on explicit depth cues. With the development of deep learning, certain approaches can efficiently learn regressing depth from numerous annotations [25]. In detail, Eigen *et al.* [19] first proposes a multi-scale fusion network to regress the depth value from the RGB images. Based on this, many researches concentrate on improving the accuracy [6, 12, 49, 50, 52, 55, 63, 78, 80, 83, 94, 96, 108, 126] and generalization [17, 28, 75, 113, 124]. Furthermore, some works [3, 24, 64] propose conducting depth regression as a classification task, and other methods choose to introduce more priors like segmentation [95], geometry[120], and better objective functions [115]. Recently, many studies [4, 20, 34, 40, 57, 76, 90, 105, 116, 117, 121, 147] propose pre-trained models on numerous depth labels that would work well in generalizable metrics. Nevertheless, collecting vast amounts of manually labeled data is resource-consuming.

2.1.2. Self-supervised Monocular Depth Estimation

To address the above limitations inherent in the supervised MDE, Zhou *et al.* [142] introduces incorporating an additional PoseNet to learn the 6-DoF camera pose from monocular videos, enabling the prediction of depths. Building on this foundation, many advanced techniques [5, 10, 13, 16, 22, 27, 32, 33, 58, 74, 84, 98, 103, 104, 106, 122] are developed to boost the performance. Other studies leverage supplementary supervisory information, such as traditional stereo matching [111] and semantic cues [47, 54]. Many-depth [112] proposes employing multiple frames available at test time and leveraging geometric constraints by constructing a cost volume, resulting in superior performance. To tackle the well-known edge-fattening issue, Bello. *et al.* [29] and Zhu *et al.* [145] utilize an occlusion mask to remove the incorrect supervision under photometric loss. Furthermore, some works investigate to estimate depths in more challenging environments like bad weather [92], night time [102] and indoor environments [43, 136]. However, these methods typically encounter performance bottlenecks and overfitting dilemmas due to limited data.

2.2. Foundation Models and Prompt Learning

Recently, CLIP [88] and ALIGN [44] leverage large-scale image-text datasets containing numerous samples to learn a joint representation by contrastive learning. By mapping image and text information into a common space to compute the similarity of the text and image, the models can gain a more comprehensive understanding of visual and textual inputs. Additionally, single-modal methods focus on learning from visual data alone: MAE [38] reconstructs masked image patches to model spatial context, and MoCo [37] builds dynamic dictionaries via momentum contrast to capture discriminative features. DINO [8, 77] uses self-distillation with multi-crop augmentation to learn hierarchical visual patterns. Furthermore, these pre-trained foundation models also show remarkable transferability for novel datasets [44, 88] and other downstream tasks [14, 61, 62, 66, 67, 73, 85, 86, 91, 118, 129, 146].

Prompt learning is tailored to optimize text prompts to facilitate better downstream performance while maintain-

ing the generalization of models. Inspired by the recent advances [56, 81, 81] in NLP, prompt learning has become prevalent in computer vision and multimodal models. Zhou *et al.*[140, 141] introduce prompt learning to CLIP, which changes the input text from the handcrafted template to learnable context vectors and gains significant improvements. Following this, many advanced techniques [11, 45, 51, 51, 71, 133] are proposed to boost performance. Additionally, researchers investigate employing prompt learning to benefit different downstream tasks [62, 119, 144]. However, there are few studies on self-supervised MDE using foundation models.

DepthCLIP [130] is a pioneering attempt to conduct zero-shot depth estimation using CLIP, but the predicted depth is ambiguous. Furthermore, Dylan *et al.* [1], Hu *et.al.* [41] and Eunjin *et.al.* [99] reveal that learnable prompts outperform the human handcraft template, but they still underperform previous MDE methods. It implies existing self-supervised MDE methods with CLIP do not fully unleash the power of the vision-language model and knowledge [125]. Therefore, we explore employing foundation models to facilitate self-supervised MDE.

3. Hybrid-depth

We present Hybrid-depth, a novel framework that first aggregates multi-grained features from CLIP (global semantics) and DINO (local spatial details) under contrastive language guidance (Sec. 3.1). Then, we further optimize the model with the help of the auxiliary camera pose information and align the hybrid-grained features with language guidance. Additionally, it could serve as a plug-and-play depth encoder to improve the capture of continuous depth variations, thereby enhancing the accuracy of self-supervised MDE methods (Sec. 3.2).

3.1. Coarse Depth Sensing

As shown in Fig. 3, we use the geometric priors in self-driving scenes, where depth increases along lane markings, to serve as a practical proxy for depth supervision because lane labels are more accessible than precise depth measurements.

Given an image I from the lane detection datasets such as TuSimple, we separately extract global semantics features f^{clip} using CLIP and the spatial details features f^{dino} with DINO, as illustrated in Fig. 4:

$$\begin{aligned} f^{clip} &= \mathbf{F}^{clip}(\mathbf{I}) \\ f^{dino} &= \mathbf{F}^{dino}(\mathbf{I}). \end{aligned} \tag{1}$$

Considering the final features from the model's last layer have limited information, we obtain multi-scale feature maps from four ResNet blocks of CLIP and DINO 2-, 5-, 8-, and 11-th layers. To match the spatial dimensions of the DINO features, we interpolate f^{clip} to the size of f^{dino} and then concatenate them to acquire the fused features f.

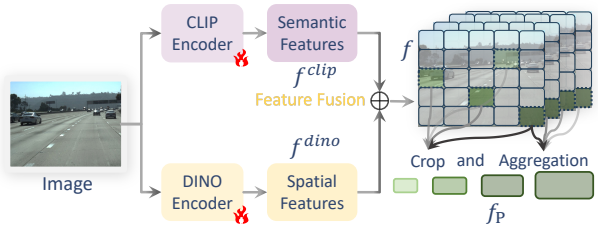

Figure 4. Patch selecting and feature concatenating.

3.1.1. Patch Selecting

We crop a set of patches $f_{\mathbf{P}} = [f_{\mathbf{P}_0}, f_{\mathbf{P}_1}, ..., f_{\mathbf{P}_{N-1}}]$ along the lane labels on the fused features f, where N is the number of patches (here, $N = 7$). These selected patches adhere to the following rules:

- For any two patches ($\mathbf{P}_i$ and $\mathbf{P}_j$, $0 \leq i < j \leq N-1$), the position of $\mathbf{P}_i$ must be higher (*e.g.*, smaller y coordinate) than that of $\mathbf{P}_j$.
- They are chosen randomly on the column and regularly internally across the row. This sample strategy could mitigate the issue of patches exhibiting nearly identical depth since the dataset is ego-view and vehicles at similar depths typically appear in the same row.

3.1.2. Depth Prompt Design

The original CLIP is not designed for depth estimation, it is necessary to customize effective depth text prompts. As mentioned above, we have successfully collected a series of patches whose depths are ordinal/ordered. The depth of different patches is ordinal, and the higher patches correspond to a more or equal depth. Therefore, we construct ranking-depth text prompts to describe the ordinal relationship of image patches. The depth text prompt is formatted as "This patch appears to be [*depth token*]", where [*depth token*] represents a set of rank distances like "*very distant, ..., close, very close*". These prompts are then fed into the text encoder to output depth text rank embeddings $\mathbf{T} = [\mathbf{T}_0, \ldots, \mathbf{T}_{N-1}] \in \mathbb{R}^{N \times C}$.

3.1.3. Contrastive Learning

Unlike the original CLIP [88], which relies solely on a single contrastive mechanism, We introduce two complementary types of contrastive learning loss to refine its ability to differentiate depth-related features:

- **Intramodal Contrastive Learning**: Within the visual modality, we enforce depth-aware representation learning by defining constraints on patch embeddings. Given a set of extracted features $f_{\mathbf{P}}$, we use the patch index to guide the feature alignment. Formally, we compute the similarity between patches as follows:

$$s_{i,j}^{intra} = f_{\mathbf{P}_i} \cdot f'_{\mathbf{P}_j}, \tag{2}$$

where $0 \leq i \leq N-1$ and $0 \leq j \leq N-1$. We introduce an intramodal contrastive loss to ensure that the self-similarity of a patch (which serves as a baseline) is higher than its similarity with patches corresponding to different depth levels. Specifically, we hope the similarity matrix $\mathbf{S}$ is a specific ordinal matrix as shown in Fig. 3:

$$\mathcal{L}_i = \sum_{i=0}^{N-1} \max\Big(0,\, s_{i',i}^{intra} - s_{i,i}^{intra}\Big), \tag{3}$$

where $0 \leq i^{'} \leq i \leq N-1$. This loss penalizes cases where a patch with a lower index (assumed to be closer) is less similar to its own representation than a patch with a higher index (assumed to be farther). It ensures that the learned features reflect the natural depth ordering present in the image.

- **Language-guided Contrastive Learning**: In addition to intramodal contrastive learning, we use language-guided contrastive learning to align visual features with corresponding depth-related text embeddings. In our framework, each patch is associated with a textual descriptor (*e.g.*, "near" or "distant") based on its relative depth position, which is determined by the patch index, which serves as a proxy for depth. This association allows us to ensure that a patch's visual features are more similar to its corresponding text embedding than to other depth descriptors. Furthermore, the depth-related text can serve as a granularity calibrator to promote the fusion of CLIP and DINO. We first compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ with elements defined as:

$$s_{i,j}^{cross} = f_{\mathbf{P}_i} \cdot \mathbf{T}_j^{\prime}. \tag{4}$$

Due to the inheritance of ranking relationships from the patches and depth text prompts, we hope the similarity matrix $\mathbf{S}$ is a specific ordinal matrix as follows:

$$\begin{aligned} & s_{i',i}^{cross} \leq s_{i,i}^{cross} \\ & \mathcal{L}_c = \sum_{i=0}^{N-1} \max\Big(0,\, s_{i',i}^{cross} - s_{i,i}^{cross}\Big), \end{aligned} \tag{5}$$

where $0 \leq i^{'} \leq i \leq N-1$. This cross-modal contrastive loss ensures that the visual features are well aligned with the semantic depth cues provided by the text, further reinforcing depth awareness.

By integrating both intramodal and language-guided contrastive learning, we enable the CLIP and DINO encoders to effectively capture and distinguish depth variations. Notably, only the visual encoder is trained, while the text encoder remains frozen during this stage.

3.2. Fine Depth Estimation

In this phrase, we obtain a dense depth map by aligning the aggregated features with learnable depth instruction and combining an auxiliary camera pose from multi-frames as shown in Fig. 3(b). Classically, previous methods [5, 27, 106, 142] generally formulate self-supervised MDE as the minimization of a photometric reprojection error during training. For the random two consecutive frames (I_a, I_b) from the training video, they utilize separate branches - DepthNet and PoseNet - to extract depth feature and pose feature. Based on these two features, they predict the depth map $\hat{D}_a$ for I_a and estimate their relative 6-DoF camera pose $M_{a\to b}$.

3.2.1. Learnable Depth Prompt

To mitigate the limitations of human language [1, 141], we replace manually crafted depth tokens like "*close, far, ...*" with N learnable tokens. The learnable depth tokens are initialized by randomly sampling 512 elements from a normal distribution with a mean of 0 and a standard deviation of 0.02. After being processed by the text encoder, we obtain the depth text embedding $\mathbf{g} \in \mathbb{R}^{N \times C}$.

3.2.2. Pixel-level Language Alignment

Given input $\mathbf{I}$ from the self-supervised MDE dataset, we first aggregate hybrid-grained features from DINO and CLIP to acquire four scales aggregated feature $\mathbf{f}^d$. To maintain the global semantic understanding of CLIP and the local spatial reasoning of DINO, we employ depth instruction as a granularity calibrator by aligning the aggregated features $\mathbf{f}$ with the corresponding learnable depth prompt:

$$\mathbf{f}^d = \mathrm{ALIGN}\left\langle \mathbf{f}, \mathbf{g} \right\rangle. \tag{6}$$

Specially, for the operation $\mathrm{ALIGN}\left\langle \mathbf{f}, \mathbf{g} \right\rangle$, we adopt following steps:

- Reshape $\mathbf{f}$: Initially, we reshape the $\mathbf{f}$ into a matrix with dimensions $C \times HW$.
- Inner Product with Depth Text Embedding: we compute the inner product between $\mathbf{f}$ and the transpose of the depth text embedding. This operation yields a new matrix $\mathbf{f}$ with dimensions $N \times HW$.
- Recover Reshaping: We reshape $\mathbf{f}$ back into the desired feature tensor $\mathbf{f}^d$ and replace the depth feature from the original DepthNet in existing self-supervised MDE methods.

Finally, we predict depth map $\hat{D}$ by up-sampling $\mathbf{f}$ following DPT [89] with the help of auxiliary camera pose.

3.2.3. Reconstruction and Smoothness Loss

Following previous methods [27, 142], we optimize the model using image reconstruction loss [137] and smoothness loss [26]. Concretely, we warp I_b into I_a to generate the reconstructed image $\hat{I}_b$ according to the predicted depth map $\hat{D}_a$ and camera pose $M_{a\to b}$ as following equation:

$$\hat{I}_b = I_a \left\langle \mathrm{proj}\left(\hat{D}_a, M_{a\to b}, K\right) \right\rangle, \tag{7}$$

Method	W × H	Train	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
GeoNet [122]	640 × 192	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Johnston *et al.* [48]	640 × 192	M	0.106	0.861	4.699	0.185	0.889	0.962	0.982
PackNet-SfM [31]	640 × 192	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
MonoFormer [2]	640 × 192	M	0.104	0.846	4.580	0.183	0.891	0.962	0.982
DIFFNET [139]	640 × 192	M	0.102	0.764	4.483	0.180	0.896	0.965	0.983
BRNet [36]	640 × 192	M	0.105	0.698	4.462	0.179	0.890	0.965	0.984
FeatureNet [97]	640 × 192	M	0.104	0.729	4.481	0.179	0.893	0.965	0.984
MonoViT [135]	640 × 192	M	0.099	0.708	4.372	0.175	0.900	0.967	0.984
Lite-Mono [128]	640 × 192	M	0.107	0.765	4.561	0.183	0.886	0.963	0.983
DualRefine [22]	640 × 192	M	0.103	0.776	4.491	0.181	0.894	0.965	0.983
Dynamicdepth [23]	640 × 192	M	0.096	0.720	4.458	0.175	0.897	0.964	0.984
SQLdepth [109]	640 × 192	M	0.094	0.697	4.320	0.172	0.904	0.967	0.984
RPrDepth [35]	640 × 192	M	0.097	0.658	4.279	0.169	0.900	0.967	**0.985**
Manydepth [112]	640 × 192	M	0.098	0.770	4.459	0.176	0.900	0.965	0.983
w/ Hybrid-depth	640 × 192	M	0.096	0.665	4.192	0.170	0.906	0.968	**0.985**
Mono-ViFI [70]	640 × 192	M	0.105	0.708	4.446	0.179	0.887	0.965	0.984
w/ Hybrid-depth	640 × 192	M	0.094	0.658	4.168	0.169	0.906	0.968	**0.985**
Monodepth2 [27]	640 × 192	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
w/ Hybrid-depth	640 × 192	M	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**	**0.970**	**0.986**

Table 1. **Quantitative results** on KITTI Eigen split [19] with the state-of-the-art self-supervised MDE methods. "$W \times H$" denotes the resolution of input images. The column of "train" specifies the way of training, where "M" represents optimizing the model using monocular video. All results are Post-Processed [26]. The best results are in **bold**; the second best is underlined. The methods integrate with Hybrid-depth modules and outperform all previous methods by a large margin on all metrics.

Method	W × H	Train	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
DepthCLIP [130]	-	0-shot	0.473	6.007	12.958	0.680	0.281	0.531	0.696
Hu *et al.* [41]	704 × 352	1-shot	0.384	4.661	12.290	0.632	0.312	0.569	0.739
Auty *et al.* [1]	640 × 192	M	0.303	6.322	-	-	0.550	0.830	0.938
Monodepth2 w/ Hybrid-depth	640 × 192	M	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**	**0.970**	**0.986**

Table 2. Comparison with previous MDE methods using CLIP. We outperform all previous methods by a large margin on all metrics.

where $\langle\rangle$ is the differentiable bilinear sampling operator following [27]. To evaluate the reconstructed images $\hat{I}_b$, we use L1-norm distance in pixel space and SSIM [110] to construct photometric error function following [26, 137]:

$$\mathcal{L}_{pe} = \frac{\beta}{2}\left(1 - \text{SSIM}\left(I_b, \hat{I}_b\right)\right) + (1-\beta)\left\|I_b - \hat{I}_b\right\|_1, \quad (8)$$

where $\beta = 0.85$ by default and SSIM computes pixel similarity over a 3×3 window. To improve the smoothness of the predicted depth map, we adopt the edge-aware smoothness loss [26] to regularize the predicted depth map:

$$\mathcal{L}_{smooth} = |\partial_x d_t^*|\, e^{-|\partial_x I_t|} + |\partial_y d_t^*|\, e^{-|\partial_y I_t|}. \quad (9)$$

Overall, the total loss $\mathcal{L}_{total}$ for training is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{pe} + \lambda * \mathcal{L}_{smooth}, \quad (10)$$

where λ is set to 0.001 by default. To reduce the trainable parameters, we only train the visual encoder, learnable tokens, and up-sample layers while freezing the text encoder.

4. Experiments

4.1. Implementation Details

Unless otherwise specified, we use the ResNet-50-based CLIP [88] visual encoder and DINOVv2 [77] encoder as the backbone in the image branch and vanilla CLIP text encoder. Our models are implemented in PyTorch [79] and trained for 10 epochs with a batch size of 16. We adopt AdamW optimizer with an initial learning rate of 1e-4 and StepLR policy. All the experiments are conducted on a single NVIDIA A800 GPU. Moreover, the number of learnable depth and pose tokens is set to 256.

4.2. Comparison with State-of-the-art

4.2.1. Results on KITTI

Table 1 presents a comparison between three classical MDE methods - Monodepth2 [27], ManyDepth [112] and Mono - VIFI [70]—augmented with our framework, and recent state-of-the-art self-supervised MDE approaches. Our proposed method not only enhances the performance of the original methods but also outperforms all other approaches by a significant margin across all evaluation metrics. In addition, Table 2 reports the same metrics for few-shot or supervised monocular depth estimation methods that utilize CLIP. Although Auty *et al.* [1] and Hu *et al.* [41] achieve MDE in a supervised manner, their performance remains substantially lower than that of current self-supervised approaches. In contrast, Hybrid-depth consistently achieves higher accuracy across all metrics, demonstrating that our

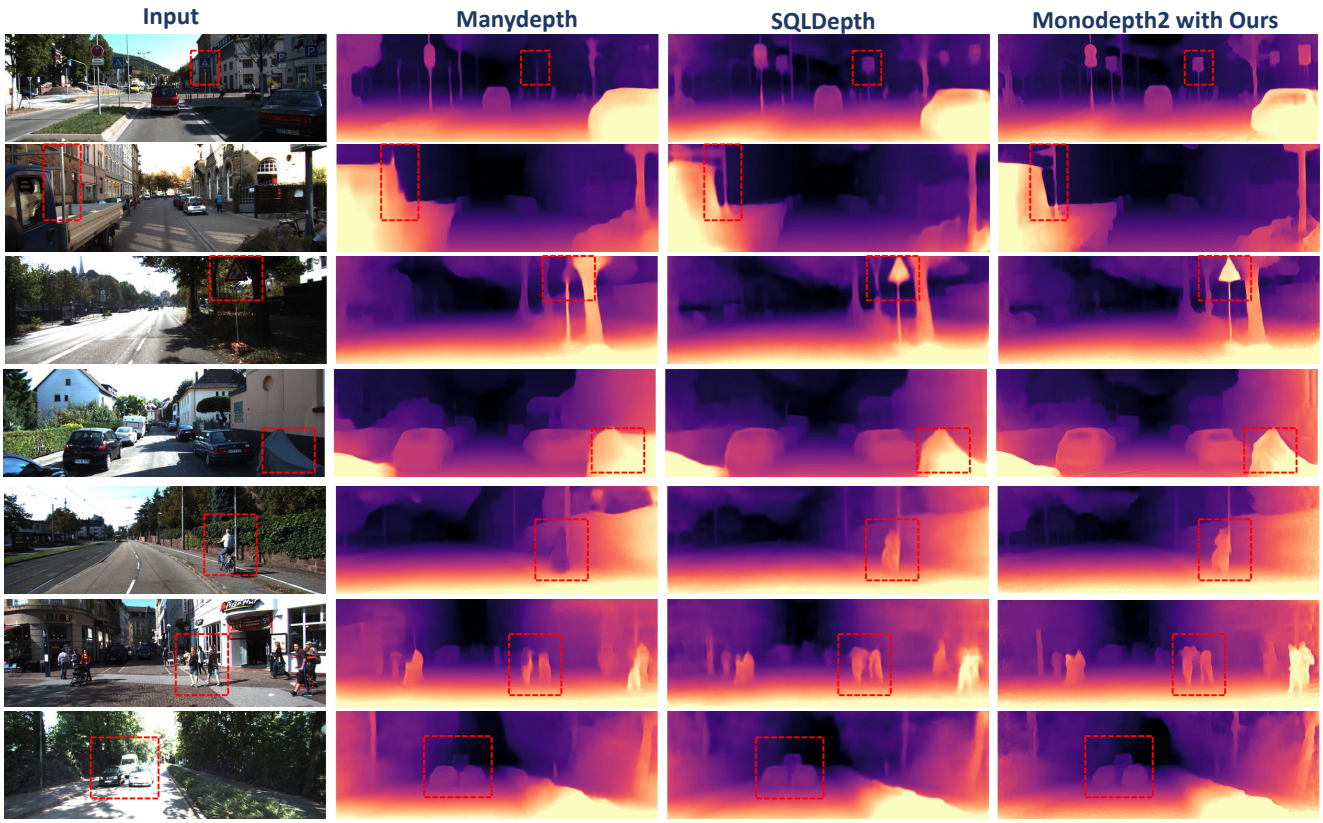

Figure 5. Qualitative comparison with Manydepth [112], SQLDepth [109] and Monodepth2 with Hybrid-depth on the KITTI dataset. Monodepth2 with Hybrid-depth accurately predicts continuous depth in the ground region while preserving sharp object edges.

method effectively leverages the power of large-scale models and pre-training with depth contrastive learning. Additionally, we find Hybrid-depth promotes the convergence of the existing self-supervised MDE methods.

4.2.2. Qualitative Results

Fig. 5 presents qualitative comparisons between Monodepth2 with Hybrid-depth and previous approaches on challenging KITTI images. The results demonstrate that Monodepth2 with Hybrid-depth achieves superior structural and edge preservation compared to its counterparts. It produces smoother depth estimations in ground regions while maintaining sharp object boundaries, effectively recovering fine details, and significantly reducing depth artifacts around object edges.

4.3. Ablation Studies

In this section, we design the experiments to investigate the following questions:

Q1: Do gains stem from the coarse depth sensing stage, not backbone improvements? To investigate the source of performance improvement, we conduct experiments as illustrated in Table. 3. We compare the results of our approach with those obtained using the same backbone but without the coarse depth sensing stage. The findings reveal that merely using a more powerful backbone does not achieve the same level of enhancement. In contrast, the coarse-to-fine learning scheme boosts performance.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑
Manydepth [112]	0.098	0.770	4.459	0.176	0.900
w/o co	0.096	0.725	4.374	0.173	0.905
w/ Hybrid-depth	0.096	0.665	4.192	0.170	0.906
Monodepth2 [27]	0.115	0.903	4.863	0.193	0.877
w/o co	0.105	0.752	4.455	0.182	0.902
w/ Hybrid-depth	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**

Table 3. Impact of coarse depth sensing stage versus backbone strength on performance, where "w/o co" denotes that directly train the models following existing self-supervised MDE methods.

Q2: Is it necessary for language-guided contrastive learning? To evaluate the effectiveness of language-guided contrastive, we conducted ablation studies by removing the cross-modal contrastive loss to compare the performance against the full model. Furthermore, we explored the role of intramodal contrastive learning. As shown in Table 4, the model without feature alignment exhibited a noticeable decline in performance across all metrics. This suggests that aligning visual features with depth-related text cues and intramodal contrastive learning effectively enhances the model's understanding of spatial relationships and improves depth prediction.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑
ManyDepth [112]	0.098	0.770	4.459	0.176	0.900
w/ L_i	0.098	0.717	4.327	0.173	0.905
w/ L_c	0.096	0.667	4.199	0.174	0.905
w/ Hybrid-depth	0.096	0.665	4.192	0.170	0.906
Monodepth2 [27]	0.115	0.903	4.863	0.193	0.877
w/ L_i	0.098	0.663	4.236	0.172	0.902
w/ L_c	0.095	0.675	4.160	0.169	0.909
w/ Hybrid-depth	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**

Table 4. Comparison of the results with different contrastive loss.

Q3: Is depth instruction necessary as a granularity calibrator? To assess the role of depth instruction as a granularity calibrator, we conducted an ablation study by removing the depth instruction component in both stages. In the first stage, we disabled the cross-modal contrastive loss aligning CLIP-DINO features with depth-related text cues, while the fusion process and intramodal contrastive learning remained unchanged. During the fine depth estimation phrase, we used a naive fusion method without language alignment for CLIP-DINO features. The results in Table 5 show a significant performance drop when depth instruction is removed, confirming its importance for calibrating granularity between CLIP and DINO features.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑
Monodepth2 [27]	0.115	0.903	4.863	0.193	0.877
w/o co-gc	0.098	0.663	4.236	0.172	0.904
w/o fi-gc	0.100	0.717	4.284	0.173	0.905
w/ Hybrid-depth	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**

Table 5. Comparison of the results when using depth instruction as a granularity calibrator in both stages, where "w/o co-gc" and "w/o fi-gc" represent feature fusion without alignment during coarse and fine stage, respectively.

Q4: Why do we need both CLIP and DINO? To address this question, we conducted ablation experiments comparing models that use only CLIP-RN50 or DINOv2 VIT-B encoder versus our dual-encoder approach as shown in Table. 6. Our results, combined with Monodepth2 [27], reveal that relying solely on a single encoder significantly degrades performance. CLIP provides robust semantic understanding, whereas DINO excels at extracting fine-grained spatial details. By integrating both encoders, our model leverages complementary strengths to facilitate depth estimation.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑
Monodepth2 [27]	0.115	0.903	4.863	0.193	0.877
w/ CLIP	0.102	0.752	4.667	0.182	0.902
w/ DINO	0.104	0.769	4.685	0.180	0.903
w/ Hybrid-depth	**0.093**	**0.596**	**4.113**	**0.167**	**0.910**

Table 6. Comparison of the result with employing CLIP, DINO, and CLIP+DINO.

Q5: What is the effect of learnable depth tokens' count? As shown in Table 7, we investigate the impact of the number of learnable depth tokens by varying the token count in Monodepth2 [27] integrated with Hybrid-depth. The results demonstrate a non-monotonic relationship: performance improves with increasing token count but degrades beyond this optimal range (e.g., 256 tokens). This trend arises due to a capacity-overfitting trade-off: while more tokens enhance model flexibility, excessive tokens may overparameterize the depth prior space, causing overfitting. Despite this, Hybrid-depth exhibits strong robustness and outperforms previous self-supervised methods, thanks to our depth-aware prompt calibration mechanism.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑
128	0.095	0.650	4.199	0.170	**0.911**
256	**0.093**	**0.596**	**4.113**	**0.167**	0.910
512	0.097	0.658	4.182	0.172	0.908
1024	0.095	0.649	4.154	0.170	**0.911**

Table 7. Comparison of the results with variant token counts.

Q6: Can Hybrid-depth benefit other 3D perception tasks such as BEV prediction? We integrate Hybrid-depth with established BEV perception methods like BEVDet [42] and FB-BEV [65]. For fair comparison, we report results using the simplest baseline configurations for both BEVDet and FB-BEV. As shown in Table 8, our approach significantly improves the performance of both methods. These findings reveal that Hybrid-depth could effectively promote 3D perception.

Method	mAP↑	mATE↓	mASE↓	mAOE↓	NDS ↑
BEVDet [42]	0.283	0.773	0.288	0.698	0.350
w/ Hybrid-depth	**0.325**	**0.734**	**0.281**	**0.623**	**0.395**
FB-BEV [65]	0.312	0.702	0.275	0.518	0.406
w/ Hybrid-depth	**0.348**	**0.673**	**0.259**	**0.464**	**0.439**

Table 8. Comparison on the nuScenes [7] *val* set. The resolution of the input image is set to 256×704.

5. Conclusions

In this paper, we present a novel paradigm named Hybrid-depth for self-supervised MDE, which progressively equips the model with the ability to estimate depth from coarse to fine. Initially, we aggregate different-grained features from CLIP and DINO for coarse depth sensing under contrastive language guidance. Then, we fine-tune the model with an auxiliary camera pose to conduct fine depth estimation, and it could serve as an efficient depth encoder to improve the capture of continuous depth variations, thereby enhancing the accuracy of self-supervised MDE methods. Additionally, we employ depth instruction as a granularity calibrator mechanism in both stages to address the issue of granularity mismatch to help hybrid-grained features to be well harmonized together. This paradigm significantly outperforms all existing methods by a large margin while facilitating 3D recognition like BEV perception.

Acknowledgements

This work was supported by Grants of NSFC 62302246, ZJNSFC LQ23F010008, Ningbo 2023Z237 & 2024Z284 & 2024Z289 & 2023CX050011 & 2025Z038, and supported by High Performance Computing Center at Eastern Institute of Technology and Ningbo Institute of Digital Twin.

References

[1] Dylan Auty and Krystian Mikolajczyk. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2039–2047, 2023. 2, 4, 5, 6

[2] Jinwoo Bae, Sungho Moon, and Sunghoon Im. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 6

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 1, 3

[4] 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. 1, 3

[5] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9): 2548–2564, 2021. 1, 3, 5

[6] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv*, 2024. 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[9] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8001–8008, 2019. 1

[10] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Continual learning of unsupervised monocular depth from videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8419–8429, 2024. 3

[11] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 4

[12] Siyu Chen, Hong Liu, Wenhao Li, Ying Zhu, Guoquan Wang, and Jianbing Wu. D3epth: Self-supervised depth estimation with dynamic mask in dynamic scenes. *arXiv preprint arXiv:2411.04826*, 2024. 3

[13] Zhiyuan Cheng, Cheng Han, James Liang, Qifan Wang, Xiangyu Zhang, and Dongfang Liu. Self-supervised adversarial training of monocular depth estimation against physical-world attacks. *arXiv preprint arXiv:2406.05857*, 2024. 3

[14] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

[15] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 2

[16] Yue-Jiang Dong, Yuan-Chen Guo, Ying-Tian Liu, Fang-Lue Zhang, and Song-Hai Zhang. Ppea-depth: Progressive parameter-efficient adaptation for self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1609–1617, 2024. 3

[17] Yue-Jiang Dong, Fang-Lue Zhang, and Song-Hai Zhang. Mal: Motion-aware loss with temporal and distillation hints for self-supervised depth estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7318–7324, 2024. 3

[18] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 829–843, 2020. 1

[19] 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. 3, 6

[20] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21795–21806, 2024. 3

[21] Mohamed Fazli Imam, Rufael Fedaku Marew, Jameel Hassan, Mustansar Fiaz, Alham Fikri Aji, and Hisham Cholakkal. Clip meets dino for tuning zero-shot classifier using unlabeled image collections. *arXiv e-prints*, pages arXiv–2411, 2024. 2

[22] Cheng Feng, Congxuan Zhang, Zhen Chen, Weiming Hu, Ke Lu, and Liyue Ge. Self-supervised monocular depth es-

timation with dual-path encoders and offset field interpolation. *IEEE Transactions on Image Processing*, 34:939–954, 2025. 3, 6

[23] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *Computer Vision – ECCV 2022*, pages 228–244, Cham, 2022. Springer Nature Switzerland. 6

[24] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 3

[25] 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. 3

[26] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 5, 6

[27] 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*, pages 3828–3838, 2019. 1, 2, 3, 5, 6, 7, 8

[28] Juan Luis Gonzalez Bello, Jaeho Moon, and Munchurl Kim. Self-supervised monocular depth estimation with positional shift depth variance and adaptive disparity quantization. *IEEE Transactions on Image Processing*, 33:2074–2089, 2024. 3

[29] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020. 3

[30] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[31] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 6

[32] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020. 3

[33] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 160–170, 2022. 3

[34] Vitor Guizilini, Pavel Tokmakov, Achal Dave, and Rares Ambrus. Grin: Zero-shot metric depth with pixel-level diffusion. *arXiv preprint arXiv:2409.09896*, 2024. 3

[35] Wencheng Han and Jianbing Shen. High-precision self-supervised monocular depth estimation with rich-resource prior. In *European Conference on Computer Vision*, pages 146–162. Springer, 2025. 6

[36] Wencheng Han, Junbo Yin, Xiaogang Jin, Xiangdong Dai, and Jianbing Shen. Brnet: Exploring comprehensive features for monocular depth estimation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022. 6

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[39] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, 2007. 3

[40] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 3

[41] Xueting Hu, Ce Zhang, Yi Zhang, Bowen Hai, Ke Yu, and Zhihai He. Learning to adapt clip for few-shot monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5594–5603, 2024. 2, 4, 6

[42] Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 8

[43] Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12787–12796, 2021. 3

[44] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[45] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 4

[46] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 2

[47] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the*

European conference on computer vision (ECCV), pages 53–69, 2018. 3

[48] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

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

[50] Numair Khan, Min H Kim, and James Tompkin. Differentiable diffusion for dense depth estimation from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8912–8921, 2021. 3

[51] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 4

[52] Dunam Kim and Seokju Lee. Clip can understand depth. *arXiv preprint arXiv:2402.03251*, 2024. 3

[53] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2

[54] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[55] Arnaud Leduc, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-depth: a scalable dataset for monocular depth estimation in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3280–3292, 2024. 3

[56] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 4

[57] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 3

[58] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pages 131–148. Springer, 2024. 3

[59] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. *arXiv preprint arXiv:2412.05435*, 2024. 1

[60] Bohan Li, Yasheng Sun, Jingxin Dong, Zheng Zhu, Jinming Liu, Xin Jin, and Wenjun Zeng. One at a time: Progressive multi-step volumetric probability learning for reliable 3d scene perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3028–3036, 2024. 1

[61] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[62] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 3, 4

[63] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3

[64] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv:2204.00987*, 2022. 3

[65] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. FB-BEV: BEV representation from forward-backward view transformations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[66] Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. Crowdclip: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2893–2903, 2023. 3

[67] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3

[68] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[69] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008. 3

[70] Jinfeng Liu, Lingtong Kong, Bo Li, Zerong Wang, Hong Gu, and Jinwei Chen. Mono-vifi: A unified learning framework for self-supervised single and multi-frame monocular depth estimation. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024. 2, 6

[71] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 4

[72] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 2

[73] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *ICML*, 2023. 3

[74] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2294–2301, 2021. 3

[75] Hoang Chuong Nguyen, Tianyu Wang, Jose M. Alvarez, and Miaomiao Liu. Mining supervision for dynamic regions in self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10446–10455, 2024. 3

[76] James Okae, Bohan Li, Juan Du, and Yueming Hu. Robust scale-aware stereo matching network. *IEEE Transactions on Artificial Intelligence*, 3(2):244–253, 2021. 1, 3

[77] 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. 2, 3, 6

[78] Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9859–9869, 2024. 3

[79] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

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

[81] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 4

[82] Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. *arXiv preprint arXiv:2411.18229*, 2024. 1

[83] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024. 3

[84] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 3

[85] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pages 28223–28243. PMLR, 2023. 3

[86] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2024. 3

[87] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint arXiv:2502.13143*, 2025. 1

[88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 6

[89] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 5

[90] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[91] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2, 3

[92] Kieran Saunders, George Vogiatzis, and Luis J Manso. Self-supervised monocular depth estimation: Let's talk about the weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8907–8917, 2023. 3

[93] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 2008. 3

[94] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[95] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance as-

sisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023. 3

[96] Shuwei Shao, Zhongcai Pei, Weihai Chen, Dingchi Sun, Peter CY Chen, and Zhengguo Li. Monodiffusion: self-supervised monocular depth estimation using diffusion model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[97] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 6

[98] Haozhe Si, Bin Zhao, Dong Wang, Yunpeng Gao, Mulin Chen, Zhigang Wang, and Xuelong Li. Fully self-supervised depth estimation from defocus clue. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9140–9149, 2023. 3

[99] Eunjin Son and Sang Jun Lee. Cabins: Clip-based adaptive bins for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4557–4567, 2024. 4

[100] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025. 2

[101] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2

[102] Madhu Vankadari, Sourav Garg, Anima Majumder, Swagat Kumar, and Ardhendu Behera. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 443–459. Springer, 2020. 3

[103] Madhu Vankadari, Samuel Hodgson, Sangyun Shin, Kaichen Zhou, Andrew Markham, and Niki Trigoni. Dusk till dawn: Self-supervised nighttime stereo depth estimation using visual foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17976–17982, 2024. 3

[104] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2811–2821, 2022. 3

[105] Ning-Hsu Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *arXiv preprint arXiv:2406.12849*, 2024. 3

[106] Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21434, 2023. 3, 5

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

[108] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818, 2022. 3

[109] Youhong Wang, Yunji Liang, Hao Xu, Shaohui Jiao, and Hongkai Yu. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5713–5721, 2024. 6, 7

[110] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[111] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019. 3

[112] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 7, 8

[113] Songlin Wei, Haoran Geng, Jiayi Chen, Congyue Deng, Cui Wenbo, Chengyang Zhao, Xiaomeng Fang, Leonidas Guibas, and He Wang. D3roma: Disparity diffusion-based depth sensing for material-agnostic robotic manipulation. In *ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2024. 3

[114] 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)*, pages 6101–6108. IEEE, 2019. 1

[115] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018. 3

[116] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 3

[117] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3

[118] Zhuoyue Yang, Junjun Pan, Ju Dai, Zhen Sun, and Yi Xiao. Self-supervised endoscopy depth estimation framework with clip-guidance segmentation. *Biomedical Signal Processing and Control*, 95:106410, 2024. 3

[119] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 4

[120] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 3

[121] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 3

[122] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 3, 6

[123] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *ICLR*, 2020. 1

[124] Ziyao Zeng, Jingcheng Ni, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, and Alex Wong. Priordiffusion: Leverage language prior in diffusion models for monocular depth estimation. *arXiv preprint arXiv:2411.16750*, 2024. 3

[125] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024. 4

[126] Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. Rsa: Resolving scale ambiguities in monocular depth estimators through language descriptions. *arXiv preprint arXiv:2410.02924*, 2024. 3

[127] Bingfeng Zhang, Siyue Yu, Jimin Xiao, Yunchao Wei, and Yao Zhao. Frozen clip-dino: a strong backbone for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025. 2

[128] 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 (CVPR)*, pages 18537–18546, 2023. 6

[129] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 3

[130] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 2, 4, 6

[131] Wenyao Zhang, Shipeng Lyu, Feng Xue, Chen Yao, Zheng Zhu, and Zhenzhong Jia. Predict the rover mobility over soft terrain using articulated wheeled bevameter. *IEEE Robotics and Automation Letters*, 7(4):12062–12069, 2022. 1

[132] Wenyao Zhang, Shipeng Lyu, Chen Yao, Feng Xue, Zheng Zhu, and Zhenzhong Jia. Analysis of robot traversability over unstructured terrain using information fusion. In *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 413–418. IEEE, 2022. 1

[133] Wenyao Zhang, Letian Wu, Zequn Zhang, Tao Yu, Chao Ma, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Unleash the power of vision-language models by visual attention prompt and multi-modal interaction. *IEEE Transactions on Multimedia*, pages 1–13, 2024. 4

[134] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 1

[135] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pages 668–678. IEEE, 2022. 6

[136] Chaoqiang Zhao, Matteo Poggi, Fabio Tosi, Lei Zhou, Qiyu Sun, Yang Tang, and Stefano Mattoccia. Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16209–16220, 2023. 3

[137] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 5, 6

[138] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2

[139] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021. 6

[140] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 4

[141] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 4, 5

[142] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 1, 3, 5

[143] Hu Zhu, Chen Yao, Zheng Zhu, Zhengtao Liu, and Zhenzhong Jia. Fusing panoptic segmentation and geometry information for robust visual slam in dynamic environments. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1648–1653, 2022. 1

[144] Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, and Chunhua Shen. Segprompt: Boosting open-world segmentation via category-level prompt learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 999–1008, 2023. 4

[145] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13116–13125, 2020. 3

[146] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023. 3

[147] Ying Zou, Zhe Chen, and Fuliang Yin. High-order multiscale attention and vertical discriminator enhanced clip for monocular depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. 3