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

Supplementary Material

A. Revisit of Contrastive Language-Image Pre-training (CLIP)

As a pre-trained vision-language model, vanilla CLIP [12] consists of two main components: a visual encoder and a text encoder, both of which are pre-trained on large-scale image-text pairs. To make it clear, we take capital letters in italic type as two-dimensional maps or images. The bold font denotes high-order tensors and the calligraphic font represents a function or a neural network module.

For the ResNet-based visual encoder, the input image will be encoded into a global visual embedding $\mathbf{I} \in \mathbb{R}^{1\times C}$. For the text encoder, the input text T_i like "a photo of a c_i" (c_i denotes the class token) is first tokenized, looking up frozen pre-trained 512-dimensional tokens for each in-vocabulary word. These tokens are then fed to a standard transformer to obtain the final text embedding. Supposing that $\mathcal{G}$ denotes the textual encoder, the text embedding $\mathbf{T}_i \in \mathbb{R}^{C\times 1}$ can be formulated as:

$$\mathbf{T}_i = \mathcal{G}(T_i) \tag{1}$$

The prediction probability of class $\boldsymbol{c_i}$ is calculated :

$$p(\boldsymbol{c_i}) = \frac{\exp(\mathrm{sim}(\mathbf{I}, \mathbf{T}_i/\tau))}{\sum_{j=1}^{K}\exp(\mathrm{sim}(\mathbf{I}, \mathbf{T}_j/\tau)}. \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity between two inputs and τ is a learnable temperature parameter.

B. Implementation Details

We replace the original DepthNet/depth encoder with the same CLIP–DINO fusion module trained with our depth-contrastive scheme, regardless of whether the encoder uses a ResNet or Transformer backbone. Then we adopt the DPT head for depth reconstruction in place of the original decoder in each method, leaving other modules like PoseNet and training strategy unchanged. 2). Specifically, for Monodepth2 and Mono-VIFI, we simply replace the original depth network with the CLIP and DINO encoders, preserving all other components, including the pose and temporal-consistency branches. 3). For ManyDepth, we construct the cost volume separately using CLIP and DINO, and replace the teacher network with Hybrid-depth (monocular version).

C. Experiments Details

C.1. Datasets

C.1.1. KITTI

This dataset contains numerous driving videos in urban scenes, and it is the most widely used dataset in self-supervised MDE approaches. Following previous work [4, 17], we employ the Eigen split [2] which has 697 images for testing, and train the model on the entire 39,810 images from the training set. Depth ranges are cropped at $0.1 \sim 80$ meters, and the input/output resolution is set to 640×192.

C.1.2. NuScenes

NuScenes [1] is a large-scale autonomous driving benchmark containing data from six cameras, one LiDAR, and five radars. There are 1000 scenarios in the dataset, which are divided into 700, 150, and 150 scenes for training, validation, and testing, respectively. Therefore, NuScenes has become the most widely used dataset in Bird-Eye-View (BEV) perception [6–8].

C.2. Evaluation Metrics

In terms of self-supervised MDE, we employ four typical error metrics to quantify the disparity between predicted and ground truth depth, as outlined in [2]. These metrics include the absolute relative error (Abs Rel), the squared relative error (Sq Rel), the root mean squared error (RMSE), and the logarithmic root mean squared error (RMSE log). Additionally, three accuracy metrics are computed, which give the fraction δ of predicted depth inside an image whose ratio and inverse ratio with the ground truth is below the thresholds: 1.25, 1.25^2, and 1.25^3.

For the 3D detection task, we report nuScenes Detection Score (NDS), mean Average Precision (mAP), as well as five True Positive (TP) metrics, including mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE).

D. Experiments

D.1. The Quantitative Result on Improvement Benchmark

To reduce the influence of noise in the sparse depth from Velodyne, we also evaluate using 93% of the Eigen split

Method	W × H	Train	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
PackNet-SfM [5]	640 × 192	M	0.078	0.420	3.485	0.121	0.931	0.986	0.996
R-MSFM6 [18]	640 × 192	M	0.088	0.492	3.837	0.135	0.915	0.983	0.995
DIFFNet [16]	640 × 192	M	0.076	0.414	3.495	0.119	0.936	0.988	0.996
MonoViT [15]	640 × 192	M	<u>0.074</u>	0.388	3.414	0.115	0.938	0.989	<u>0.997</u>
Lite-Mono [14]	640 × 192	M	0.082	0.455	3.685	0.127	0.923	0.985	0.996
Mono-ViFI [9]	640 × 192	M	0.080	0.400	3.497	0.121	0.930	0.987	<u>0.997</u>
D-HRNet [10]	640 × 192	M	0.077	0.423	3.496	0.119	0.935	0.987	0.996
RA-Depth [11]	640 × 192	M	<u>0.074</u>	<u>0.363</u>	<u>3.349</u>	<u>0.114</u>	<u>0.940</u>	<u>0.990</u>	<u>0.997</u>
Monodepth2 [4]	640 × 192	M	0.090	0.545	3.942	0.137	0.914	0.983	0.995
w/ Hybrid-depth	640 × 192	M	**0.072**	**0.335**	**3.265**	**0.110**	**0.944**	**0.991**	**0.998**

Table 1. Performance comparison on KITTI [3] using improved ground truth from [13], with a resolution of 640 × 192. The best results are in **bold**; the second best is <u>underlined</u>. The methods integrate with Hybrid-depthmodules outperform all previous methods by a large margin on all metrics.

with the improved ground truth from [13] as shown in Table .1, which contains 652 test frames.

D.2. More Ablation Studies

Q1: Does the size of the visual encoder affect the performance? As shown in Fig. 1, the performance metric at $\delta < 1.25$ improves as the size of the visual encoder increases, indicating that enhanced representation learning leads to better overall performance. Moreover, our method consistently outperforms previous approaches.

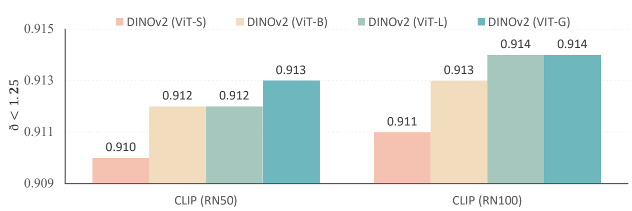

Figure 1. The metric at $\delta < 1.25$ with variant backbone size.

References

[1] 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. 1

[2] 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. 1

[3] 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. 2

[4] 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

[5] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 2

[6] 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. 1

[7] 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] Hongsi Liu, Jun Liu, Guangfeng Jiang, and Xin Jin. Mssf: A 4d radar and camera fusion framework with multi-stage sampling for 3d object detection in autonomous driving. *arXiv preprint arXiv:2411.15016*, 2024. 1

[9] 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

[10] 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. 2

[11] He Mu, Hui Le, Bian Yikai, Ren Jian, Xie Jin, and Yang Jian. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *ECCV*, 2022. 2

[12] 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. 1

[13] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 2

[14] 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. 2

[15] 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. 2

[16] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021. 2

[17] 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

[18] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12777–12786, 2021. 2