SD-ViT: Performance and Robustness Enhancements of MonoViT for Multi-View Depth Estimation

Yubo Wang, Chi Zhang, Jianhang Sun
Changchun University of Science and Technology
2022101084@mails.cust.edu.cn

Abstract

MonoViT is an enduring method in unsupervised estimation known for its robustness in disturbed data. Our aim was to explore the performance and robustness of MonoViT in full-supervised multi-view Settings. We tested both SurroundDepth training styles in the nuscenes data set using SurroundDepth as the multi-view training framework. The robustness probing sets provided by the track4 of the 2024 RoboDrive Challenge [1] demonstrate its robustness in multi-view depth estimation.

1. Introduction

At present, the self-supervised deep estimation architecture relies on sunny weather scenarios to train deep neural networks. However, in many places, this assumption is too strong for practical usages, such as autonomous driving [2–4]. For example, in the UK (2021), there were 149 days of rainfall [5]. In order for these architectures to be effective in real-world applications, it is necessary to create models that can summarize all weather conditions, time of day, and image quality. In recent years, Dosovitskiy et al. [6] proposed the ViT model, which introduces a transformer structure into computer vision and outperforms CNN in image classification tasks [7]. Thanks to its powerful modeling ability, transformer-based visual structures quickly occupy the rankings of various tasks, including object detection and semantic segmentation [8–13]. At present, research has begun to compare the robustness between ViT and CNN, and through experiments, it has been found that ViT has a stronger recognition ability on general disturbances than CNN. However, this study only draws preliminary empirical conclusions and lacks a specific analysis of each component and design unit of the ViT model. On the other hand, a large number of ViT variants have been proposed, such as Swin, PVT, etc. In last year's competition, MonoViT and its derivative methods performed exceptionally well in testing data. With the effectiveness of Transformer+CNN in foreground and background relationships, scale change perception, and other directions, their generalization performance in different weather conditions was also astonishing. Compared with the optimal method of the same period, their advantages were obvious.

Technical Report of the 2024 RoboDrive Challenge.
Track 4: Robust Depth Estimation.

2. Approach

Our proposed solution consists of a training pipeline and an image restoration module.

2.1. Training Pipeline

In our work, we tested the robustness of four different depth estimation models, with MonoViT performing best. The specific structure is shown in Figure 1 below.

MonoViT, presented at the 2022 3D Vision International Conference 3DV, has garnered considerable attention since its publication due to its remarkable accuracy and generalizability, poised to become a new benchmark method!

Compared to MonoDepth2 [14], the key innovation of MonoViT [15] lies in its deep estimation network framework design. Essentially, it employs a CNN+Transformer architecture in lieu of the previously prevalent ResNet. This integration of CNN's fine-grained detail perception with Transformer's extensive long-range feature extraction yields state-of-the-art performance on the KITTI dataset.

The encoder incorporates the MPVIT [10] structure, introduced in CVPR 2022, which primarily facilitates the seamless integration of CNN and Transformer frameworks, enhancing image feature extraction capabilities significantly. The decoder, drawing inspiration from HR-Depth's multi-scale feature fusion approach, effectively integrates features across various scales. The specific structure is shown in Figure 2.

The scheme's accuracy on the KITTI dataset is particularly impressive, achieving an Abs Rel error below 0.1 for the first time and a precision metric 1 surpassing 0.9.

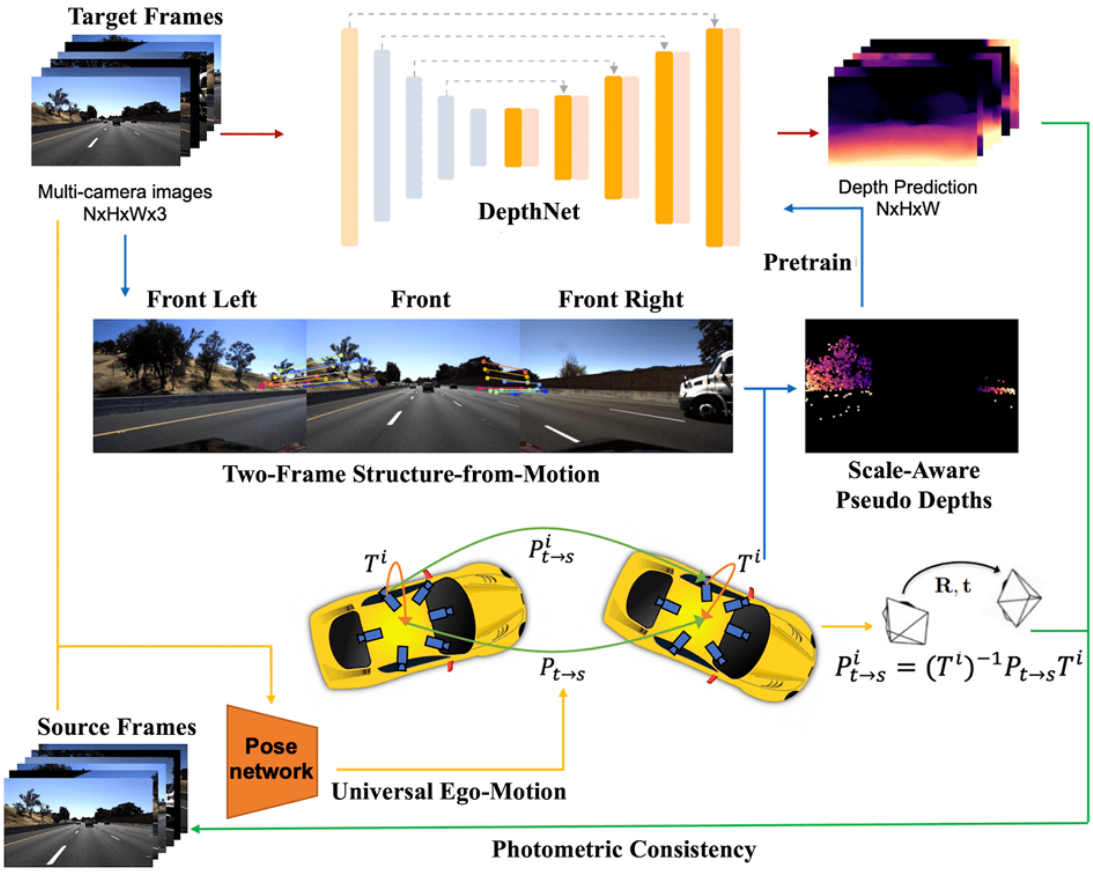

Figure 1. Overall Structure

We used two types of SurroundDepth [16] training to train the depth model as well as its reconstructed loss function.

2.2. Restoration Module

We used the Restormer to restore the image of the test data set, but we did not train the new model and used the official test model.

Restormer [17] is another masterpiece by the authors of MPRNet [18] and MIRNet [19] in the field of image restoration, and also another SOTA of Transformer technology in the low-level field. Two improvements, MDTA and GDFN, were proposed to address the challenges of Transformer in high-resolution image restoration, greatly alleviating the issues of computational complexity and GPU cache usage. The proposed solution refreshed the SOTA performance of multiple image restoration tasks.

3. Experiments

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [1] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [2] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [20–22], RoboBEV [23–25], and Robo3D [26, 27]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. For more details, please refer to the corresponding GitHub repositories.

3.2. Implementation Details

The framework is implemented based on the MMDetection3D codebase [28]. Our model consists of multiple NVIDIA A100 public servers for training, and there is no fixed number of training cards. Most of the tasks submitted are single or two cards. We believe that this does not affect

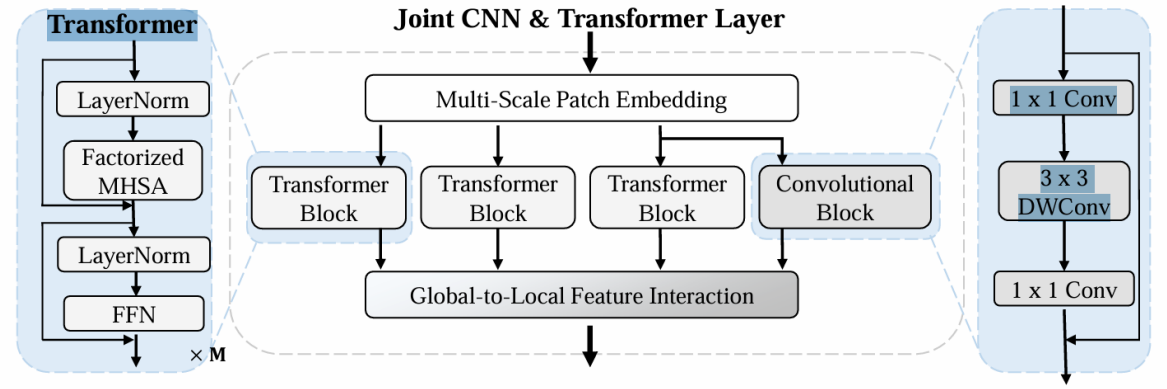

Figure 2. MPVIT

the training results, and we use the parameters of the baseline model. Resources are limited, so there is no change to the small parameters.

3.3. Main Results

Table 1 shows our ablation experiments, we did not use the training methods in the baseline model, in the first phase of the submission, and we found that this did not help improve the score. In the table, we replace several depth estimation models, and at the end, we add TTA.

	abs_rel	**sq_rel**	**rmse**	**rmse_log**
ZeroDepth	0.3853	5.7303	9.4417	0.4453
CADepth	0.3816	6.0043	9.6141	0.4497
MonoViT	0.2725	2.4170	8.1429	0.3734
MonoViT+TTA	0.2644	2.3196	7.9612	0.3632

4. Conclusion

Our method has proved the effectiveness and robustness of MonoViT in multi-view depth estimation in different data sets. We believe that improving the applicability of Transformer-like structures in dense estimation and attention modules that can improve the acquisition of network dynamic information is the key to optimizing OOD data, which is also the direction of our future work.

References

[1] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The robodrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[3] Adrien Gaidon, Greg Shakhnarovich, Rares Ambrus, Vitor Guizilini, Igor Vasiljevic, Matthew Walter, Sudeep Pillai, and Nick Kolkin. The dense depth for autonomous driving (ddad) challenge. https://sites.google.com/view/mono3d-workshop, 2021.

[4] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multi-modal data-efficient 3d scene understanding for autonomous driving. *arXiv preprint arXiv:2405.05258*, 2024.

[5] 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.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[9] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[10] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7287–7296, 2022.

[11] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[12] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[13] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[14] 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.

[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.

[16] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on Robot Learning*, pages 539–549. PMLR, 2023.

[17] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[18] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713, 2021.

[19] Le Chang, Guangyan Zhou, Othman Soufan, and Jianguo Xia. mirnet 2.0: network-based visual analytics for mirna functional analysis and systems biology. *Nucleic acids research*, 48(W1):W244–W251, 2020.

[20] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust

out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Benoit Cottereau, Lai Xing Ng, and Wei Tsang Ooi. The robodepth benchmark for robust out-of-distribution depth estimation under corruptions. https://github.com/ldkong1205/RoboDepth, 2023.

[22] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Ding Zhao, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The robodepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[23] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

[24] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[25] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robobev benchmark for robust bird's eye view detection under common corruption and domain shift. https://github.com/Daniel-xsy/RoboBEV, 2023.

[26] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[27] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robo3d benchmark for robust and reliable 3d perception. https://github.com/ldkong1205/Robo3D, 2023.

[28] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.