Fusing Features Across Scales: A Semi-Supervised Attention-Based Approach for Robust Depth Estimation

Ziyan Wang Chiwei Li Shilong Li Chendong Yuan Songyue Yang
Wentao Liu Peng Chen Bin Zhou
Beihang University
Wangzyecho@buaa.edu.cn sy2310123@buaa.edu.cn lishilong@buaa.edu.cn
19374271@buaa.edu.cn yangsy@buaa.edu.cn zyliuwt@buaa.edu.cn
cpeng@buaa.edu.cn binzhou@buaa.edu.cn

Abstract

Most existing depth estimation models are trained on "clean" datasets, resulting in a lack of robustness when encountering out-of-distribution data. To address this issue, the BUAATrans team has developed a new multi-branch network architecture called FFASDepth. This architecture uses DINOv2 and ResNet as backbones to perform multi-scale feature extraction. Our method introduces a novel channel-attention-based fusion technique that employs the input image as an embedding vector. This allows for the effective redistribution and fusion of features from both branches. Additionally, we have incorporated semi-supervised learning augmentations, namely CutFlip and AugMix. These augmentations enhance the generalization capabilities of the model. By combining these innovative strategies, FFASDepth significantly improves the robustness of depth estimation models, ensuring more reliable performance in varied and challenging environments.

1. Introduction

In the field of computer vision, depth estimation is a critical task, essential for applications such as autonomous driving [1–5] and 3D reconstruction [6–8]. However, most current learning-based depth estimation models are primarily trained and tested on clean datasets. These models often fail to fully consider the various challenges encountered in real-world applications, such as adverse weather and lighting conditions, external interference, and hardware failures [9]. This limitation often results in a lack of necessary robustness and generalization in real-world environments, potentially leading to serious safety risks.

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

To address this issue, we propose a new model variant based on the SurroundDepth [10] model—Fusing Features Across Scales Depth Estimation (FFASDepth), aimed at enhancing the robustness of autonomous driving perception models in out-of-distribution (OoD) scenarios. By utilizing a multi-branch encoder and fusing features across different scales, FFASDepth significantly enhances the model's adaptability to OoD situations. Additionally, we adopted a semi-supervised data augmentation to further enhance the accuracy and reliability of depth estimation.

FFASDepth builds upon and refines the SurroundDepth architecture, employing a multi-branch strategy to enhance depth estimation robustness and accuracy. Utilizing DinoV2 [11] and FeatUp [12], one branch learns and super-resolves unsupervised image features, restoring high-level spatial information. Simultaneously, another branch leverages ResNet [13] to extract low-level edge features. These features are fused using SENet [14] technology, which adds a simplified ResNet to encode input images into embedding vectors that act as a subtle cue for channel fusion. The channel-attention mechanism effectively manages feature prioritization, addressing information overload and guiding precise integration across branches.

To further enhance adaptability, FFASDepth uses the CutFlip operation [15] and AugMix [16]. These innovations collectively advance the model's performance in varied imaging conditions and out-of-distribution scenarios.

Our approach secured second place in the 2024 RoboDrive Challenge [17], a result that not only validates the exceptional performance of the FFASDepth model in handling OoD scenarios but also demonstrates its effectiveness and foresight in practical applications. These achievements mark significant progress in the field of autonomous driving depth estimation.

2. Related work

Monocular depth estimation based on deep learning can be divided into three main learning paradigms:

Supervised learning: In this paradigm, the model is trained by learning the mapping relationship between input images and their corresponding depth maps, with the primary goal of minimizing the difference between predicted depth and true depth. Representative studies include: Liu et al. [18] combine deep CNNs and continuous CRFs for monocular depth estimation; Laina et al. [19] use a fully convolutional residual network modified from the ResNet-50 architecture; Aich et al. [20] propose a bidirectional attention network (BANet) for monocular depth estimation. In our research, we use real values generated from point cloud data to supervise the generation of depth feature maps [21, 22].

Self-supervised learning: In the field of depth estimation, self-supervised learning typically utilizes the temporal continuity of image sequences to train models. For example, Zhou et al. [23] estimate depth and pose simultaneously using a single-view CNN; Casser et al. [24] address monocular depth and ego-motion through video sequences; Tosi et al. [25] propose an integrated framework for depth estimation, optical flow, semantic, and motion segmentation; Zhao et al. [26] develop MonoViT, a framework combining CNNs and visual transformers. In our work, we inherit the pose network architecture from SurroundDepth to train depth features.

Semi-supervised learning: This approach first trains an initial model of depth estimation using limited labeled data, then enhances the model's performance using unlabeled data. Representative works include: Kuznietsov et al. [27] develop a learning model based on sparse true depth; He et al. [28] propose a wearable monocular depth estimation system for stereo images; Ramirez et al. [29] integrate semantic information by adding a semantic segmentation decoder; Zhao et al. [30] use a synthetic data method combining image style transformation and depth estimation modules. In our research, we supervise the network model enhanced by data augmentation using depth maps obtained from training on clean datasets.

3. Approach

As illustrated in Figure 1, FFASDepth comprises three primary components. The first component features a multi-branch network architecture. The second component introduces a novel module that effectively fuses the features from the two branches, guided by the input embedding vector. The third component employs data augmentation techniques based on CutFlip and AugMix.

Multi-branch network architecture. In building upon the research presented in SurroundDepth [10], we integrate a similar ResNet-based encoder architecture within a branch of our feature extraction module. This encoder excels at delineating edge features from images; however, it is susceptible to introducing noise, particularly when processing corrupted data.

To mitigate this issue, we incorporate a self-supervised, transformer-based Dinov2 model. This model is adept at capturing high-level semantic information during training, thereby enhancing the overall robustness of our system. Despite these strengths, the Dinov2 model's coarse-grained approach to handling corrupted images limits its capacity to generate precise depth maps independently. Furthermore, due to its tendency to aggregate information across extensive areas, the features generated often lack the requisite spatial resolution for executing detailed prediction tasks such as depth estimation. To address this deficiency, we employ a super-resolution technique for feature restoration as proposed in the FEATUP [12]. Specifically, we utilize a bootstrap upsampler based on the Joint Bilateral Upsampler (JBU) [31] stack. This method leverages high-resolution signals as a guide to reinstating high-frequency details, and applying spatial weights within the neighborhoods of low-resolution feature maps. Subsequently, we downsample these enhanced features to construct a feature pyramid that aligns with the dimensions of the pyramid extracted via the ResNet encoder. This integrated approach ensures the preservation of spatial information, crucial for accurate depth tasks.

Features Across Scales Fusion. We propose a sophisticated feature fusion module, designed to enhance the integration of features derived from different neural network models, notably DINOv2 and ResNet. At the heart of this architecture lies the Feature Fusion Module, which employs a series of Squeeze-and-Excitation (SE) blocks [32] to perform attention-based fusion of features. Each SE block is meticulously engineered to refine the feature maps by leveraging both average and maximum pooling strategies, thereby capturing complementary contextual information from the input feature maps. The output from these pooling layers is subsequently processed through a sequence of fully connected layers, which apply a sigmoid activation function to generate a dynamic gating mechanism. Furthermore, a residual network implementation is utilized to encode the embedding vectors of the input image, enabling the implicit expression of the image's noise information for subsequent feature reassignment. The adjusted features from both sources are then weighted by the image feature embeddings. This weighting step is critical, as it allows the model to perform a weighted average based on the embedded image features, rather than a simple element-wise addition or multiplication. This approach ensures that the fusion process is guided by the semantic and contextual relevance of the features within the image, thereby enhancing the model's ability to focus on pertinent features while effectively disregarding irrelevant ones.

Data Augmentation. In this section, a simple CutFlip

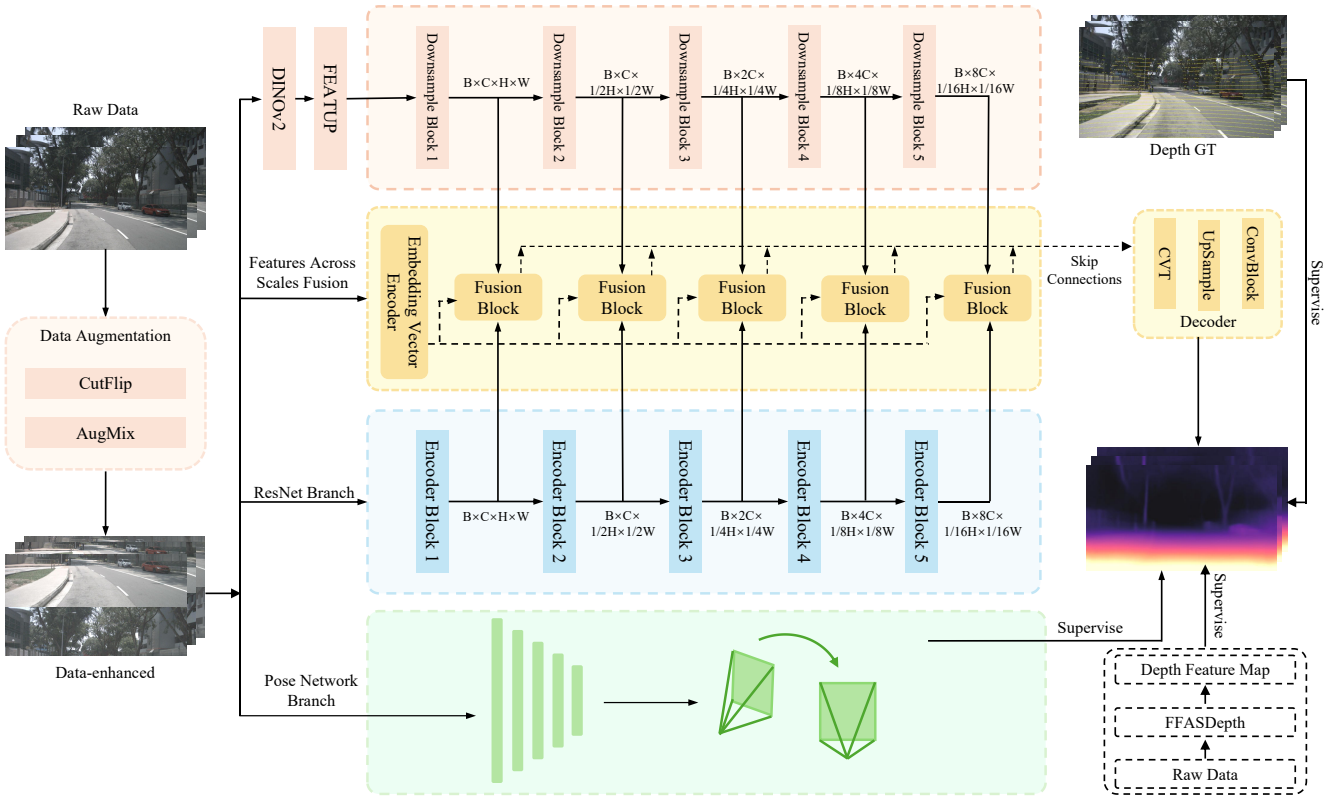

Figure 1. Overview of FFASDepth.

method is employed to enhance the diversity of the data. One of the significant risks associated with deep estimation tasks is the model's substantial reliance on vertical image positions. To encourage the model to focus on more valuable cues, the CutFlip technique divides the input sample vertically into an upper and a lower part, and then flips these two parts along the vertical axis. This weakens the relationship between depth and vertical image position, allowing the model to fit a greater variety of data types.

Furthermore, we apply the AugMix technique [16] in the data enhancement section to improve the robustness and uncertainty estimation of the image classification model. A set of augmentation transformations is selected, including 'autocontrast', 'equalize', 'posterize', 'rotate', 'solarize', 'shear', and 'translate', where the transformations do not overlap with the damage algorithm of the test set. These are identified as possible augmented atomic operations. Firstly, K weights are randomly generated according to the Dirichlet distribution for mixing different images. Then, three enhancement transforms are randomly selected to form multiple enhancement chains. One of the enhancement chains is randomly selected to enhance the image and mix the enhanced image. Finally, the weights are randomly generated according to the beta distribution, and then the image obtained above is mixed with the original image. In comparison to other enhancement methods, such as CutMix and MixUp, it addresses the issue of image distortion while maintaining image diversity.

Training pipeline. We set up a total of three steps to complete the training of our model in stages.

Step One: We follow the Self-supervised pre-training method in SurroundDepth by extracting correspondences for images with neighboring viewpoints using SIFT descriptors and converting them into sparse depths with true scales using triangulation, and using these sparse depths to pre-train the depth estimation network so that it can predict depth maps with true scales. We minimize the photometric reprojection error L_r [33]. This loss can be calculated as follows,

$$L_r = \min_{t'} \mathrm{pe}\left(I_t, I_{t'\to t}\right),$$
$$\mathrm{pe}\left(I_a, I_b\right) = \frac{\alpha}{2}\left(1 - \mathrm{SSIM}\left(I_a, I_b\right)\right) + (1-\alpha)\left\|I_a, I_b\right\|_1. \tag{1}$$

Moreover, we apply the following smoothness loss [34],

$$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|}, \tag{2}$$

In this step, the total loss is,

$$L_{step_1} = \frac{1}{N}\sum_{s=i}^{N}\left(\alpha L_r + \lambda L_s\right) \tag{3}$$

where α and λ denote adjustable hyperparameters.

Step Two: We use a Resnet and DinoV2 dual branching-based network for supervised learning under sparse point cloud data. We minimize the L1 error of the sparse point cloud concerning the generated depth map. This loss can be calculated as follows,

$$L_{l1} = \frac{1}{N} \sum_{i=1}^{N} \left| d_i - \hat{d}_i \right| \tag{4}$$

where d_i and $\hat{d}_i$ denote ground depth and estimated depth. In this step, the total loss is,

$$L_{step_2} = \frac{1}{N} \sum_{s=i}^{N} (\alpha L_r + \lambda L_s + \beta L_{l1}) \tag{5}$$

where α, λ and β denote adjustable hyperparameters.

Step Three: We integrate the data augmentation technique to bolster the robustness of our model. Specifically, we use AugMix to employ the augmentation chain strategy to produce a set of strongly augmented images. Concurrently, we utilize a simple image enhancement algorithm, which involves fine-tuning the brightness and contrast, to create a distinct set of weakly augmented images. Our model synthesizes depth maps across three distinct classes of images, denoted as D, D_w, and D_s. To enhance the stability of the model's estimation, we calculate the Kullback-Leibler (KL) divergence between these image classes and incorporate it as a consistency loss.

$$M = (D + D_{\mathrm{w}} + D_{\mathrm{s}})/3 \tag{6}$$

$$\begin{gathered} L_{KL}(D, D_w, D_s) = \\ KL(D \mid M) + KL(D_w \mid M) + KL(D_s \mid M) \end{gathered} \tag{7}$$

During the training period, the total loss adds up to the four losses mentioned above,

$$L_{step_3} = \frac{1}{N} \sum_{s=i}^{N} (\alpha L_r + \lambda L_s + \beta L_{l1} + \gamma L_{KL}) \tag{8}$$

where α, λ, β and γ denote adjustable hyperparameters.

4. Experiments

4.1. Datasets

We conducted experiments using the publicly accessible nuScenes dataset [1]. This dataset is officially divided into training, validation, and testing subsets. Specifically, we utilized 20,096 training items and 6,019 validation items from the nuScenes dataset. The dataset comprises images with a resolution of 900 × 1600 pixels. During data loading, we resized each image to various scales. We follow the protocol in the 2024 RoboDrive Challenge [17] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [1] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [9, 35, 36], RoboBEV [37–39], and Robo3D [40, 41]. 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.

4.2. Implementation Details

The model is implemented based on the MMDetection3D codebase [42]. We implemented the FFASDepth framework using PyTorch. Six NVIDIA GTX 4090 GPUs were used for model training, each with a batch size of 6. We used the FFASDepth framework for training with an epoch of 5, based on the provided Surrounddepth model pose parameters. Throughout the training process, we implemented a cosine annealing strategy to modulate the learning rate across different epochs. The learning rate initiated at 5e-5 and gradually diminished to zero. Selecting the best parameters from the previous step, we added the AugMix [16] and trained again with an epoch of 3.

4.3. Comparative Study

As shown in Table 1, our proposed FFASDepth model outperforms the baseline model in all metrics, indicating that our model can effectively enhance the robustness of the model to cope with different corruptions. We compared various modules of our model with other common and similar modules. In terms of the framework, compared to the single-branch ResNet framework provided by SurroundDepth and and single-branch DinoV2 framework, our proposed multi-branch framework that integrates DinoV2 and ResNet improved the accuracy by 10% on the corruptions dataset, effectively extracting deep feature information. In the upsampling module, we chose to compare DinoV2 with Vision Transformers for Dense Prediction (DPT) [43] and found that DinoV2 is more effective at extracting structural feature information in damaged datasets, achieving a 10% improvement in accuracy compared to DPT. for the fusion module, we adopted three different methods to fuse features extracted from different branches. These methods include direct addition of H × W values under the same channel(Addition), channel merging followed by convolution to revert to the original channels(Concatenate), and our proposed FFASDepth method. Experimental results show that our FFASDepth fusion method achieves a 5% to 10% increase in accuracy compared to the other two simple addition methods, effectively combining the features extracted by DinoV2 and ResNet. Overall, our model demonstrates superior performance and robustness on datasets with corruptions through strategic choices in framework, upsampling, and fusion modules.

Table 1. Quantitative results for different modules are presented. Note: During our comparative experiments, we ensured that the model frameworks within each module remained consistent. While, differences existed between the frameworks used in different comparisons, primarily reflected in whether depth ground truth Supervision and AugMix were employed.

comparison	Method	Abs Rel↓	Sq Rel↓	RMSE↓	log RMSE↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
		Supervision (×)			**AugMix (×)**			
Framework	ResNet	0.304	3.060	8.528	0.400	0.544	0.784	0.891
	DinoV2	0.287	3.260	8.076	0.370	0.580	0.819	0.914
	MultiBranch	0.271	2.699	7.867	0.358	0.588	0.822	0.918
		Supervision (✓)			**AugMix (×)**			
Upsampler Module	DPT	0.263	2.447	8.537	0.384	0.577	0.793	0.899
	Featup	0.242	2.247	7.697	0.353	0.631	0.837	0.918
		Supervision (✓)			**AugMix (✓)**			
Fusion Module	Addition	0.215	1.865	6.880	0.315	0.684	0.866	0.934
	Concatenate	0.223	1.946	7.217	0.326	0.668	0.858	0.931
	FFAS	0.204	1.745	6.490	0.300	0.709	0.882	0.944

Table 2. Quantitative results of ablation experiments with different modules in the FFASDepth model are presented. Notations: "Supervision" indicates that we add lidar depth ground truth in training; "AugMix" indicates that we use AugMix in the image loader.

Method	Abs Rel↓	Sq Rel↓	RMSE↓	log RMSE↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Surrounddepth	0.304	3.060	8.528	0.400	0.544	0.784	0.891
FFAS	0.271	2.699	7.867	0.358	0.588	0.822	0.918
FFAS + AugMix	0.238	2.055	7.392	0.338	0.641	0.840	0.924
FFAS + Supervision	0.242	2.247	7.697	0.353	0.631	0.837	0.918
FFAS + Supervision + AugMix	**0.204**	**1.745**	**6.490**	**0.300**	**0.709**	**0.882**	**0.944**

4.4. Ablation Study

To further validate the effectiveness of our proposed modules, we extracted relevant results of the FFASDepth model from Table 1 and recombined these with the findings from our ablation study, which are detailed in Table 2. As indicated in Table 2, the addition of the Supervision and AugMix modules individually led to an approximate 15% improvement in various metrics. When these modules were used in conjunction, the precision improved by about 40%. The results of the ablation study demonstrate that each module in our model effectively contributes to depth estimation in corrupted datasets, thereby significantly enhancing the model's robustness.

5. Conclusion

In this study, to improve the accuracy of depth estimation in corruptions datasets, we synthesized existing research on monocular depth estimation and introduced a novel method that combines multiple sub-networks and channel self-attention mechanisms. We extracted both deep and surface-level feature information from images using various networks, employing channel self-attention to weight these features effectively for accurate depth information retrieval. Subsequently, these images were enhanced for training purposes, aimed at boosting the model's robustness. The results affirm that our strategic model composition significantly improves depth estimation performance on corruption datasets, not only enhancing depth prediction accuracy but also markedly bolstering the model's resilience to dataset flaws. Ultimately, we achieved a second place ranking in Track 4 of this competition.

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. *arXiv preprint arXiv:1903.11027*, 2019.

[2] Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, and Xin Yang. Adaptive fusion of single-view and multi-view depth for autonomous driving. *arXiv preprint arXiv:2403.07535*, 2024.

[3] 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.

[4] Somnath Lahiri, Jing Ren, and Xianke Lin. Deep learning-based stereopsis and monocular depth estimation techniques: A review. *Vehicles*, 6(1):305–351, 2024.

[5] 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.

[6] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15804–15815, October 2021.

[7] 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.

[8] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3480–3495, 2024.

[9] 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.

[10] 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. *arXiv preprint arXiv:2204.03636*, 2022.

[11] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[12] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024.

[13] 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.

[14] Souvik Kundu, Shun Lu, Yuke Zhang, Jacqueline Liu, and Peter A. Beerel. Learning to linearize deep neural networks for secure and efficient private inference. *arXiv preprint arXiv:2301.09254*, 2023.

[15] Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimatione. *IEEE Transactions on Multimedia*, 2023.

[16] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[17] 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.

[18] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2016.

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision*, pages 239–248. IEEE, 2016.

[20] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11746–11752, 2021.

[21] 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 mod-

els. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[22] 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.

[23] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6612–6619, 2017.

[24] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[25] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[26] 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 *International Conference on 3D Vision*, 2022.

[27] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2215–2223, 2017.

[28] Li He, Chuangbin Chen, Tao Zhang, Haife Zhu, and Shaohua Wan. Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision. *IEEE Access*, 6:41337–41345, 2018.

[29] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantic for semi-supervised monocular depth estimation. In *14th Asian Conference on Computer Vision (ACCV)*, 2018.

[30] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9780–9790, 2019.

[31] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Trans. Graph.*, 26(3):96–es, jul 2007.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018.

[33] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation, 2019.

[34] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency, 2017.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.