Dense Depth Estimation by Fusion of Millimeter-Wave Radar Point Cloud and RGB Image Information

Xiaolong Feng, Qian Liu, Yansong Liao, Xipeng Liu, Peng Lu

Artificial Intelligence Research Institute,Xiamen University,Fujian,China
{36920231153189,36920231153215,36920231153213,36920231153216,23320231154414}@stu.xmu.edu.cn

Abstract

Depth estimation plays a pivotal role in the context of autonomous driving, offering vital distance information crucial for environmental awareness, obstacle detection, path planning, and ensuring safe driving practices. Nevertheless, traditional camera-based solutions face inherent limitations with their 2D output, lacking direct depth information. To address this shortfall, active sensors such as LiDAR and radar become indispensable for providing comprehensive assistance. Recent advancements in computer vision, particularly with visual transformer networks, have showcased remarkable performance across various tasks, notably excelling in depth prediction compared to conventional deep learning methods. This study explores the potential of visual transformers, specifically BatchFormer, in seamlessly integrating monocular images with radar-reflected point clouds to achieve robust monocular dense depth estimation. The efficacy of this innovative depth estimation approach was rigorously evaluated using the mini edition of nuScenes dataset.

Introduction

In recent years, autonomous driving technology has made significant advancements, leading the way for the future of transportation and traffic safety. Mainstream autonomous driving systems are typically equipped with a variety of sensors, including LiDAR, cameras, radar, and ultrasonic sensors, to perceive the surrounding environment. Dense depth estimation plays a crucial role in autonomous driving, as it predicts depth information for each pixel from RGB images, providing vital environmental perception data. This enables vehicles to accurately identify and track obstacles, pedestrians, and other vehicles on the road, thereby improving path planning, collision avoidance, and ensuring safe and efficient autonomous driving experiences. The capability of dense depth estimation enhances the perception abilities of autonomous driving systems, providing essential information for autonomous decision-making and behavior planning, bringing autonomous driving technology closer to achieving safe driving on real-world roads.

Challenges in dense depth estimation from RGB images include accurately reconstructing three-dimensional depth information from two-dimensional images and sensitivity to lighting conditions. Integrating information from diverse sensors, like LiDAR and millimeter-wave radar, helps address these challenges. LiDAR and millimeter-wave radar provide direct distance measurements for objects, improving the accuracy of RGB image-based dense depth estimation and enhancing performance in various lighting conditions.

As an active sensor, LiDAR can accurately measure object distances and is not affected by lighting conditions. Integrating LiDAR point cloud information contributes to improving dense depth estimation based on RGB images under both strong and weak lighting conditions. However, LiDAR is sensitive to weather conditions and performs poorly in extreme weather, such as rain, snow, or fog. Additionally, its high cost limits its widespread adoption.

LiDAR excels at measuring object distances accurately but is weather-sensitive and expensive. In contrast, millimeter-wave radar is cost-effective and thrives in extreme weather conditions. Researchers are increasingly focusing on addressing challenges related to integrating millimeter-wave radar data with RGB images and developing powerful deep learning algorithms(Lo and Vandewalle 2021, 2023). The recent introduction of 4D millimeter-wave radar has led to denser datasets with height measurement capabilities, promising more robust performance in dense depth estimation when fused with RGB images.

The main contributions of this paper include:

- We design a method for applying visual transformers to fuse images and sparse radar reflections for depth estimation. And a transformer is added to the network module for feature extraction with a batchformer to improve performance.
- Add a additional loss function for denser, dilated LiDAR ground-truth data and propose a novel loss for infinitely far regions based on semantic segmentation.

Related Work

The task of estimating depth can be categorized into two groups: depth prediction and depth completion. Depth prediction involves using camera data and ground-truth depth information during training to generate dense depth maps. In contrast, depth completion takes advantage of additional sparse depth measurements as input, typically obtained from LiDAR or radar sensors.

Camera Depth Prediction

Contemporary research primarily focuses on depth estimation using monocular camera setups, avoiding stereo information. This is typically achieved through supervised models. Notable approaches include VA-DepthNet (Liu et al. 2023b) , which excels in camera depth prediction on the KITTI benchmark, and Single Image Depth Prediction Made Better (Liu et al. 2023a) , which estimates multivariate Gaussian distributions for depth at each pixel. iDisc (Piccinelli, Sakaridis, and Yu 2023)leverages high-level environmental patterns for supervised monocular depth prediction. There is also CADepth-Net(Yan et al. 2021), a competitive self-supervised approach using image disparity between consecutive frames.

Depth Completion

we give an overview of the field of depth completion. Current approaches addressing this task are presented, and different sensors and data sources are discussed.

LiDAR Depth Completion LiDAR data is widely used in computer vision tasks due to its dense and accurate point cloud representation. However, it has limitations, such as high cost and performance issues in adverse weather conditions(Masoumian et al. 2022; Wang 2021). Recent studies have explored using LiDAR data for depth completion, with approaches like CompletionFormer(Zhang et al. 2023), DynSPN(Lin et al.), and SemAttNet(Nazir et al. 2022) emerging as state-of-the-art methods on the KITTI dataset(Geiger et al. 2013). CompletionFormer combines convolutional attention with vision transformer blocks, DynSPN combines a spatial propagation network with dynamic affinity matrices, and SemAttNet fuses RGB images, LiDAR scans, and semantic segmentation (Soydaner 2022).

Radar Depth Completion Radar sensors, compared to LiDAR, offer a more cost-effective solution for mass production and better performance in harsh weather conditions. While radar lacks the same accuracy and point cloud density as LiDAR, ongoing research explores its potential in depth estimation. Most existing approaches use fully convolutional models, with recent studies incorporating transformer architectures for improved efficiency.

Radar, Semantic Segmentation, and Depth Completion Some studies have examined combining semantic segmentation with radar data for depth completion. While these studies show the benefits of radar information for depth estimation, further quantification of the added value is needed. Innovative approaches use a coarse depth map initially predicted from camera and radar features and employ a separate branch for semantic segmentation to refine depth maps, even in challenging conditions.

Visual Transformers

Visual transformers, introduced by Vaswani et al(Vaswani et al. 2017)., have gained popularity for their ability to handle long-term dependencies. Dosovitskiy et al(Dosovitskiy et al. 2020). made significant advancements by replacing the conventional encoder with a transformer model, resulting in remarkable performance improvements in various vision tasks. Despite computational challenges in the initial implementation, recent work has addressed these issues, making transformer-based models more efficient and practical for real-world applications.

Our method

Our architecture adheres to established practices for depth estimation, employing the widely-recognized U-net architecture .This approach, based on a fully convolutional autoencoder, incorporates spatial skip connections to leverage fine-grained features. Noteworthy is the replacement of the convolutional encoder block in the U-net with the batch transformer backbone, a strategic decision in our model architecture design. This modification, detailed by (Hou et al. 2022) .The overall architecture is shown in Figure 1.

U-net Architecture

Encoder The encoder consists of four Transformer blocks designed to reduce spatial dimensions. In the original work (Yang et al. 2022), various embedding sizes and head counts were proposed as hyperparameters. Larger values for these hyperparameters often lead to better performance, albeit at the expense of increased computational complexity. Notably, in this context, we downsample the feature maps by a quarter in the first layer, as opposed to the more conventional half, aiming to reduce the computational load of subsequent attention layers.

Decoder Consisting of five corresponding upsampling stages, our approach applies bicubic interpolation layers in each step for rapid and effective upsampling. Following this, a more substantial dense block is employed for post-processing(Lee et al. 2022).

Starting from the second upsampling stage, a deep computation block is utilized to generate intermediate depth maps. The primary purpose of the first convolutional layer within each block is to reduce the dimensions or complexity of the extensive feature maps, typically containing a large number of channels (often in the hundreds). This reduction aims to transform the feature map into a new representation with a significantly smaller number of channels, typically 32 channels in this context. Another convolutional layer predicts depth values by computing a single feature map. Subsequently, a sigmoid function is applied to obtain a depth map within the (0, 1) range. These intermediate depth maps are then concatenated with the features of the current stage, serving as guiding inputs for the subsequent stages of the decoder.

Semantic segmentation Branch

We incorporated a semantic segmentation branch to investigate the possible advantages of utilizing semantic data in depth estimation. This branch is seamlessly integrated into the third stage of the depth decoder and functions alongside skip connections from the corresponding attention layers. Its primary goal is to produce two intermediary segmentation maps, which are displayed in Figure 1. These segmentation feature maps are then combined with depth feature maps and

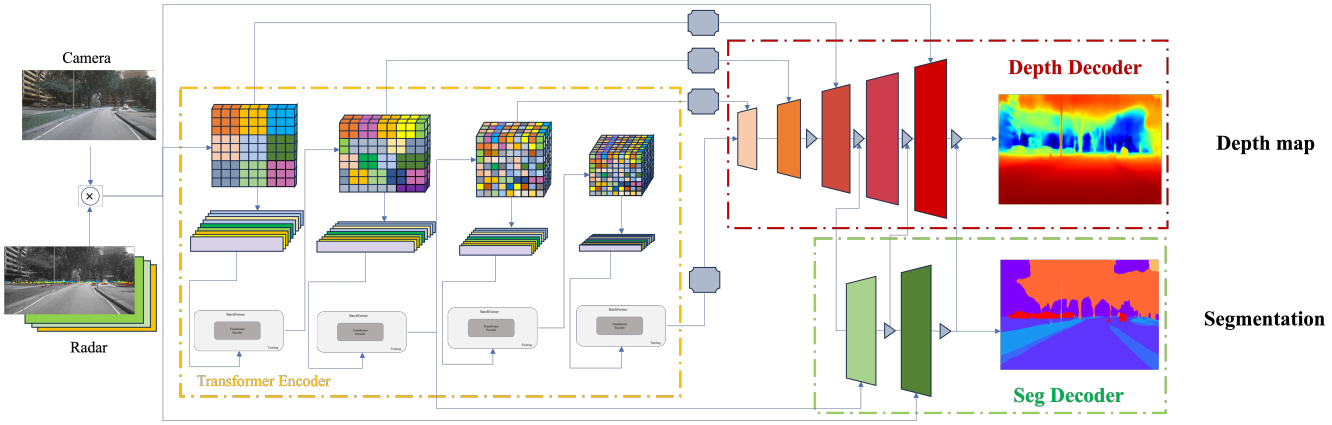


Figure 1: model architecture

forwarded to the depth activation block, resulting in the generation of the depth map. For streamlining this workflow, we reduced the number of categories to 21, aligning with the Cityscapes dataset and adhering to mseg's standard implementation.

Batch Former

In order to extend batched attention mechanisms to pixel-level feature maps, we introduced the BatchFormerv2 architecture(Hou et al. 2022).

$$\mathbf{Z}_i = softmax(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{C}})\mathbf{V}_i, \mathbf{Z} = concat(\mathbf{Z}_1, \ldots, \mathbf{Z}_N) \quad (1)$$

For a specific block with spatial dimensions H and W, the number of image blocks, N, is given by $N = H \times W$. During training, for each spatial position $i = 1, \ldots, N$, the features of a batch of blocks at the current position are treated as a sequence. That is, there are N sequences, each with a length of B. Subsequently, all these sequences are fed into a shared Transformer block.

Total Loss Functions

To optimize the depth maps, we utilize an RMSE (Root Mean Squared Error) loss function. Furthermore, we apply this reconstruction loss to a dilated point cloud, created through a 5×5 dilation with a single iteration. This augmentation significantly increases the number of relevant points, providing more robust guidance for the model to achieve superior reconstruction.

In the segmentation branch, we utilize a Focal loss (Lin et al. 2017). This modified version of the standard Cross-Entropy loss assigns greater importance to challenging examples. This approach enables the model to focus more on the minority class, leading to improved performance. This strategy is particularly beneficial for datasets with imbalanced classes, such as those encountered in autonomous driving scenarios.

Both loss functions were incorporated within a Perceptual loss (Johnson, Alahi, and Fei-Fei 2016)inspired framework, designed to enforce similarity to the target ground truth depth map in the initial layers of the decoder.

The ultimate loss function is a variant of the Contrastive loss function (Arora et al. 2019), which we refer to as the Infinity loss. The concept is to leverage pre-existing corresponding segmentation maps to identify regions with infinite depth values, such as the sky. For each pixel's value in the predicted depth map [formulate] located in these regions, it is adjusted towards zero (considering an inverted depth map where closer objects have values closer to 1). Any other pixel in the predicted depth map with a value below a certain threshold is elevated above this threshold using a Hinge loss approach (Lin 2004).For an individual pixel in the predicted depth map,the loss is given by

$$\begin{aligned}\mathcal{L}_{infinity}(\hat{y}, y, ft) = \\ (1-y) \cdot \max(0, \hat{y}) + y \cdot \max(0, ft - \hat{y})\end{aligned} \quad (2)$$

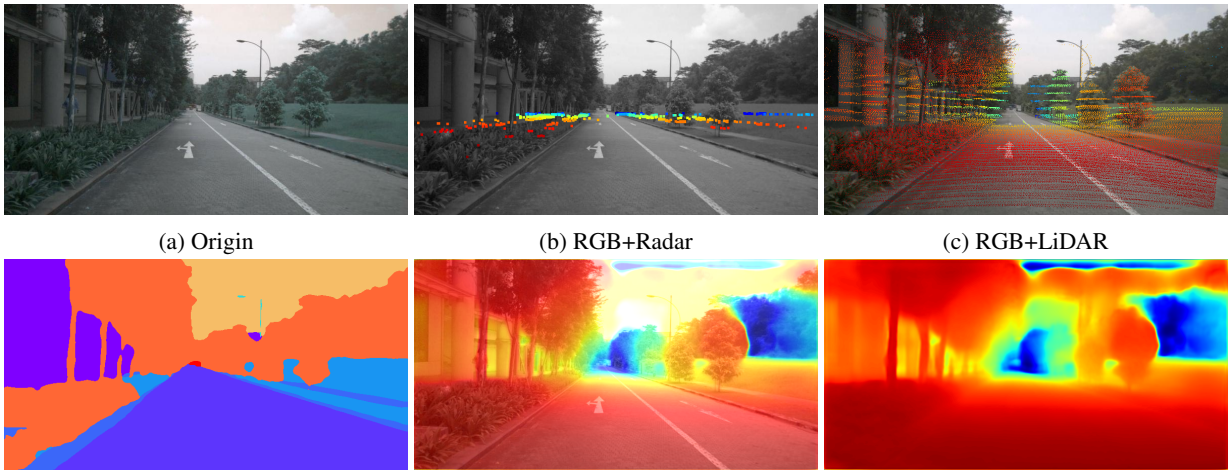
where $\hat{y}$ is the predicted value,$y \in 0, 1$ where 0 is an index for sky-segmented pixel ,1 others.ft is the true threshold for skies,chosen as low $ft = 1e^{-2}$,to penalize only the pixels that have been "segmented" as infinitedly far away.

Experiment

Datasets

The input data for our foundational model is diverse, comprising six feature maps—three sourced from the camera and three from the radar. Specifically, these include an RGB image from a monocular camera, alongside radar data containing information on distance, radial velocity, and radar flow. Our data preprocessing and ground truth generation pipeline follow the methodology outlined in (Y. Long and Narayanan 2021). We parse the description files of the nuScenes dataset to categorize the data into day-clear and challenging scenes. Subsequently, the samples are partitioned into training, validation, and test sets, maintaining a distribution ratio of 0.8, 0.1, and 0.1, respectively.

RGB Images To enhance the efficiency and precision of our analysis, we perform cropping on the original image data from the nuScenes dataset. The resolution is reduced from



(a) Origin (b) RGB+Radar (c) RGB+LiDAR

(d) Pred Seg (e) Depth On RGB (f) Depth Pred

Figure 2: Inputs and outputs of the model.

900×1600 to 416×800. This adjustment is made to focus exclusively on the pertinent section of the environment containing LiDAR ground-truth data, all the while retaining a relatively high resolution to facilitate accurate depth estimation, especially for fine details.

Radar Data Given the sparse and limited nature of raw radar data, we adopt a super-resolution technique, as proposed by Long et al(Y. Long and Narayanan 2021), to enhance the density of radar data points for each frame. This involves leveraging a consecutive sequence of frames spanning a total time of 0.3 seconds before the current frame, effectively accumulating all reflections. It is crucial to compensate for ego-motion within this timeframe. Radar sensors feature a function that enables instantaneous velocity measurement in the radial direction using the Doppler effect. However, this approach introduces a challenge, as it cannot measure the tangential speed of moving objects in the scene. Additionally, we utilize radar flow to gain a better understanding of dynamic scenes. In our approach, where a dedicated segmentation branch is employed to enrich object-level comprehension, we opt not to use MER (multi-channel enhanced radar)(Y. Long and Narayanan 2021)for the sake of clarity and conservation of computing resources.

Ground-Truth Data During the training process, obtaining ground-truth information for pixel-wise depth and semantic segmentation of the camera images is imperative. The depth-ground-truth data for each frame constitutes a feature map derived from the combination of 21 subsequent and four previous scans from a 32-beam LiDAR. These scans are accumulated and then projected onto the image plane of the RGB image, accounting for proper ego-motion and external calibration parameters. In the realm of semantic segmentation, moving vehicles undergo segmentation and motion compensation.It's noteworthy that the chosen maximum depth is 100 meters, which may result in higher errors compared to the standard 50 meters employed in other works. To address this, ground-truth values are inverted to circumvent issues such as infinity at longer distances.

Implementation Details

Experimental Settings A single 3080 GPU is utilized for both the training and testing phases in all experiments.The initial learning rate is set at $4e^{-5}$, and during the fine-tuning stage, it progressively descends from $2e^{-5}$ to $8e^{-6}$. In our exploration of various novel optimizers, we opted for the DiffGrad optimizer(S. R. Dubey and Chaudhuri 2019), incorporating the concept of leveraging the gradient norm from previous iterations(S. R. Dubey and Chaudhuri 2023) and a dynamic weight-decay coefficient. Additionally, we employ the OneCyclicLR scheduler(Smith and Topin 2018), characterized by a robust learning rate warm-up that claims to achieve super convergence. In our study, we implement a 5 epoch-long warm-up, equivalent to about 20,000 optimizer steps.

Metrics In the initial phase of our experiments, we employ the standard evaluation metric RMSE, known for its resilience to outliers compared to regular RMSE, Mean Absolute Error (MAE), and Absolute Relative Error (Abs-REL). The latter calculates the mean percentage of the prediction error.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2} \tag{3}$$

Where y_i represents the true values, $f(x_i)$ denotes the predicted values, and m signifies the quantity of test data.

Result Analysis

The model comprises a semantic segmentation branch and a depth prediction branch, where the loss is calculated as the weighted sum of both branches. The scenario pertains to a well-defined daytime field of view.Figure 2 illustrates exemplary inputs and outputs of our model, demonstrating an achieved RMSE error of 5.4.The LiDAR depth ground truth,

derived from a total of 25 LiDAR scans, is projected onto the RGB camera image. We present both a supervised segmentation prediction aiming to reconstruct the 'quasi-ground-truth'. The baseline model employed in this study is a U-net, incorporating convolutional layers and skip connections.

Conclusion

In this study, we explored the synergistic effects of integrating visual transformers, radar data, and semantic segmentation into the domain of monocular depth estimation. Our investigation revealed the transformative potential of transformers for forthcoming applications. Our experiments demonstrated that radar data significantly influences the final outcomes by enabling the model to learn correspondences and encode them directly into the weights. Despite the promising outcomes achieved, there exist promising avenues for future research to enhance performance. Primarily, two key approaches stand out for augmenting the effectiveness of our method: enhancing the quality of ground-truth data or refining the model itself.

References

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Engelmore, R.; and Morgan, A., eds. 1986. *Blackboard Systems*. Reading, Mass.: Addison-Wesley.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Hou, Z.; Yu, B.; Wang, C.; Zhan, Y.; and Tao, D. 2022. Batchformerv2: Exploring sample relationships for dense representation learning. *arXiv preprint arXiv:2204.01254*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Lee, S.; Yi, E.; Lee, J.; and Kim, J. 2022. Multi-Scaled and Densely Connected Locally Convolutional Layers for Depth Completion. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8360–8367. IEEE.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Y. 2004. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1): 73–82.

Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; and Yang, H. ???? Dynamic spatial propagation network for depth completion. arXiv 2022. *arXiv preprint arXiv:2202.09769*.

Liu, C.; Kumar, S.; Gu, S.; Timofte, R.; and Van Gool, L. 2023a. Single Image Depth Prediction Made Better: A Multivariate Gaussian Take. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17346–17356. Vancouver, BC, Canada: IEEE.

Liu, C.; Kumar, S.; Gu, S.; Timofte, R.; and Van Gool, L. 2023b. Va-depthnet: A variational approach to single image depth prediction. arXiv:2302.06556.

Lo, C.-C.; and Vandewalle, P. 2021. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3343–3347. IEEE.

Lo, C.-C.; and Vandewalle, P. 2023. RCDPT: Radar-Camera Fusion Dense Prediction Transformer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Masoumian, A.; Rashwan, H. A.; Cristiano, J.; Asif, M. S.; and Puig, D. 2022. Monocular depth estimation using deep learning: A review. *Sensors*, 22(14): 5353.

Nazir, D.; Pagani, A.; Liwicki, M.; Stricker, D.; and Afzal, M. Z. 2022. Semattnet: Toward attention-based semantic aware guided depth completion. *IEEE Access*, 10: 120781–120791.

Piccinelli, L.; Sakaridis, C.; and Yu, F. 2023. iDisc: Internal Discretization for Monocular Depth Estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21477–21487. Vancouver, BC, Canada: IEEE.

S. R. Dubey, S. K. R. S. M. S. K. S., S. Chakraborty; and Chaudhuri, B. B. 2019. diffgrad: an optimization method for convolutional neural networks,. *IEEE transactions on neural networks and learning systems,*, 31(11).

S. R. Dubey, S. K. S.; and Chaudhuri, B. B. 2023. Adanorm: Adaptive gradient norm correction based optimizer for cnns,. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 5284–5293.

Smith, L. N.; and Topin, N. 2018. Super-convergence: Very fast training of neural networks using large learning rates,.

Soydaner, D. 2022. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, 34(16): 13371–13385.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P. 2021. Research on comparison of LiDAR and camera in autonomous driving. In *Journal of Physics: Conference Series*, volume 2093, 012032. IOP Publishing.

Y. Long, X. L. M. C. P. C., D. Morris; and Narayanan, P. 2021. Radar-camera pixel depth association for depth completion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20(1): 12507–12516.

Yan, J.; Zhao, H.; Bu, P.; and Jin, Y. 2021. Channel-wise attention-based network for self-supervised monocular depth estimation. In *2021 International Conference on 3D vision (3DV)*, 464–473. London, United Kingdom: IEEE.

Yang, J.; An, L.; Dixit, A.; Koo, J.; and Park, S. I. 2022. Depth estimation with simplified transformer. *arXiv preprint arXiv:2204.13791*.

Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; and Mattoccia, S. 2023. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.