VPOcc: Exploiting Vanishing Point for 3D Semantic Occupancy Prediction

Junsu Kim[1], Junhee Lee[1], Ukcheol Shin[2], Jean Oh[2] and Kyungdon Joo[1]
[1]UNIST, [2]Carnegie Mellon University
{joonsu0109, junhee98, kyungdon}@unist.ac.kr, {ushin, hyaejino}@andrew.cmu.edu
https://vision3d-lab.github.io/vpocc/

Abstract

*Camera-based 3D semantic occupancy prediction aims to estimate dense voxel grids of 3D scenes from 2D images. It is gaining attention due to its resource efficiency compared to 3D sensors (*i.e.*, LiDAR) for understanding 3D scenes. However, due to the perspective projection, camera-based methods inherently suffer from a 2D-3D discrepancy problem, where closer objects appear larger in 2D images. To address this issue, we propose a novel framework, VPOcc, that leverages a vanishing point (VP) to mitigate the 2D-3D discrepancy. As a pixel-level solution, we introduce VPZoomer, which warps images by counteracting the perspective effect through a VP-based homography transform. In addition, as a feature-level solution, we propose VP-guided cross-attention (VPCA) that performs perspective-aware feature aggregation. Lastly, spatial volume fusion (SVF) module fuses two feature volumes extracted from original and warped images to complement each other. By effectively incorporating VP, our framework achieves improved performance in both IoU and mIoU metrics on SemanticKITTI and SSCBench-KITTI360 datasets.*

1. Introduction

Camera-based 3D semantic occupancy prediction, which estimates semantic voxel grids of 3D scenes using only RGB images, is becoming important for the safe navigation of robots and autonomous vehicles. However, camera-based methods face challenges such as missing metric-scale depth [1], feature dimension mismatch [2], and occlusion [3], which are all related to perspective geometry.

2D images are created by projecting 3D scenes onto image planes via perspective projection, causing a 2D-3D discrepancy based on distance from the camera. As shown in Fig. 1, the nearby object (left green box) and the distant object (right red box), which are similar in 3D size, appear differently sized in the 2D image. We break down this problem into two perspectives: pixel and feature levels. The pixel-level discrepancy indicates a *discrepancy in the number of pixels* between near and far object regions in the 2D image. Additionally, it causes a feature-level discrepancy, *an imbalance in feature granularity*, as a fixed kernel captures fine local details in near areas but broad global features in distant areas.

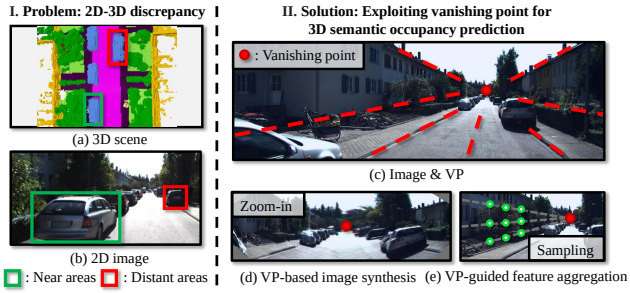

Figure 1. **Overview of the proposed VPOcc**. *Problem*: Due to perspective projection, 3D objects of similar size appear differently in a 2D image depending on their distance from the camera. *Solution*: Our framework leverages a vanishing point (VP), incorporating VP-based image synthesis and VP-guided feature aggregation to mitigate the imbalance caused by perspective projection.

Although deformable mechanisms [4] have been proposed to address the fixed receptive field problem, they may not fully account for perspective geometry. We address this using a vanishing point (VP), where 3D parallel lines converge in 2D images (see Fig. 1-II). In road scenes, a dominant VP often indicates distant areas, allowing us to roughly distinguish between near and far regions.

In this paper, we propose VPOcc, a framework leveraging VP for 3D semantic occupancy prediction to mitigate the perspective effect. Our framework includes: (1) VP-Zoomer, which creates VP-guided warped images by counteracting the perspective effect at the pixel level. (2) VP-guided cross-attention (VPCA), which samples perspective-aware offset points using VP to address the imbalance in feature granularity at the feature level. (3) Spatial volume fusion (SVF), which integrates feature volumes from the original and warped images to complement each other. As a result, we improve performance on the SemanticKITTI [5] and SSCBench-KITTI360 [6] datasets.

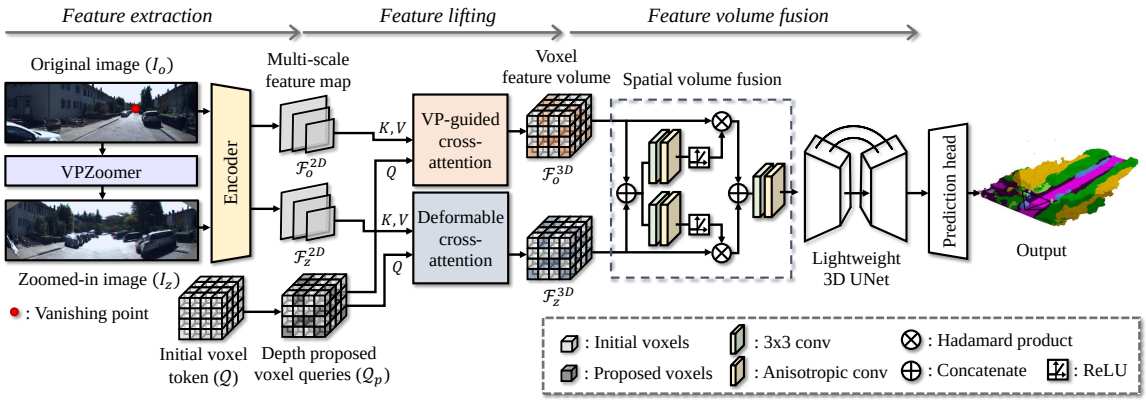

Figure 2. **Overall architecture of VPOcc**. In the feature extraction step, a zoomed-in image is generated using VPZoomer, and multi-scale feature maps $\mathcal{F}_o^{2D}$ and $\mathcal{F}_z^{2D}$ are extracted from I_o and I_z. During the feature lifting, the depth proposed voxel query $\mathcal{Q}_p$ is employed with VP-guided cross-attention (VPCA) on $\mathcal{F}_o^{2D}$ and deformable cross-attention on $\mathcal{F}_z^{2D}$ to construct the voxel feature volumes $\mathcal{F}_o^{3D}$ and $\mathcal{F}_z^{3D}$, respectively. In the feature volume fusion stage, both $\mathcal{F}_o^{3D}$ and $\mathcal{F}_z^{3D}$ are fused using the spatial volume fusion (SVF) module and refined via the 3D UNet-based decoder. Finally, the prediction head estimates the 3D semantic voxel map of the entire scene.

2. Related Work

Camera-based 3D semantic occupancy prediction [7] aims to estimate a complete 3D scene as a voxel grid of occupancy and semantics from 2D images, which indicate only a partial scene. Previous works primarily focus on transforming 2D features into 3D voxel grids. MonoScene [1] directly projects 2D features onto 3D voxels, while VoxFormer [2] aggregates 2D image features into 3D voxel space using deformable cross-attention. OccFormer [8] employs a local-global transformer decoder and Symphonies [3] proposes Serial Instance-Propagated Attentions. In contrast, we enhance 3D scene understanding from limited 2D images by incorporating VP to address perspective geometry.

For VP, prior works have explored the use of VP to improve 2D image understanding. VP-based image resampling has been employed to enhance small-object detection in 2D [9], while VP-guided motion prediction has been used to improve video segmentation accuracy [10]. Unlike these approaches, we propose a novel method that leverages VP to enhance 3D scene understanding from 2D images.

3. Method

3.1. Overview

Feature extraction. Given an input image $I_o \in \mathbb{R}^{H\times W\times 3}$ and a corresponding VP, we generate a zoomed-in image $I_z \in \mathbb{R}^{H\times W\times 3}$ using the proposed VP-based zoom-in module, VPZoomer (see Sec. 3.2). It enlarges the far areas and shrinks the near areas, simultaneously creating symmetric images around the image center via warping the original image. By utilizing both I_o and I_z, which contain differently scaled scenes based on the distance from the camera, we can leverage balanced pixel information across the scene by addressing the 2D-3D discrepancy at the pixel level. Subsequently, we use an encoder initialized with pre-trained MaskDINO [11] following Symphonies [3] to create multi-scale feature maps $\mathcal{F}_o^{2D}$ and $\mathcal{F}_z^{2D}$ for each of I_o and I_z.

Feature lifting. We construct a depth proposed voxel query $\mathcal{Q}_p$, following Symphonies [3]. First, we initialize voxel token $\mathcal{Q} \in \mathbb{R}^{X\times Y\times Z\times C}$ with learnable embeddings. We back-project depth values from the image plane into 3D space using camera parameters, converting depth values into 3D points. Then, voxel tokens overlapping with these 3D points are employed as $\mathcal{Q}_p$. With $\mathcal{Q}_p$, we aggregate multi-scale image features via cross-attention to construct 3D voxel feature volumes. In this process, our VP-guided cross-attention (VPCA) module, which samples points towards VP, is employed on $\mathcal{F}_o^{2D}$ to generate $\mathcal{F}_o^{3D} \in \mathbb{R}^{X\times Y\times Z\times C}$. With this perspective-aware 2D point sampling, we can aggregate 2D features with a more suitable granularity for 3D representation. Additionally, we utilize the general deformable cross-attention (DCA) [4] to generate $\mathcal{F}_z^{3D} \in \mathbb{R}^{X\times Y\times Z\times C}$ by aggregating 2D image features from $\mathcal{F}_z^{2D}$, which are from the image distorted by zoom-in. As a result, we obtain the voxel feature volumes $\mathcal{F}_o^{3D}$ and $\mathcal{F}_z^{3D}$ from the original and zoomed-in images.

Feature volume fusion. $\mathcal{F}_o^{3D}$ and $\mathcal{F}_z^{3D}$ are fused through the spatial volume fusion (SVF) module and a lightweight 3D UNet-based decoder. By fusing them, each voxel aggregates the balanced number of features and uniform feature granularity from the image, regardless of its distance from the camera. Lastly, the fused voxel feature volume is passed through a prediction head, which consists of 3D convolutions for upsampling to the target size, a 3D ASPP [12] block, and a 3D convolution layer to generate the 3D semantic voxel grids.

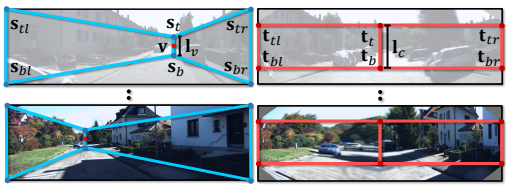

Figure 3. **Illustration of VPZoomer**. *Left:* The original image I_o with source areas ($\mathcal{S}_L$, $\mathcal{S}_R$) of blue trapezoids. *Right:* The zoomed-in image I_z with target areas ($\mathcal{T}_L$, $\mathcal{T}_R$) of red rectangles.

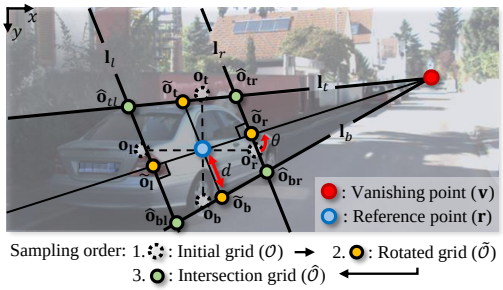

Figure 4. **VP-guided point sampling**. Centered on the reference point $\mathbf{r}$, we first generate the initial grid $\mathcal{O}$ with the offset d and rotate it by an angle θ to obtain $\tilde{\mathcal{O}}$. Next, we identify the intersection grid $\hat{\mathcal{O}}$ at the cross-point of lines from $\tilde{\mathcal{O}}$ and VP $\mathbf{v}$. As a result, a set of sampling points $\mathcal{P}$ is composed of $\{\tilde{\mathcal{O}}, \hat{\mathcal{O}}, \mathbf{r}\}$.

Training objective. Following MonoScene [1], we use scene-class affinity loss $\mathcal{L}_{\text{scal}}$ for class-wise metrics, with $\mathcal{L}^{\text{sem}}_{\text{scal}}$ and $\mathcal{L}^{\text{geo}}_{\text{scal}}$ for geometry and semantics. We also use cross-entropy loss $\mathcal{L}_{\text{ce}}$, weighted by class frequencies, for occupancy prediction. The total loss is formulated as:

$$\mathcal{L} = \mathcal{L}^{\text{geo}}_{\text{scal}} + \mathcal{L}^{\text{sem}}_{\text{scal}} + \mathcal{L}_{\text{ce}}. \tag{1}$$

3.2. VPZoomer: VP-based image warping

Due to perspective projection, 3D scenes projected onto 2D image planes exhibit pixel density imbalance based on their distance from the camera. To mitigate this, our VPZoomer generates the zoomed-in image I_z warped toward VP, resulting in the effects of compressing near areas, enlarging far areas, and ensuring horizontal symmetry (see Fig. 3).

Given the original image I_o and VP $\mathbf{v} = [v_x, v_y]^\top$, VPZoomer warps two source areas with trapezoidal shapes (blue) to target areas of rectangular shapes (red) by 2D transformation. I_z is composed by the following process:

$$\begin{aligned} I_z &= \mathbf{M} \odot I_l + (1 - \mathbf{M}) \odot I_r, \\ \text{where } I_r &= \mathbf{H}_R(I_o),\ I_l = \mathbf{H}_L(I_o), \end{aligned} \tag{2}$$

$\mathbf{M}$ is a binary mask that activates the left-half area ($W/2$), $1 - \mathbf{M}$ activates the remaining half area, and $\mathbf{H}_L(\cdot)$ and $\mathbf{H}_R(\cdot)$ represent warping functions.

3.3. VP-guided cross-attention

Recently, deformable cross-attention between 3D voxel queries and 2D image features has been commonly employed to lift 2D features into 3D [2, 3]. However, this approach only implicitly adjusts the offsets of sampling points without considering geometry between 2D and 3D (*i.e.*, perspective projection). To resolve this, we introduce the VP-guided cross-attention (VPCA) module, which samples points in a trapezoidal shape towards VP.

Method	IoU	mIoU	road	sidewalk	building	car	truck	vegetation
SemanticKITTI								
MonoScene [1]	34.16	11.08	54.70	27.10	14.40	18.80	3.30	14.90
VoxFormer [2]	42.95	12.20	53.90	25.30	19.80	20.80	3.50	22.40
OccFormer [8]	34.53	12.32	55.90	30.30	15.70	21.60	1.20	16.80
Symphonies [3]	42.19	15.04	58.40	29.30	24.70	23.60	3.20	24.20
VPOcc (Ours)	**44.58**	**16.15**	**58.90**	**32.60**	**27.42**	**25.32**	**6.13**	**26.87**
SSCBench-KITTI360								
MonoScene [1]	37.87	12.31	48.35	28.13	32.89	19.34	8.02	26.15
VoxFormer [2]	38.76	11.91	52.99	27.21	31.18	17.84	4.56	14.69
OccFormer [8]	40.27	13.81	54.30	31.53	36.42	22.58	9.89	31.00
Symphonies [3]	44.12	18.58	54.94	32.76	35.11	**30.02**	**25.07**	**38.33**
VPOcc (Ours)	**46.39**	**19.80**	**58.68**	**37.87**	**42.62**	28.73	21.71	37.77

Table 1. **Quantitative results on SemanticKITTI [5] and SSCBench-KITTI360 [6] test set.** IoU and mIoU are computed over all classes, with selected dominant (*i.e.*, road, sidewalk, vegetation, building) and rare classes (*i.e.*, car, truck) also reported.

Feature aggregation by cross-attention. Following Fig. 4, VPCA samples image features $\mathcal{F}^{2D}_o(\mathbf{p}_s)$, where $\mathbf{p}_s \in \mathcal{P}$, and d is determined by distance between $\mathbf{r}$ and $\mathbf{v}$, with θ directed toward $\mathbf{v}$. Next, using the depth-proposed voxel queries $\mathcal{Q}_p$, VPCA aggregates image features via a cross-attention mechanism and generates the voxel feature volume $\mathcal{F}^{3D}_o$ as expressed as follows:

$$\text{VPCA}(\mathcal{Q}_p, \mathbf{p}_s, \mathcal{F}^{2D}_o) = \sum_{i=1}^{N} \mathbf{A}_i \mathbf{W} \mathcal{F}^{2D}_o(\mathbf{p}^i_s), \tag{3}$$

where N is the numbers of sampling points, $\mathbf{A}_i$ represents the attention weight and $\mathbf{W}$ denotes the embedding projection weight. Otherwise, we utilize the general deformable cross-attention (DCA) [4] on $\mathcal{F}^{2D}_z$ to aggregate 2D image features from the distorted image by warping.

3.4. Spatial volume fusion

We propose the spatial volume fusion (SVF) module to fuse $\mathcal{F}^{3D}_z$ and $\mathcal{F}^{3D}_o$, extending 3D-CVF [13]. As shown in Fig. 2, the module employs 3×3 and anisotropic convolutions [14] to integrate the two feature volumes through attention masks. It leads to a voxel feature volume that is effectively fused both locally and spatially.

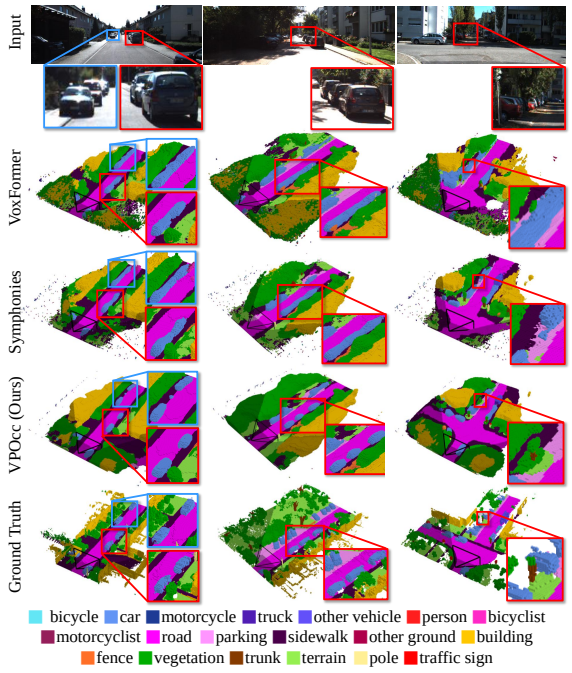

Figure 5. **Qualitative results on SemanticKITTI validation set**.

4. Experiments

We conduct experiments on SemanticKITTI [5] and SSCBench-KITTI360 [6] datasets, which contain 20 and 19 classes. We use IoU and mean IoU (mIoU) to evaluate completion and segmentation. We use estimated depth and VP from images using pre-trained MobileStereoNet [15] and NeurVPS [16], respectively.

4.1. Performance comparison

Quantitative results. In Tab. 1, we compare existing 3D semantic occupancy prediction methods across different datasets. We outperform the previous metrics in both IoU (+1.63) and mIoU (+1.11) on the SemanticKITTI test set, as well as IoU (+2.27) and mIoU (+1.22) on the SSCBench-KITTI360 test set. Unlike previous methods (*e.g.*, VoxFormer [2], Symphonies [3]) that excel in only a single metric, VPOcc outperforms in both IoU and mIoU metrics.

Qualitative results. Fig. 5 compares results of different methods on the SemanticKITTI validation set. Boxed areas highlight the ability of VPOcc to distinguish objects along the road by effectively leveraging VP to mitigate 2D-3D discrepancies caused by perspective effects.

4.2. Ablation studies

Architecture composition. Tab. 2-(a) validates the effectiveness of our VPOcc modules: VPZoomer, VPCA, and SVF. VPZoomer addresses pixel density discrepancies in method (1), and VPCA mitigates feature granularity imbalances caused by perspective projection, resulting in both IoU (+0.21) and mIoU (+0.48) improvements in method (2). Additionally, the parameter count is reduced as VPCA is composed of fewer parameters compared to the general DCA. Finally, SVF spatially fuses the two feature volumes to construct the perspective-aware feature volume, leading to gains in both IoU (+0.80) and mIoU (+0.57).

Method	VPZoomer	VPCA	SVF	IoU	mIoU	Params (M)
(1)	✓			43.97	15.31	105.78
(2)	✓	✓		44.18	15.79	105.72
VPOcc (Ours)	✓	✓	✓	**44.98**	**16.36**	110.60

(a) **Ablation study for the proposed framework.**

Range	$0m-17m$		$17m-34m$		$34m-51.2m$	
Metric	IoU	mIoU	IoU	mIoU	IoU	mIoU
MonoScene [1]	39.05	12.49	38.52	12.22	31.83	8.57
VoxFormer [2]	**48.60**	12.95	46.83	13.85	35.85	9.54
OccFormer [8]	38.66	15.38	38.51	13.85	31.24	10.69
Symphonies [3]	39.71	16.28	47.19	16.36	38.04	11.19
VPOcc (Ours)	44.97	**17.12**	**49.66**	**18.11**	**39.74**	**13.22**

(b) **Depth-wise performance evaluation.**

Table 2. **Additional experiments of our proposed framework on SemanticKITTI validation set.**

Depth-wise performance evaluation. Tab. 2-(b) shows that our method improves performance across depth ranges, particularly in mIoU. This results from effectively utilizing VP to distinguish pixel distances from the camera, incorporating perspective geometry.

5. Conclusion

We propose VPOcc, the camera-based 3D semantic occupancy prediction framework that leverages VP to address the 2D-3D discrepancy from perspective projection. VPZoomer creates VP-based warped images to balance pixel density, and VP-guided cross-attention (VPCA) conducts perspective-aware feature aggregation for uniform feature granularity. Additionally, spatial volume fusion (SVF) resolves 2D-3D discrepancies by effectively fusing feature volumes. By incorporating these components, VPOcc achieves superior performance in both IoU and mIoU.

References

[1] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *CVPR*, 2022. 1, 2, 3, 4

[2] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *CVPR*, 2023. 1, 2, 3, 4

[3] H. Jiang, T. Cheng, N. Gao, H. Zhang, T. Lin, W. Liu, and X. Wang, "Symphonize 3d semantic scene completion with contextual instance queries," in *CVPR*, 2024. 1, 2, 3, 4

[4] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv*, 2020. 1, 2, 3

[5] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of lidar sequences," in *ICCV*, 2019. 1, 3, 4

[6] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu, Y. Wang, H. Zhao, Z. Yu, and C. Feng, "Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving," in *IROS*, 2024. 1, 3, 4

[7] H. Xu, J. Chen, S. Meng, Y. Wang, and L.-P. Chau, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *Information Fusion*, vol. 114, p. 102671, 2025. 2

[8] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2304.05316*, 2023. 2, 3, 4

[9] A. Ghosh, N. D. Reddy, C. Mertz, and S. G. Narasimhan, "Learned two-plane perspective prior based image resampling for efficient object detection," in *CVPR*, 2023. 2

[10] D. Guo, D.-P. Fan, T. Lu, C. Sakaridis, and L. Van Gool, "Vanishing-point-guided video semantic segmentation of driving scenes," in *CVPR*, 2024. 2

[11] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *CVPR*, 2023. 2

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2017. 2

[13] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *ECCV*, 2020. 3

[14] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3d semantic scene completion," in *CVPR*, 2020. 3

[15] F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, "Mobilestereonet: Towards lightweight deep networks for stereo matching," in *WACV*, 2022. 4

[16] Y. Zhou, H. Qi, J. Huang, and Y. Ma, "NeurVPS: Neural vanishing point scanning via conic convolution," in *NeurIPS*, 2019. 4