Monocular 3D Object Detection with Viewpoint-Invariant Inter-Object Estimation for Better Contextual Behavior Understanding

Minghan Zhu
University of Michigan
minghanz@umich.edu

Abstract

We introduce a model developed for monocular 3D object detection with an emphasis on inter-object estimation. Inter-object estimation, for example, of the relative pose between objects, allows a self-driving car to better understand the behavior of road users in the surrounding environment, which is essential for realizing automated driving that is both safe and consistent with the expectations of the local traffic. Furthermore, we observe that the inter-object estimation is invariant to viewpoint changes and is intuitive to understand from vision signals. Combining the architecture of transformers and graph networks, our model learns inter-object features from images to better model the inter-object relations. We further propose new metrics to evaluate inter-object estimation. Through experiments on the nuScenes dataset, our model outperforms the baseline in both 3D object detection and inter-object estimation.

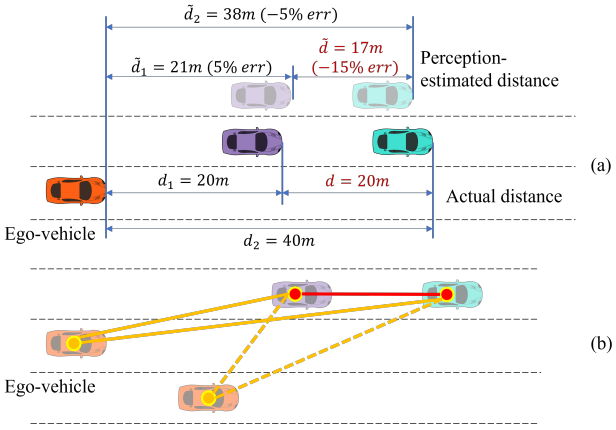

Figure 1. Inter-object relative poses is subject to accumulated error from the individual estimations. In Fig. 1 (a), a 5% error in the estimated distance from ego-vehicle to two vehicles can result in a 15% error in the estimated distance between those two vehicles. In this work, we focus on the inter-object estimation, as highlighted in red in Fig. 1 (b), which is invariant to ego-vehicle's viewpoint.

1. Introduction

For self-driving cars, knowing the pose of objects in the surrounding environment allows safe navigation without crashing into obstacles. However, safety is not the only requirement for self-driving. Since self-driving cars are expected to operate in a traffic environment with human-driving vehicles before they completely replace human-driving, it is important for self-driving algorithms to accommodate the human-driving conventions, which may change spatially and temporally. For example, in a highway car-following scenario, a self-driving car is expected to maintain a headway (measured in distance and/or time) similar to other vehicles, which changes depending on the geographic location and the time of the day. Failure to observe and adhere to the local consensus of the traffic could result in discomfort and safety hazards for both the self-driving car and the surrounding traffic. Such an ability to *blend-in* the local traffic with similar driving behaviors to the average human drivers is called *roadmanship* in literature [7, 19, 26]. It requires an understanding of the interaction among the road users in the environment, for which an accurate estimation of the relative pose between the road users is essential.

However, estimating the relative pose between the road users is non-trivial. Estimating the pose of individual objects and calculating the relative pose through subtraction could lead to magnified error. An illustration is shown in Fig. 1 (a). Regarding this, we propose to model the inter-object relative poses in the monocular 3D object detection network. With the perception algorithm built with an awareness of object relations, we expect this to help the self-driving algorithms achieve better roadmanship. On the other hand, there could be some advantages of directly estimating the relative poses. First, it is intuitive from vision to tell the relative pose between two vehicles if they are close to each other, for example, when they are platooning or parked side-by-side in a parking lot. Second, while the absolute pose estimation heavily depends on the camera's

intrinsic and extrinsic parameters, the projective geometry, and the depth estimation of the model, the inter-object relative pose is viewpoint-invariant, as illustrated in Fig. 1 (b), and thus could circumvent the complexity brought by the camera geometry and benefit the network performance [29].

Overall, this work has the following contributions:

- We propose viewpoint-invariant inter-object relative pose estimation as a new learning target for the monocular 3D object detection task.
- We develop an inter-object estimation module that improves the monocular 3D object detection, the relative pose estimation, and the behavior understanding.
- We propose new metrics to evaluate the inter-object relative pose estimation.

2. Related work

2.1. Network architectures

We can classify the monocular 3D object detection models regarding the network architecture. The first type is based on fully convolutional neural networks [10, 17, 28]. The input RGB images are processed by the convolutional layers, and the feature maps embedding the desired information to be estimated gradually emerge. The second network type is based on the transformer architecture [16, 24]. For monocular 3D object detection, the transformer module is mainly used in the decoder stage, where the objects are modeled as queries. The queries start from an image-independent initialization. They gradually gather information from image features and exchange information among the queries through the attention mechanism. In the end, each query predicts the class, location, size, and other target properties of an object. The query-based object modeling allows for object-centric non-local information gathering, which has become increasingly popular recently.

2.2. Geometric constraints

Since monocular 3D object detection is an underdetermined problem, another important aspect of the models is how they leverage geometry constraints to regularize the estimation. Many existing methods have explored incorporating various geometric constraints to refine the estimation or assist the network training. Some of the representative geometric constraints are listed here:

- **Prior object shape models** [1, 13, 18]. In earlier work, CAD models of objects are commonly used, which provides prior knowledge on the shape and size of objects, and the network predictions are encouraged to align with the CAD models.
- **Dense shape prediction** [2, 5, 11]. Other than CAD models, the object shape can also be learned by the network itself. The shape, together with the pose, should render an appearance consistent with the actual image.
- **Projection of 3D bounding box vertices** [14, 15, 27, 28]. Predicting the 2D projection of the vertices of the 3D bounding boxes in the image is common in monocular 3D object detection. These projected points carry information about the distance and pose of the actual 3D bounding boxes.
- **Road plane** [8, 23]. Most outdoor driving scenes are approximately flat roads. Thus, some work assumes that all objects lay on the same 2D plane and use it as a constraint to refine the 3D detections.
- **Relative pose** [6, 25]. Relative pose between objects is also explored in existing work as a constraint for the individual pose of objects. However, the formulation in existing approaches depends on the camera configurations, thus not viewpoint invariant.

2.3. Equivariance and invariance

Due to the absence of depth information in the monocular images, explorations of 3D equivariance and invariance is very limited in monocular 3D object detection. The majority of them focus on scale-invariance and equivariance [12, 21] of the 2D feature map and build an approximate inverse-proportional relation between the object depth and 2D appearance scale. [4] is closest to us as it investigates the viewpoint-equivariance by augmenting more camera viewpoints during training. In comparison, our target is viewpoint-invariant, which is simpler and less costly.

3. Methodology

3.1. Overview

We develop a monocular 3D object detection network with inter-object estimation capacity. Our network is built based on the state-of-the-art model Focal-PETR [22]. An overview is illustrated in Fig. 2. It is a transformer-based architecture, which uses queries to represent objects, thus making it straightforward to model the inter-object relations by taking the pair of involved queries as input.

3.2. Inter-object estimation module

Our inter-object estimation module is built upon query-based object representations. The basic idea is to input pairs of object-query features and yield inter-object estimations. The inter-object estimation module comprises four parts, object-pair feature fusion, inter-object regression, inter-level feature refinement, and differentiable pose-graph optimization. Due to the quadratic complexity in computing the pair-wise features, we filter the object queries based on the classification score and only pass the positive predictions into the inter-object estimation module.

Object-pair feature fusion A pair of objects is defined as a reference object i and a target object j, with their poses

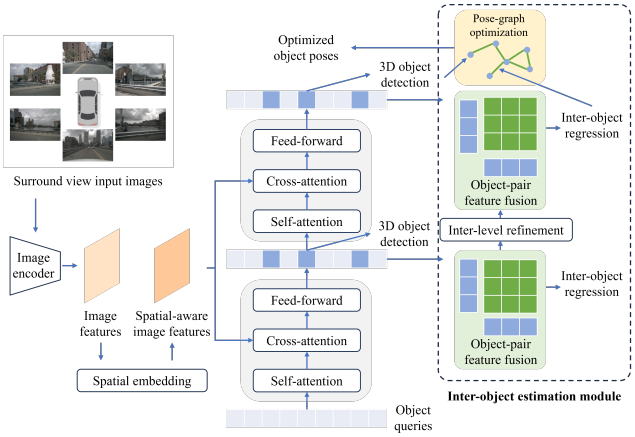

Figure 2. The overview of our network design. It follows the overall architecture of Focal-PETR [22] with our proposed inter-object estimation module on the right, highlighted in the dashed box. We only illustrate two levels of attention modules and the corresponding inter-object estimations for simplicity. The actual implementation of the model has six levels.

$P_i, P_j \in \mathrm{SE}(3)$. We denote the transformation,

$$P_{ij} = P_i^{-1} P_j, \tag{1}$$

as the relative pose for the object pair. The relative pose is irrelevant to the observer (ego vehicle) and thus is *viewpoint-invariant*.

For each pair of objects, we use a spatial alignment module as in [22] to obtain the inter-object features,

$$F_{ij} = \mathcal{T}_w(F_i) * F_j + \mathcal{T}_b(F_i), \tag{2}$$

where $F_i, , F_j \in \mathbb{R}^c$ are features of the object queries, $\mathcal{T}_w$ and $\mathcal{T}_b$ are MLP mappings: $\mathbb{R}^c \rightarrow \mathbb{R}^c$, and $*$ denotes the element-wise multiplication.

Inter-object regression Given the inter-object features, the inter-object regression is conducted using MLPs. Besides the relative pose, we define a *correlation mask* to highlight the object pairs that are closely correlated. Our observation is that the relative pose between objects is easy to perceive from images only when the objects are close to each other. We define the correlation mask as follows:

$$M_{ij} = M_{ij}^{class} * M_{ij}^{distance}, \tag{3}$$

where

$$M_{ij}^{class} = \begin{cases} 1 & \text{if } C_i = C_j, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

in which C_i is the classification label of object i.

Denote $P \in \mathrm{SE}(3)$ as $P = (x_P, y_P, z_P, \theta_P)$, where θ is the yaw angle (object pitch and roll are assumed to be zero), then we have

$$M_{ij}^{distance} = \begin{cases} 1 & \text{if } d_{P_{ij}} \leq d_{threshold}, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where $d_{P_{ij}} = \sqrt{x_{P_{ij}}^2 + y_{P_{ij}}^2 + z_{P_{ij}}^2}$, and $d_{threshold}$ is to filter out the object pairs that are far away.

We use an MLP to regress the correlation mask M_{ij}.

Inter-level feature refinement The above object-pair feature fusion and inter-object estimation are conducted after each level of the attention blocks. We can propagate the earlier layers' features to the later ones to enhance the representation. We use MLPs to process the object-pair features from the last level and add them to the object-pair features of the next level. We take the union of positive queries at different levels for inter-object estimation so that the inter-object features can be propagated along different levels.

Differentiable pose-graph optimization With relative pose estimation between objects, we can build a pose graph among the objects. Given a set of objects with estimated pose $\hat{P}_i, i \in \{1, ..., n\}$ and their predicted correlation mask $\hat{M}_{ij}$ and relative pose $\hat{P}_{ij}$, the optimization variables are the poses of the objects, and the constraints are the individually estimated poses and the relative poses. Only the subset of relative pose constraints with predicted correlation mask $\hat{M}_{ij} > 0.5$ are included as edges in the pose graph. The network regresses a weight vector for each of the individual object-pose estimation ($\hat{W}_i$) and the relative pose estimation ($\hat{W}_{ij}$) to weight the cost terms. The pose-graph optimization is made differentiable using the Theseus library [20], so that the regressed pose, relative pose, and weights are supervised by the optimized object poses $\tilde{P}_i$'s:

$$\{\tilde{P}_i\}_{i \in G} = \arg\min_{\{P_i\}} \sum_i \hat{W}_i r^2(P_i, \hat{P}_i) + \sum_{i,j} \hat{W}_{ij} r^2(P_i^{-1} P_j, \hat{P}_{ij}), \tag{6}$$

where G denotes the subset of object detections with associated edges. r^2 denotes the squared residual function:

$$r^2(P, \hat{P}) = [(x_P - x_{\hat{P}})^2, (y_P - y_{\hat{P}})^2, (z_P - z_{\hat{P}})^2, (round(\theta_P - \theta_{\hat{P}}))^2], \tag{7}$$

where $round$ is to round the angular error to $[-\pi, \pi)$.

Loss functions We follow the same loss functions used in Focal-PETR [22]. Furthermore, we use L1 loss to supervise $\hat{P}_{ij}$ and $\tilde{P}$. Yaw angle θ is parameterized with $(sin(\theta), cos(\theta))$ in the regression and loss function. $\hat{M}_{ij}$ is supervised using focal loss.

Table 1. Results on the nuScenes [3] validation set. ↑ means higher is better. ↓ means lower is better. The best is highlighted in **bold**.

Method	$NDS\uparrow$	$mAP\uparrow$	$mATE\downarrow$	$mASE\downarrow$	$mAOE\downarrow$	$mAVE\downarrow$	$mAAE\downarrow$	$mARTE\downarrow$	$mAROE\downarrow$	$mARPE\downarrow$
Focal-PETR [22]	0.363	0.330	0.758	0.280	0.685	1.155	0.303	0.437	0.245	0.577
Ours	**0.384**	**0.334**	**0.757**	**0.276**	**0.672**	**0.879**	**0.245**	**0.424**	**0.237**	**0.561**

Table 2. Ablation study on the inter-object estimation targets.

Mask	Distance	Relative pose	$NDS\uparrow$
×	×	×	0.363
✓	×	×	0.380
✓	✓	×	0.382
✓	×	✓	**0.384**

4. Experiments

4.1. Dataset and metrics

We validate our approach on the large-scale autonomous driving dataset, nuScenes [3]. Besides the official metrics of nuScenes to evaluate the 3D object detection, we further propose new metrics to **evaluate the accuracy of inter-object relative pose estimations**. Specifically, $mARTE$, $mAROE$, $mARPE$ (mean Average Relative Translation / Orientation / Pose Error) are proposed. First, we define the relative translation from object i to object j, $RT_{ij} = [x_j - x_i, y_j - y_i, z_j - z_i] \in \mathbb{R}^3$, and relative orientation, $RO_{ij} = round(\theta_j - \theta_i) \in [-\pi, \pi)$. Then, the relative translation / orientation errors RTE_{ij} and ROE_{ij} are defined as:

$$RTE_{ij} = \left\| RT_{ij}^{est} - RT_{ij}^{gt} \right\|_2, \tag{8}$$

$$ROE_{ij} = \left| round(RO_{ij}^{est} - RO_{ij}^{gt}) \right|, \tag{9}$$

where est and gt denote the estimation and ground truth, respectively. We also define the relative pose error to combine the translation and orientation errors together, $RPE_{ij} = \sqrt{RTE_{ij}^2 + ROE_{ij}^2}$. Finally, the mean average operation (e.g., $\{RTE_{ij}\}_{ij} \rightarrow mARTE$) is defined similarly as the true-positive metrics in nuScenes (e.g., $mATE$), which is to take the average of the cumulative mean at each recall level per class and finally take the mean overall the classes. Notice that the above inter-object metrics are symmetric, i.e., $RTE_{ij} = RTE_{ji}$, and they are defined in the ego-vehicle's frame. Thus, each pair of objects is counted only once in the metrics. The pair-wise confidence is defined as the confidence of the object with lower confidence in the pair, and the errors of all pairs with the same lower-confidence object are averaged before calculating the Precision-Recall curve.

4.2. Implementation details

We use the ResNet-50 [9] backbone with input image size 704×256. All networks are trained with batch size 16 for 24 epochs across 8 NVIDIA A40 GPUs. $d_{threshold} = 20\text{m}$.

4.3. Experimental results

We present the quantitative result in Tab. 1. The numbers are evaluated on the nuScenes validation set. Our approach achieves superior performance in all reported metrics. The improvement in the mAP metric shows that the 3D localization of the detection objects is improved. Moreover, an interesting outcome is that the $mAVE$ and $mAAE$ drop considerably with our method. While the relative pose estimation seems not to be directly related to the velocity estimation or the attribute (e.g., moving vs. parked) estimation, the improvement is not a coincidence. Notice that our model predicts a correlation mask to highlight the object pairs of the same class within a short distance. They are likely to have *similar velocities and attributes*, for example, all platooning on a highway or parked on the sideway. It implies that the inter-object estimation module enhances the similarity of the features of object pairs with high correlation, encouraging more consistent velocity and attribute predictions between the pairs. It shows that the inter-object estimation improves the understanding of objects' behavior.

The evaluation of the inter-object relative pose estimation is presented in the last three columns of Tab. 1. Our proposed method achieves lower error, indicating that the estimated object poses from our method are more consistent regarding their relative poses.

Ablation study We further show the effect of different inter-object estimation targets on the overall performance in Tab. 2. It shows that mask prediction is very important in inter-object estimation. Adding this single sub-task considerably improves the overall performance. The distance and relative pose estimation can further improve performance.

5. Conclusion

We introduce inter-object estimation to monocular 3D object detection. By regressing the correlation and relative pose in object pairs and optimizing the obtained pose graph, we improve over the competitive baseline model, Focal-PETR, in 3D object detection, relative pose estimation, and surprisingly, the behavior understanding of the objects. The ablation study validates the effect of each inter-object estimation target in the overall performance improvement. We believe that our perception solution can better prepare the self-driving algorithms to adapt to the behavior of the surrounding road users and achieve improved roadmanship.

References

[1] Ivan Barabanau, Alexey Artemov, Evgeny Burnaev, and Vyacheslav Murashkin. Monocular 3d object detection via geometric reasoning on keypoints. *arXiv preprint arXiv:1905.05618*, 2019. 2

[2] Deniz Beker, Hiroharu Kato, Mihai Adrian Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Monocular differentiable rendering for self-supervised 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 514–529. Springer, 2020. 2

[3] 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. 4

[4] Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9213–9222, 2023. 2

[5] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2

[6] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 2

[7] Laura Fraade-Blanar, Marjory S Blumenthal, James M Anderson, and Nidhi Kalra. *Measuring automated vehicle safety: Forging a framework*. 2018. 1

[8] Jiaqi Gu, Bojian Wu, Lubin Fan, Jianqiang Huang, Shen Cao, Zhiyu Xiang, and Xian-Sheng Hua. Homography loss for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1080–1089, 2022. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[11] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11867–11876, 2019. 2

[12] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683. Springer, 2022. 2

[13] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3559–3568, 2018. 2

[14] Peixuan Li and Huaici Zhao. Monocular 3d detection with geometric constraint embedding and semi-supervised training. *IEEE Robotics and Automation Letters*, 6(3):5565–5572, 2021. 2

[15] Peixuan Li, Huaici Zhao, and Pengfei Liu. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. 2

[16] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2

[17] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 2

[18] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[19] Huei Peng, Minghan Zhu, and John K Lenneman. Automated vehicles and roadmanship, how safe robot cars can read the road like humans do. *Mcity White Papers*, 2022. 1

[20] Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky TQ Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, et al. Theseus: A library for differentiable nonlinear optimization. *Advances in Neural Information Processing Systems*, 35:3801–3818, 2022. 3

[21] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Elisa Ricci, and Peter Kontschieder. Towards generalization across depth for monocular 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 767–782. Springer, 2020. 2

[22] Shihao Wang, Xiaohui Jiang, and Ying Li. Focal-petr: Embracing foreground for efficient multi-camera 3d object detection. *arXiv preprint arXiv:2212.05505*, 2022. 2, 3, 4

[23] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2

[24] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[25] Jiwei Xiao, Ruiping Wang, and Xilin Chen. Holistic pose graph: modeling geometric structure among objects in a

scene using graph inference for 3d object prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12717–12726, 2021. 2

[26] Songan Zhang, Lu Wen, Huei Peng, and H Eric Tseng. Quick learner automated vehicle adapting its roadmanship to varying traffic cultures with meta reinforcement learning. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1745–1752. IEEE, 2021. 1

[27] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 2

[28] Minghan Zhu, Lingting Ge, Panqu Wang, and Huei Peng. Monoedge: Monocular 3d object detection using local perspectives. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 643–652, 2023. 2

[29] Minghan Zhu, Shizong Han, Hong Cai, Shubhankar Borse, Maani Ghaffari Jadidi, and Fatih Porikli. 4d panoptic segmentation as invariant and equivariant field prediction. *arXiv preprint arXiv:2303.15651*, 2023. 2