Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving

Yi-Nan Chen[1] Hang Dai[2*] Yong Ding[1*]
[1]School of Micro-Nano Electronics, Zhejiang University
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[*]Corresponding authors{hang.dai@mbzuai.ac.ae, dingy@vlsi.zju.edu.cn}.

Abstract

Pseudo-LiDAR 3D detectors have made remarkable progress in monocular 3D detection by enhancing the capability of perceiving depth with depth estimation networks, and using LiDAR-based 3D detection architectures. The advanced stereo 3D detectors can also accurately localize 3D objects. The gap in image-to-image generation for stereo views is much smaller than that in image-to-LiDAR generation. Motivated by this, we propose a Pseudo-Stereo 3D detection framework with three novel virtual view generation methods, including image-level generation, feature-level generation, and feature-clone, for detecting 3D objects from a single image. Our analysis of depth-aware learning shows that the depth loss is effective in only feature-level virtual view generation and the estimated depth map is effective in both image-level and feature-level in our framework. We propose a disparity-wise dynamic convolution with dynamic kernels sampled from the disparity feature map to filter the features adaptively from a single image for generating virtual image features, which eases the feature degradation caused by the depth estimation errors. Till submission (November 18, 2021), our Pseudo-Stereo 3D detection framework ranks 1st on car, pedestrian, and cyclist among the monocular 3D detectors with publications on the KITTI-3D benchmark. The code is released at https://github.com/revisitq/Pseudo-Stereo-3D.

1. Introduction

Detecting the 3D objects from monocular image enables the machine to perceive and understand the 3D real world, which has a wide applications including virtual reality, robotics and autonomous driving. Monocular 3D detection is a challenging task because of the lack of accurate 3D information in a single image. However, the huge potential in such a cheap and easy-to-deploy solution to 3D detection attracts more and more researchers. Remarkable progress has been made in Pseudo-LiDAR detectors [11, 29, 34, 43, 52] that use a pre-trained depth estimation network to generate Pseudo-LiDAR representations, *e.g.* pseudo point clouds and pseudo voxels, and then feed them to LiDAR-based 3D detectors. It shows that enhancing the capability of perceiving depth can improve monocular 3D detection performance. However, there is a huge performance gap between Pseudo-LiDAR and LiDAR-based detectors because of the errors in image-to-LiDAR generation [32].

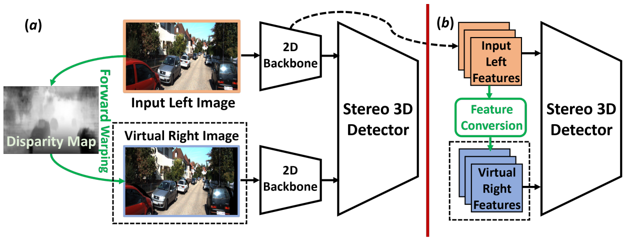

Figure 1. Overview of our Pseudo-Stereo 3D detection framework with novel virtual view generation methods: (a) Image-level to use the generated disparity map for forward warping the input left image into a virtual right image, (b) Feature-level to convert the left features into virtual right features. A feature conversion baseline is to clone the left features as the special case in stereo views.

Apart from LiDAR-based detectors, the stereo 3D detectors [9, 17] can also accurately localize 3D objects. Also, the gap in image-to-image generation for stereo views is much smaller than that in image-to-LiDAR generation, which is a cross-modality conversion. Instead of Pseudo-LiDAR, we propose a novel Pseudo-Stereo 3D detection framework for monocular 3D detection. Our Pseudo-Stereo 3D detection framework generates a virtual view from a single input image to compose Pseudo-Stereo views with the generated virtual view and the input view. Then, we feed the Pseudo-Stereo views to stereo 3D detectors for detecting 3D objects from the single input image. We use one of the most advanced stereo 3D detectors, LIGA-Stereo [17], as the base detection architecture. Thus, the virtual view generation is the key to our Pseudo-Stereo 3D detection framework.

We take KITTI-3D as an example only to explain how to generate a virtual view. Note that the virtual view does not require the ground-truth actual view in the dataset for training. In KITTI-3D, the monocular 3D detection is performed on the left image from the stereo views. Our aim is to construct Pseudo-Stereo views by generating the virtual right view from the input left view in either image-level or feature-level for monocular 3D detection. As shown in Figure 1, we propose two types of virtual view generation: (a) image-level to generate the virtual right image from the input left image and (b) feature-level to convert the left features into virtual right features. In image-level, we convert the estimated depth map from the input left image into disparities and use them to forward warp the input left image into a virtual right image to compose the Pseudo-Stereo views with the input left view. In feature-level, we propose a disparity-wise dynamic convolution with dynamic kernels sampled from disparity feature map to filter the left features adaptively for generating virtual right features, which eases the feature degradation caused by the depth estimation errors. Also, a simple feature conversion is to clone the left features as the virtual right features, which is the special case of stereo views that the virtual right view is the same as the left view. We summarize our **contributions**:

- We propose a Pseudo-Stereo 3D detection framework with three novel virtual view generation methods, including image-level generation, feature-level generation and feature-clone, for detecting 3D objects from a single image, achieving significant improvements in monocular 3D detection. The proposed framework with feature-level virtual view generation ranks 1^{st} among the monocular 3D detectors with publications across three object classes on KITTI-3D benchmark.
- In our framework, we analyze two major effects of learning depth-aware feature representations, including the estimated depth map and the depth loss as the depth guidance. It is very interesting to find that the depth loss is effective in feature-level virtual view generation only and the estimated depth map is effective in both image-level and feature-level for depth-aware feature learning.
- In our feature-level virtual view generation method, we propose a disparity-wise dynamic convolution with dynamic kernels from disparity feature map to adaptively filter the features from a single image for generating virtual image features, which avoids the feature degradation caused by the depth estimation errors.

2. Related Works

The architectures for monocular 3D object detection can be mainly categorized into two groups: Pseudo-LiDAR based methods [11, 34, 43] that use pre-trained depth networks to generate pseudo LiDAR representations, *e.g.* pseudo point clouds and pseudo voxels, and then feed them to LiDAR-based 3D detectors, and the rest monocular 3D detection methods that use 2D feature learning from a single image with optional 3D cues matching, concatenating or guiding for 3D perception [24, 26, 31, 38, 39, 54, 57].

Monocular 3D Detection. There are a few monocular 3D detectors use 2D feature learning in 2D backbone with optional 3D cues concatenated or matched to 2D features for 3D perception. Chabot *et al*. [6] estimate the similarity between the detected vehicle and a pre-defined 3D vehicle shape template used as 3D cues in 2D backbone. They solve the 3D location and 3D rotation angle of the detected vehicle in a standard 2D/3D matching algorithm [22]. Barabanau *et al*. [2] use 2D features to predict the rotation and key points of a car in a 2D backbone. Then, they use a geometric reasoning between the key points and the corresponding points in CAD models to get the depth and 3D locations of the car. But it is difficult to get CAD models of all object classes. GrooMeD-NMS [21] extracts 2D features for monocular 3D detection, with a differentiable NMS selecting the best 3D box candidate. GS3D [23] uses a specifically designed 2D backbone to extract the surface features for tackling the representation ambiguity between 2D bounding box and 3D bounding box. MonoEF [56] employs a 2D backbone with a camera extrinsic parameter aware module to deconstruct camera extrinsic parameters from 3D detection parameters. M3D-RPN [4] extracts 2D image feature to predict both 2D and 3D bounding boxes directly by minimizing the distance error between the 2D projection of the predicted 3D bounding box and the predicted 2D bounding box. Following M3D-RPN [4], many works [28, 33, 35] enhance the 2D feature learning with the 2D-3D detection head for monocular 3D detection.

Some methods aggregate the 2D image feature and the depth features extracted from the depth map to get 2D depth-aware features [12, 33]. D4LCN [12] employs a depth-guided convolution with the weights and the receptive fields learning from the estimated depth for depth-aware feature extraction. DDMP-3D [47] uses a depth-conditioned propagation based on graph to learn 2D depth-aware features for monocular 3D detection. DD3D [33] adds a depth prediction head to the 3D detection head and uses a depth loss to learn 2D features that are sensitive to depth information for monocular 3D detection. DD3D [33] also pre-trains the depth prediction head on a large-scale dataset and fine-tunes the overall network on monocular 3D detection task. Other methods extract 2D features and construct 3D feature volume from the transformation of 2D features to improve 3D perceiving capacity. CaDDN [38] uses the estimated depth distributions to construct a frustum feature grid. Then, the frustum feature is converted into a voxel grid using known camera calibration parameters to construct 3D voxel feature volumes. ImVoxelNet [39] uses 2D backbone to extract 2D image feature and projects the 2D features into 3D feature volumes following [31]. Then, the 3D feature volumes go through 3D backbone to enhance the 3D features for monocular 3D detection.

Pseudo-LiDAR. Pseudo-LiDAR architecture converts the estimated depth map from a single image into Pseudo 3D data representations [5, 46] which are then fed to 3D backbone to learn point-wise, voxel-wise or bird's eye view (BEV) features for monocular 3D detection. RefinedMPL [46] uses PointRCNN [41] for point-wise feature learning in a supervised or an unsupervised scheme from pseudo point clouds prior. AM3D [30] uses a PointNet [36] backbone for point-wise feature extraction from pseudo point clouds, and employs a multi-modal fusion block to enhance the point-wise feature learning. MonoFENet [1] enhances the 3D features from the estimated disparity for monocular 3D detection. Decoupled-3D [5] recovers the missing depth of the object using the coarse depth from 3D object height prior with the BEV features that are converted from the estimated depth map. However, the performance and the generalization capability of these methods rely on the accuracy of image-to-LiDAR generation, which has a huge gap between the two data modalities.

3. Preliminaries of Stereo 3D Detector

Volume-based stereo 3D detectors aim to generate 3D anchor space from stereo image [37] and localize 3D objects from 3D feature volume [9, 17, 48]. DSGN [9] follows the widely used 3D cost volume construction in stereo matching [16, 45, 50] with 3D geometric information encoding. The depth loss in stereo matching branch helps learn depth-aware features for the detection branch, improving the detection accuracy. Wang *et al*. [48] use a direct construction of 3D cost volume to reduce the computational cost. Based on DSGN [9], LIGA-Stereo [17] achieves significant improvements against other methods [3, 9, 53] in stereo 3D detection. Thus, we use LIGA-Stereo [17] as our base stereo 3D detection architecture and feed the Pseudo-Stereo views to LIGA-Stereo. We focus on how to generate the virtual right view from the input left view and learn Pseudo-Stereo features that are sensitive to depth information. Thus, we introduce the stereo image feature extraction and the 3D feature volume construction in LIGA-Stereo [17].

Stereo Image Feature Extraction. Given an image pair (I_L, I_R) from stereo views, the LIGA-Stereo [17] first extracts the left features and the right features via a ResNet-34 [18] with shared weights as the 2D image backbone. The strides of the output feature map in the five blocks are 2, 2, 4, 4 and 4, respectively. The channels of the output feature map in the five blocks are 64, 64, 128, 128 and 128. Then, we denote the left and the right features as F'_L and F'_R that are the input to the spatial pyramid pooling (SPP) module [7] with shared weights for getting the final left features F_L and right features F_R. The strides of the final left features F_L and right features F_R are 1, and the channels of the final left features F_L and the right features F_R are 32.

The 3D Feature Volume Construction. With the left features F_L and the right features F_R, the stereo volume V_{st} is built by concatenating the left features F_L with the re-projected right features $F_{R->L}$ at every candidate depth level. Thus, the stereo volume construction can be formulated with camera parameters as:

$$V_{st}(u,v,w) = concat[F_L(u,v), F_{R->L}(u,v)] \quad (1)$$

$$F_{R->L}(u,v) = F_R(u - \tfrac{f \cdot b}{d(w) \cdot S}, v) \quad (2)$$

$$d(w) = w \cdot v_d + z_{min} \quad (3)$$

where (u, v) are the pixel coordinates, $w \in [0, 1, ...]$ indicates the depth index, S is the stride of the feature map, v_d is the depth interval, z_{min} indicates the minimal depth value, f is the camera focal length, and b represents the baseline of the stereo camera pair. After the stereo volume V_{st} is filtered by a stereo network 3D Hourglass [17], we get a re-sampled stereo volume V'_{st} and a depth distribution volume P_{st}. The P_{st} describes the depth probability distribution of pixels for all the candidate depth levels described in $d(w)$. A **depth loss** is computed between the depth map regressed from the re-sampled stereo volume V'_{st} and the ground-truth depth map to guide the depth-aware learning of V'_{st}. With the camera calibration parameters, we can transform the volume in stereo space V'_{st} to the volume in 3D space V_{3d} by concatenating the semantic features from left image penalized by the depth probability P_{st} and the re-sampled stereo volume V'_{st}. Following SECOND [51], the 3D feature volume V_{3d} is collapsed to a bird's eye view (BEV) feature map F_{BEV} by merging the dimension of height and channel. Finally, a 2D aggregation network is used to get a refined BEV feature map F'_{BEV} that is used for the regression of 3D detection parameters.

4. Method

As shown in Figure 2, we propose three novel methods to generate the virtual right view from the input left view and construct the Pseudo-Stereo views in (a) image-level in Section 4.2, (b) feature-level in Section 4.3, and (c) feature-clone as the baseline of feature-level generation in Section 4.4. We describe the loss function in Section 4.5. In Section 4.6, we analyze the depth-aware feature learning in the proposed Pseudo-Stereo 3D detection framework.

4.1. Pseudo-Stereo 3D Detection Framework

We use LIGA-Stereo [17] as our base stereo 3D detection architecture and replace the stereo image feature extraction block with our Pseudo-Stereo image feature extraction block. As shown in Figure 2, we propose three virtual right view generation methods and extract Pseudo-Stereo image features from the input left view and the generated virtual right view. Then, we feed the Pseudo-Stereo image features to LIGA-Stereo for detecting 3D objects from the input left image only.

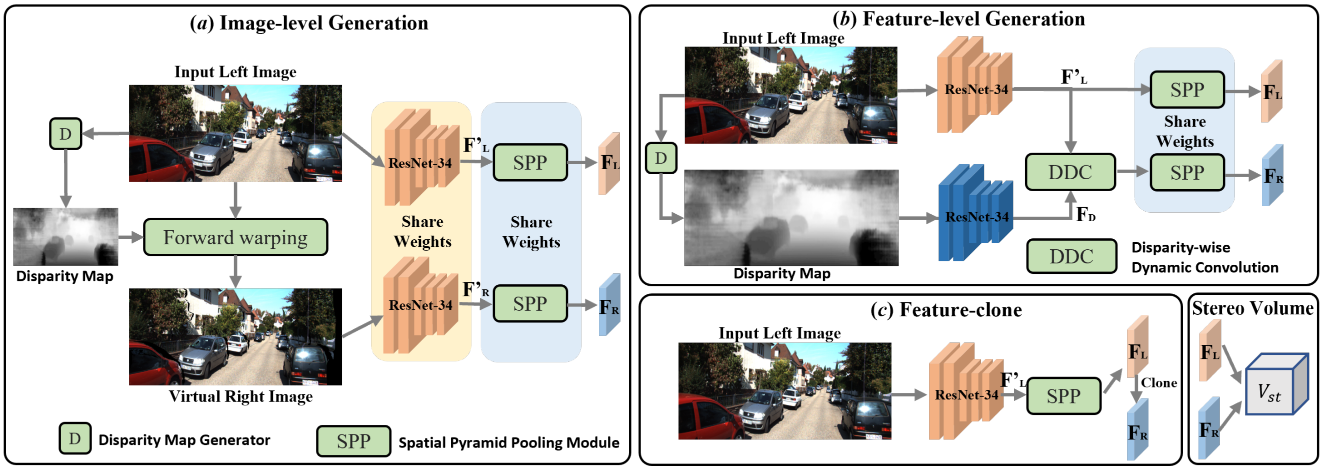

Figure 2. Overview of our virtual view generation methods: (a) Image-level that uses the generated disparity map for forward warping the input left image into a virtual right image, (b) Feature-level that converts the left features into virtual right features via the proposed disparity-wise dynamic convolution (DDC), (c) Feature-clone that simply duplicates the left features as the virtual right features.

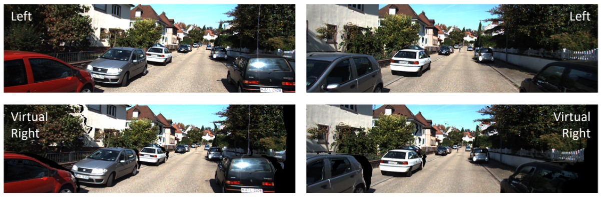

Figure 3. Left image (top) and the generated virtual right image (bottom) using our image-level virtual view generation method.

4.2. Image-level Generation

In image-level, we generate a virtual right image $\hat{I}_R$ from the input left image I_L using the estimated disparity map as shown in Figure 2(a). Then, we extract Pseudo-Stereo image features from the Pseudo-Stereo pair $(I_L, \hat{I}_R)$. With a pair of pixel correspondences x_l and x_r in the left image I_L and the right image I_R, the disparity d between the pair of pixel correspondences can be computed as:

$$d = x_l - x_r \tag{4}$$

Given the depth value z for the pixel x_l in the input left image and the camera calibration parameters, the relationship between the disparity d and its corresponding depth value z can be formulated as:

$$d = \frac{f \cdot b}{z} \tag{5}$$

where f and b are the camera focal length and the baseline of the stereo camera pair, respectively.

To get the virtual right image, we first use a pre-trained DORN [13] to estimate the depth map Z from the input left image I_L. Then, we convert the depth map Z to a disparity map D according to Eqn. 5 with camera parameters. Following Eqn. 4, we use the disparity map to forward warp the left image I_L [40] into the virtual right image $\hat{I}_R$ as shown in Figure 3. To address 'blurry' edges, occlusions and collisions, we sharpen the disparity map by identifying the flying pixels and applying a Sobel edge filter response of greater than 3 to the disparity map on those flying pixels [19, 49]. In image-level generation, we embed the estimated depth information extracted from the left image into the virtual right image for Pseudo-Stereo 3D detection. Then, we use a ResNet-34 [18] with shared weights followed by a spatial pyramid pooling (SPP) module with shared-weights to extract the left features F_L and the virtual right features $\hat{F}_R$ from the Pseudo-Stereo pair $(I_L, \hat{I}_R)$. We can use the Pseudo-Stereo image features to construct the stereo volume V_{st} as detailed in Section 3.

4.3. Feature-level Generation

Generating the virtual right image is a time-consuming process because of the forward warping [19, 49]. To overcome this, we propose a differentiable feature-level method for generating the virtual right features from the left features and the disparity features as shown in Figure 2(b). We convert the estimated depth map into a disparity map and use two ResNet-34 [18] to extract the left features F'_L from the left input image and the disparity features F_D from the disparity map. The two ResNet-34 are not with shared weights.

Instead of computing the offsets to compensate the left view as the virtual right view, we propose a disparity-wise dynamic convolution (DDC) to filter the left feature map $F'_L \in \mathbb{R}^{W \times H \times C}$ adaptively by the dynamic kernels from the disparity feature map $F_D \in \mathbb{R}^{W \times H \times C}$ for generating the virtual right feature map $\hat{F'}_R \in \mathbb{R}^{W \times H \times C}$, where W, H and C are the width, height and channel of the feature map, respectively. As shown in Figure 4, the adaptive filtering process use a 3×3 sliding window to cover all the

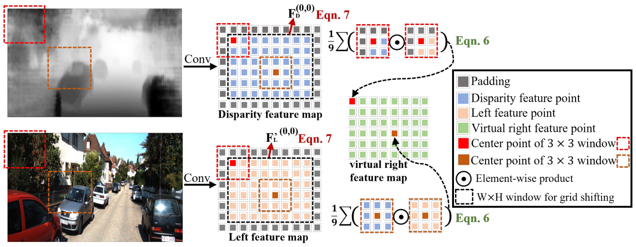
Conv
Disparity feature map
$F_D^{(0,0)}$ Eqn. 7
Conv
Left feature map
$F_L'^{(0,0)}$ Eqn. 7
$\frac{1}{9}\sum(\ \odot\ )$ Eqn. 6
virtual right feature map
$\frac{1}{9}\sum(\ \odot\ )$ Eqn. 6
Padding
Disparity feature point
Left feature point
Virtual right feature point
Center point of 3 × 3 window
Center point of 3 × 3 window
Element-wise product
W×H window for grid shifting

Figure 4. An illustration of disparity-wise dynamic convolution.

feature points in F_D and F'_L:

$$\hat{F}_R(i,j) = \frac{1}{3{\times}3}\sum_{u'}\sum_{v'} F_D(u',v') \cdot F'_L(u',v') \tag{6}$$

where $u' \in \{i-1, i, i+1\}$ and $v' \in \{j-1, j, j+1\}$ are the coordinates of the feature points in the sliding window. We need to apply $W * H$ times sliding window to cover the whole feature map, which is not efficient. Instead, we use a grid shifting operation that can cover the whole feature map with 9 times shifting. After padding, we shift a $W \times H$ window following the direction and the step size represented in a 3×3 grid $\{(g_i, g_j)\}$, $g \in \{-1, 0, 1\}$ on F_D and F'_L for getting the kernel $F_D^{(g_i,g_j)} \in \mathbb{R}^{W\times H\times C}$ and the feature map $F_L'^{(g_i,g_j)} \in \mathbb{R}^{W\times H\times C}$, respectively. When $g_i = 0$ and $g_j = 0$, the $W \times H$ window covers the original feature map that is without padding as shown in black dot box of Figure 4. Thus, we can get the virtual right features $\hat{F}'_R$ by filtering $F_L'^{(g_i,g_j)}$ adaptively by the kernels $F_D^{(g_i,g_j)}$:

$$\hat{F}'_R = \frac{1}{3{\times}3}\sum_{g_i,g_j} F_L'^{(g_i,g_j)} \odot F_D^{(g_i,g_j)} \tag{7}$$

where the grid shifting operation is applied nine times to cover the whole feature map. For more details, please refer to supplementary materials. We feed F'_L and $\hat{F'}_R$ to SPP module with shared weights using strides of 4 for getting the final left features F_L and virtual right features $\hat{F}_R$.

Compared with the image-level generation, the feature-level generation is faster without forward warping and more adaptive using the proposed disparity-wise dynamic convolution. Also, by embedding the estimated depth information into high dimensional feature space and using the embedded depth information to filter the left features, it mitigates the degradation of depth-aware representations and the depth-aware representations can strengthen the embedded depth information with extra depth guidance, achieving significant improvements in monocular 3D detection.

4.4. Image Feature Clone

We clone the left features as the virtual right features as shown in Figure 2(c). This can be seen as the special case of Pseudo-Stereo views that the virtual right view is the same as the left view. Also, feature cloning is used as the baseline of feature-level generation. With different Pseudo-Stereo views, the proposed framework can improve the representations of 3D feature volume with the paired pixel-correspondence constraints or the feature-correspondence constraints converted from the estimated depth map. However, cloning feature does not need a pre-trained depth estimation network in our Pseudo-Stereo 3D detection framework, leading to better generalization capability.

4.5. Loss Function

Since we use LIGA-Stereo [17] as our base stereo 3D detection architecture and replace the original stereo image feature extraction block with the proposed three variants of Pseudo-Stereo image feature generation block as shown in Figure 2, we employ the same loss function as LIGA-Stereo [17], including the detection loss L_d for the regression of all detection parameters, the depth loss L_{depth} as the additional depth guidance for the re-sampled stereo volume V'_{st} and the knowledge distillation loss L_{kd} to transfer the structural knowledge from a LiDAR-based detector as described in LIGA-Stereo [17]. The overall loss can be formulated as:

$$L = \lambda_d L_d + \lambda_{dep} L_{depth} + \lambda_{kd} L_{kd} \tag{8}$$

where λ_d, λ_{dep}, and λ_{kd} are the regularization weights for the detection loss, the depth loss, and the knowledge distillation loss, respectively. The knowledge distillation adopted in LIGA-Stereo is well studied in [17]. Please refer to LIGA-Stereo [17] for more details. We focus on how to generate the virtual right view from the input left view and improve the capability of perceiving depth in features.

4.6. Learning Depth-aware Features

The depth-aware feature learning in our framework lies in two aspects: the estimated depth map and the depth loss. we convert the estimated depth map as the disparity map and use it in either image space or feature space. By comparing the performance of the two methods, we can study the effect of the estimated depth map used for depth-aware feature learning in both image-level and feature-level. The depth loss L_{depth} is used as the additional depth guidance for the re-sampled stereo volume V'_{st} to improve the depth awareness in features, improving monocular 3D detection performance. Although both the estimated depth map and the depth loss can improve the depth awareness in feature learning, the interaction between the two factors is not well studied for monocular 3D detection before this work.

For the image-level generation in Section 4.2, we use a Pseudo-Stereo image pair to extract the Stereo image features, where the virtual right image is generated from the left image with the help of the estimated depth map. The monocular depth estimation is an ill-posed problem, which makes it difficult to get high-quality depth maps for virtual right image generation. Thus, the pixel-correspondence constraints of the generated Pseudo-Stereo pairs may have large offsets against the ground truth because of the depth estimation errors. Learning with the virtual right image warped from the inaccurate pixel-correspondences causes feature degradation. Since there is a huge gap between the ground-truth depth and the degraded feature, forcing the network to fit the ground-truth depth map using the depth loss with the degraded feature impairs the overall performance. For feature-level generation in Section 4.3, the virtual right features are generated from the left features and the disparity features. Unlike image-level generation, where the forward-warping is a non-learning process from image to image, the feature-level generation is an adaptive learning process with disparity-wise dynamic convolutions from feature to feature. Also, the estimated depth information is embedded into high dimensional feature space and the embedded depth information is used to filter the left features in the feature-level generation. This eases the degradation of depth-aware representations, mitigating the gap between the ground-truth depth and the feature. Thus, the feature representations can evolve and refine the depth information with extra depth guidance, for example, a depth loss. For feature-clone in Section 4.4, we duplicate the left features as the Pseudo-Stereo image features without the estimated depth map. The depth loss alone can improve the depth awareness of features, improving the detection performance.

5. Experiments

5.1. Dataset and Evaluation Metric

Dataset. KITTI 3D object detection benchmark [15] is the most widely used benchmark for 3D object detection. It comprises 7481 training images and 7518 test images, along with the corresponding point clouds captured around a mid-size city from rural areas and highways. KITTI-3D provides 3D bounding box annotations for 3 classes, *Car*, *Cyclist* and *Pedestrian*. Commonly, the training set is divided into training split with 3712 samples and validation split with 3769 samples following that in [12], which we denote as KITTI $train$ and KITTI val, respectively. All models in ablation studies are trained on the KITTI $train$ and evaluated on KITTI val. For the submission of our methods, the models is trained on the 7481 training samples. Each object sample is assigned to a difficulty level, Easy, Moderate or Hard according to the object's bounding box height, occlusion level and truncation.

Evaluation Metric. We use two evaluation metrics in KITTI-3D, $i.e.$, the IoU of 3D bounding boxes or BEV 2D bounding boxes with average precision (AP) metric, which are denoted as AP_{3D} and AP_{BEV}, respectively. Following the monocular 3D detection methods [2,12,54], we conduct the ablation study on *Car*. KITTI-3D uses the $AP|_{R40}$ with 40 recall points instead of $AP|_{R11}$ with 11 recall points from October 8, 2019. We report all the results in $AP|_{R40}$.

5.2. Experiment Settings

Input Setting. We use the pre-trained model of DORN [14] to estimate the depth map. Then, we transform the depth maps into disparity maps with the camera calibration parameters. The virtual right images in image-level generation are generated before training to reduce the training time. For feature-level generation, the disparity map is normalized by μ= 33.20, σ=15.91. The μ and σ indicate the mean and variance of disparity map calculated from the training set.

Training Details. The network is trained with an AdamW [25] optimizer, with β_1=0.9, β_2=0.999. We train the network with 4 NVIDIA RTX 3090 GPUs. The batch size is set to 4. For the regularization weights of the training loss, λ_d=1.0, λ_{kd}=1.0. The regularization weight λ_{dep} for depth loss L_{depth} is set to 0 or 1, representing whether the depth loss is used or not. We use a single model to detect objects in different classes (*Car*, *Cyclist* and *Pedestrian*) together. Other hyper-parameters are set as the same as LIGA-Stereo [17].

Exp.	Methods	L_{depth}		AP_{3D}/AP_{BEV}	
			Easy	Moderate	Hard
1	Image-level	✓	31.43 / 41.82	21.53 / 29.00	18.47 / 25.21
2	Image-level		**31.81 / 42.87**	**22.36 / 30.16**	**19.33 / 26.38**
3	Feature-level	✓	**35.18 / 45.50**	**24.15 / 32.03**	**20.35 / 27.57**
4	Feature-level		22.04 / 31.10	16.18 / 22.55	14.31 / 20.56
5	Feature-clone	✓	**28.46 / 37.66**	**19.15 / 25.78**	**16.56 / 22.47**
6	Feature-clone		24.33 / 32.99	17.09 / 23.77	14.61 / 20.81

Table 1. Ablation studies of three proposed Pseudo-Stereo variants and L_{depth} at IOU threshold 0.7. Exp. is the experiment tag.

5.3. Ablation Study

As shown in Table. 1, we conduct ablation studies on the KITTI *val* for the three proposed Pseudo-Stereo variants: image-level, feature-level and feature-clone generation.

Image-level. As shown in Exp.1 and Exp.2 in Table. 1, with the depth loss, the overall performance of the image-level generation method decreases by (-0.38%, -0.83%, -0.86%) for AP_{3D} and (-1.05%, -1.16%, -1.17%) for AP_{BEV}. The pixel-correspondence constraints of the generated Pseudo-Stereo pairs may have large offsets against the ground truth because of the depth estimation errors. Learning with the virtual right image warped from the inaccurate pixel-correspondences causes the feature degradation. Forcing the degraded feature to fit the ground-truth depth map with the depth loss impairs the overall performance.

Feature-level. As can be seen from Exp.3 and Exp.4 in Table. 1, the feature-level generation with the depth loss achieves significant improvements on AP_{3D} (**+13.04%**, **+7.97%**, **+6.04%**) and AP_{BEV} (**+14.4%**, **+9.48%**, **+7.01%**). The forward-warping used in image-level generation is a non-learning process from image to image, while

Methods	Reference	AP_{3D} Easy	AP_{3D} Moderate	AP_{3D} Hard	AP_{BEV} Easy	AP_{BEV} Moderate	AP_{BEV} Hard
MonoDIS [44]	ICCV 2019	10.37	7.94	6.40	17.23	13.19	11.12
AM3D [30]	ICCV 2019	16.50	10.74	9.52	25.03	17.32	14.91
M3D-RPN [4]	ICCV 2019	14.76	9.71	7.42	21.02	13.67	10.23
D4LCN [12]	CVPR 2020	16.65	11.72	9.51	22.51	16.02	12.55
MonoPair [10]	CVPR 2020	13.04	9.99	8.65	19.28	14.83	12.89
MonoFlex [55]	CVPR 2021	19.94	13.89	12.07	28.23	19.75	16.89
MonoEF [56]	CVPR 2021	21.29	13.87	11.71	29.03	19.70	17.26
GrooMeD-NMS [21]	CVPR 2021	18.10	12.32	9.65	26.19	18.27	14.05
CaDDN [38]	CVPR 2021	19.17	13.41	11.46	27.94	18.91	17.19
DDMP-3D [47]	CVPR 2021	19.71	12.78	9.80	28.08	17.89	13.44
MonoRUn [8]	CVPR 2021	19.65	12.30	10.58	27.94	17.34	15.24
DFR-Net [58]	ICCV 2021	19.40	13.63	10.35	28.17	19.17	14.84
MonoRCNN [42]	ICCV 2021	18.36	12.65	10.03	25.48	18.11	14.10
DD3D [33]	ICCV 2021	23.22	16.34	14.20	30.98	22.56	20.03
Ours-im	–	19.79	13.81	12.31	28.37	20.01	17.39
Ours-fld	–	**23.74**	**17.74**	<u>15.14</u>	**32.84**	**23.67**	**20.64**
Ours-fcd	–	<u>23.61</u>	<u>17.03</u>	**15.16**	<u>31.83</u>	<u>23.39</u>	<u>20.57</u>

Table 2. Performance for *Car* of three methods on KITTI *test* at IOU threshold 0.7. The best results are **bold**, the second best <u>underlined</u>.

the feature-level generation is an adaptive and differentiable learning process with disparity-wise dynamic convolutions from feature to feature. The virtual right features are generated from the left features and the disparity features. Thus, the feature degradation caused by the depth estimation errors is mitigated by embedding the estimated depth information into high dimensional feature space and using the embedded depth representations to filter the left features dynamically in the feature-level generation. The gap between the ground-truth depth and the feature is mitigated. With the extra depth guidance from the depth loss, the depth representations can strengthen the embedded 3D measurements in feature-level, achieving much better performance.

Feature-clone. From Exp.5 and Exp.6 in Table. 1, it shows that the feature-clone achieves significant improvements with the depth loss on AP_{3D} (+4.13%, +2.06%, +1.95%) and AP_{BEV} (+4.67%, +2.01%, +1.66%). This lies in the fact that feature-clone does not require depth estimation network and the depth loss alone can improve the awareness of depth information in features.

Estimated Depth Map. From the comparison among Exp.2, Exp.4 and Exp.6 in Table. 1, with the estimated depth map used in both image-level (Exp.2) and feature-level (Exp.4), the models achieve better performance than the model without using the estimated depth map (Exp.6), which implies that the estimated depth map is effective in both image-level and feature-level in our framework. In image-level, the degraded feature caused by inaccurate pixel-correspondences is not forced to fit the ground-truth depth without the depth loss, and the estimated depth map improves the capability of perceiving depth information in the image input level, improving performance for monocular 3D detection. In feature-level, the feature degradation is eased by the proposed DDC in high dimensional feature space and the estimated depth map improves the capability of perceiving depth information in features, thereby achieving better performance than the image-level methods.

DDC. By comparing the performance feature-level and feature-clone methods with and without depth loss, we find that the depth loss is essential in feature-level generation to monocular 3D detection. We use the feature-clone method with depth loss as the baseline and add DDC to the baseline (Exp.3 in Table. 1) to shown the effect. Feature-level generation with the proposed DDC achieves significant improvements against the baseline, indicating that the proposed DDC is effective in generating the virtual right view in feature level for monocular 3D detection. This lies in the fact that the proposed DDC uses the embedded depth representations to dynamically filter the left features, deriving depth-aware feature learning and achieving significant improvements in monocular 3D detection.

5.4. Quantitative and Qualitative Results

We evaluate the three proposed Pseudo-Stereo variants: image-level generation, feature-level generation and feature clone, on KITTI *test* and *val* set. From the above ablation studies, we choose the strategy with better performance for each method: image-level generation without depth loss (Ours-im), feature-level generation with depth loss (Ours-fld) and feature-clone with depth loss (Ours-fcd).

Results on *test* set. Table. 2 shows the performance comparison for *Car* on KITTI *test* server and Table. 3 shows the performance comparison for *Pedestrian* and *Cyclist* on KITTI *test* server. DD3D [33], GUPNet [27] and MonoPSR [20] rank 1^{st} on *Car*, *Pedestrian* and *Cyclist*, respectively, for monocular 3D detection in KITTI-3D benchmark be-

Methods	*Pedestrian* AP_{3D}/AP_{BEV} Easy	Moderate	Hard	*Cyclist* AP_{3D}/AP_{BEV} Easy	Moderate	Hard
D4LCN [12]	4.55 / 5.06	3.42 / 3.86	2.83 / 3.59	2.45 / 2.72	1.67 / 1.82	1.36 / 1.79
MonoPSR [20]	6.12 / 7.24	4.00 / 4.56	3.30 / 4.11	8.37 / 9.87	4.74 / 5.78	3.68 / 4.57
CaDDN [38]	12.87 / 14.72	8.14 / 9.41	6.76 / 8.17	7.00 / 9.67	3.41 / 5.38	3.30 / 4.75
MonoFlex [54]	9.43 / 10.36	6.31 / 7.36	5.26 / 6.29	4.17 / 4.41	2.35 / 2.67	2.04 / 2.50
GUPNet [27]	14.95 / 15.62	9.76 / 10.37	8.41 / 8.79	5.58 / 6.94	3.21 / 3.85	2.66 / 3.64
Ours-im	8.26 / 9.94	5.24 / 6.53	4.51 / 5.72	4.72 / 5.76	2.58 / 3.32	2.37 / 2.85
Ours-fld	**16.95 / 19.03**	**10.82 / 12.23**	**9.26 / 10.53**	**11.22 / 12.80**	**6.18 / 7.29**	**5.21 / 6.05**
Ours-fcd	14.33 / 17.08	9.18 / 11.04	7.86 / 9.59	9.80 / 11.92	5.43 / 6.65	4.91 / 5.86

Table 3. Performance for *Pedestrian* and *Cyclist* on KITTI *test* at IOU threshold 0.5. The best results are **bold**, the second best underlined.

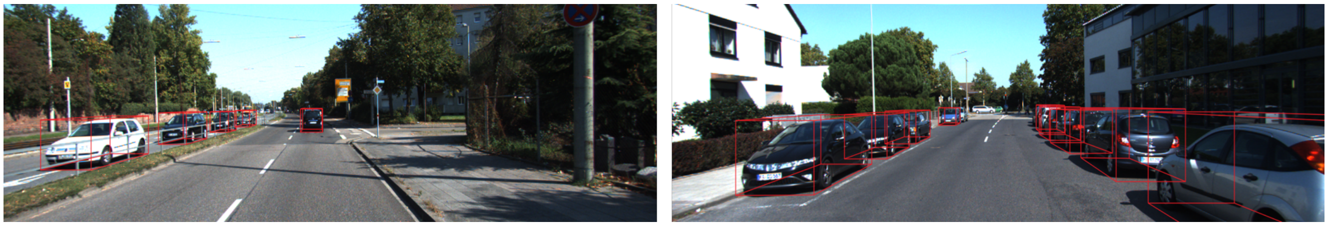

Figure 5. Qualitative results of the best model (Ours-fld) on KITTI *val* set with red 3D bounding boxes.

Methods	AP_{3D} Easy	Moderate	Hard
D4LCN [12]	22.32	16.20	12.30
DDMP-3D [47]	28.12	20.39	16.34
CaDDN [38]	23.57	16.31	13.84
MonoFlex [55]	23.64	17.51	14.83
GUPNet [27]	22.76	16.46	13.72
Ours-im	31.81	22.36	19.33
Ours-fld	**35.18**	**24.15**	**20.35**
Ours-fcd	28.46	19.15	16.56

Table 4. Performance for *Car* on KITTI *val* set at IOU threshold 0.7. The best results are **bold**, the second best underlined.

fore this work. As shown in Table. 2 and Table. 3, Ours-fld achieves better performance than DD3D [33], GUPNet [27] and MonoPSR [20] across all three object classes on both AP_{3D} and AP_{BEV} for monocular 3D detection using single model only. Moreover, our three methods achieve 18/18 best results and 15/18 second-best results across all three object classes on both AP_{3D} and AP_{BEV}. Note that Ours-fld achieves 17 out of 18 best results except the hard level of car, where the best is Our-fcd. This implies that the proposed Pseudo-Stereo 3D detection framework is very effective in monocular 3D detection.

Results on *val* set. As shown in Table. 4, Ours-fld outperforms state-of-the-art methods on KITTI *val* set. Figure 5 shows the qualitative results of Our-fld, the best model, on KITTI *val* set.

Generalization Capability. Usually, there is a large gap between the monocular 3D detection performance on val set and test set because of over-fitting. As shown in Table. 2 and Table. 4, the performance gap for Ours-fcd is much smaller than Ours-im and Ours-fld. This lies in the fact that feature-clone method does not require the estimated depth map for training, leading to better generalization capability. Note that we provide both options in our framework.

6. Conclusion

We propose a Pseudo-Stereo 3D detection framework with three novel virtual view generation methods, including image-level generation, feature-level generation and feature-clone, for detecting 3D objects from a single image, achieving significant improvements in monocular 3D detection. The proposed framework with our feature-level virtual view generation method ranks 1st among the monocular 3D detectors with publications across three object classes on KITTI-3D benchmark. In feature-level virtual view generation, we propose a disparity-wise dynamic convolution with dynamic kernels from disparity feature map to filter the features adaptively from a single image for generating virtual image features, which eases the feature degradation and achieves significant improvements. We analyze two major effects of depth-aware feature learning in our framework.

Broader Impacts. Our Pseudo-Stereo 3D detection framework has the potential to provide a new perspective of monocular 3D detection with Pseudo-Stereo views to our community. Also, our analysis of depth-aware feature learning in Pseudo-Stereo frameworks may give an inspiration to mitigate the performance gap between monocular and stereo 3D detectors.

Acknowledgement: This work was supported by the National Key Research and Development Program of China (2018YFE0183900). Hang Dai would like to thank the support from MBZUAI startup fund (GR006).

References

[1] Wentao Bao, Bin Xu, and Zhenzhong Chen. Monofenet: Monocular 3d object detection with feature enhancement networks. *IEEE Transactions on Image Processing*, 2019. 3

[2] 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, 6

[3] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. *arXiv preprint arXiv:2005.09927*, 2020. 3

[4] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9287–9296, 2019. 2, 7

[5] Yingjie Cai, Buyu Li, Zeyu Jiao, Hongsheng Li, Xingyu Zeng, and Xiaogang Wang. Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. *arXiv preprint arXiv:2002.01619*, 2020. 3

[6] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2040–2049, 2017. 2

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 3

[8] 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. 7

[9] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. *arXiv preprint arXiv:2001.03398*, 2020. 1, 3

[10] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. *arXiv preprint arXiv:2003.00504*, 2020. 7

[11] Xiaomeng Chu, Jiajun Deng, Yao Li, Zhenxun Yuan, Yanyong Zhang, Jianmin Ji, and Yu Zhang. Neighbor-vote: Improving monocular 3d object detection through neighbor distance voting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5239–5247, 2021. 1, 2

[12] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. *arXiv preprint arXiv:1912.04799*, 2019. 2, 6, 7, 8

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 4

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 3

[17] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3153–3163, 2021. 1, 3, 5, 6

[18] 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. 3, 4

[19] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 4

[20] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11867–11876, 2019. 7, 8

[21] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8973–8983, 2021. 2, 7

[22] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 2

[23] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2019. 2

[24] 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

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncer-

tainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3111–3121, October 2021. 2

[27] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 7, 8

[28] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3dssd: Monocular 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6145–6154, 2021. 2

[29] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[30] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6851–6860, 2019. 3, 7

[31] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 2

[32] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[33] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 2, 7, 8

[34] Liang Peng, Fei Liu, Senbo Yan, Xiaofei He, and Deng Cai. OCM3D: object-centric monocular 3d object detection. *CoRR*, abs/2104.06041, 2021. 1, 2

[35] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. *arXiv preprint arXiv:2104.09035*, 2021. 2

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[37] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7607–7615. IEEE, 2019. 3

[38] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2, 7, 8

[39] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. *arXiv preprint arXiv:2106.01178*, 2021. 2

[40] Loren Arthur Schwarz. Non-rigid registration using free-form deformations. *Technische Universität München*, 6, 2007. 4

[41] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 3

[42] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. *arXiv preprint arXiv:2104.03775*, 2021. 7

[43] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Peter Kontschieder, and Elisa Ricci. Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3225–3233, 2021. 1, 2

[44] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1991–1999, 2019. 7

[45] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 3

[46] Jean Marie Uwabeza Vianney, Shubhra Aich, and Bingbing Liu. Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving. *arXiv preprint arXiv:1911.09712*, 2019. 3

[47] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 454–463, 2021. 2, 7, 8

[48] Yan Wang, Bin Yang, Rui Hu, Ming Liang, and Raquel Urtasun. Plume: Efficient 3d object detection from stereo images. *arXiv preprint arXiv:2101.06594*, 2021. 3

[49] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision*, pages 722–740. Springer, 2020. 4

[50] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3

[51] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[52] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *ICLR*, 2020. 1

[53] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *ICLR*, 2020. 3

[54] 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 (CVPR)*, pages 3289–3298, June 2021. 2, 6, 8

[55] 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. 7, 8

[56] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021. 2, 7

[57] Zhikang Zou, Xiaoqing Ye, Liang Du, Xianhui Cheng, Xiao Tan, Li Zhang, Jianfeng Feng, Xiangyang Xue, and Errui Ding. The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2713–2722, October 2021. 2

[58] Zhikang Zou, Xiaoqing Ye, Liang Du, Xianhui Cheng, Xiao Tan, Li Zhang, Jianfeng Feng, Xiangyang Xue, and Errui Ding. The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2713–2722, 2021. 7