# Decoupled Pseudo-labeling for Semi-Supervised Monocular 3D Object Detection

Jiacheng Zhang[1*] Jiaming Li[1*] Xiangru Lin[2*] Wei Zhang[2]
Xiao Tan[2] Junyu Han[2] Errui Ding[2] Jingdong Wang[2] Guanbin Li[1,3†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Vision Technology (VIS), Baidu Inc., China
[3]GuangDong Province Key Laboratory of Information Security Technology
{zhangjch58,lijm48}@mail2.sysu.edu.cn, {liguanbin}@mail.sysu.edu.cn
{linxiangru,zhangwei99,tanxiao01,hanjunyu,dingerrui,wangjingdong}@baidu.com

## Abstract

*We delve into pseudo-labeling for semi-supervised monocular 3D object detection (SSM3OD) and discover two primary issues: a misalignment between the prediction quality of 3D and 2D attributes and the tendency of depth supervision derived from pseudo-labels to be noisy, leading to significant optimization conflicts with other reliable forms of supervision. To tackle these issues, we introduce a novel decoupled pseudo-labeling (DPL) approach for SSM3OD. Our approach features a Decoupled Pseudo-label Generation (DPG) module, designed to efficiently generate pseudo-labels by separately processing 2D and 3D attributes. This module incorporates a unique homography-based method for identifying dependable pseudo-labels in Bird's Eye View (BEV) space, specifically for 3D attributes. Additionally, we present a Depth Gradient Projection (DGP) module to mitigate optimization conflicts caused by noisy depth supervision of pseudo-labels, effectively decoupling the depth gradient and removing conflicting gradients. This dual decoupling strategy—at both the pseudo-label generation and gradient levels—significantly improves the utilization of pseudo-labels in SSM3OD. Our comprehensive experiments on the KITTI benchmark demonstrate the superiority of our method over existing approaches.*

## 1. Introduction

Monocular 3D Object Detection (M3OD) is designed to detect objects in 3D space using a single 2D RGB image as input, playing a pivotal role in contemporary 3D perception systems across applications like autonomous driving and robotic navigation. The major challenge in current M3OD methods lies in their dependence on precise annotations, a labor-intensive and costly process. To overcome this obstacle, Semi-Supervised Monocular 3D Object Detection (SSM3OD) has emerged as a promising solution. It capitalizes on the abundance of readily available unlabeled images to enhance the performance of M3OD detectors. In line with prevalent semi-supervised learning techniques [1, 30, 31, 39], pseudo-labeling and consistency regularization are two kinds of widely used technology in SSM3OD [16, 37]. This paper specifically explores the pseudo-labeling technique within the realm of SSM3OD.

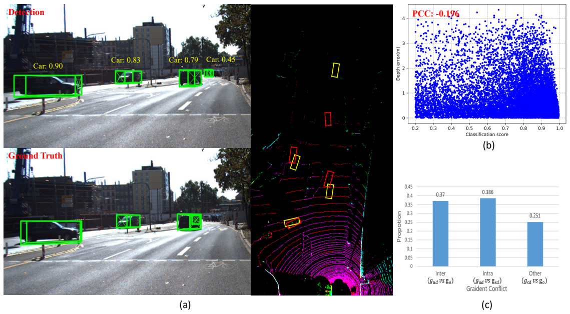


Figure 1. (a) Visualization of Pseudo-Labels and Ground Truth in Image Plane and Bird's Eye View (BEV) Plane. Red: Ground Truth. Yellow: Detected Bounding Boxes. (b) Statistical analysis of classification scores and depth errors (PCC: Pearson Correlation Coefficient). (c) The proportion of different types of gradient conflicts occurring. $g_{sd}$: Gradient of ground truth depth loss. $g_{ud}$: Gradient of pseudo-label depth loss. $g_o$: Gradient of other attribute supervision loss. Gradient conflicts between $g_i$, $g_j$ when $\cos(g_i, g_j) < 0$.

M3OD is inherently a multi-task challenge, encompassing a range of both 2D (e.g. classification) and 3D (e.g. depth) attribute predictions. We observe that there is a significant disparity between the 2D and 3D attributes.

[*] Equally-contributed authors. Work done during an internship at Baidu.

[†] Corresponding author.

As illustrated in Fig. 1 (a), the detection with high confidence scores has subpar depth and orientation predictions in considerable cases. Taking the depth attribute as an example, our analysis reveals a weak correlation (PCC: -0.196) between the quality of depth prediction and the associated classification confidence, as depicted in Fig. 1 (b). This issue stems from the perspective projection, which complicates the distinction of 3D attribute quality on the 2D image plane, as illustrated in Fig. 1 (a). However, most existing SSM3OD methods [16, 37] overlook this disparity and only rely on the accuracy of 2D attributes(e.g. confidence score) to achieve pseudo-label generation, which leads to unreliable supervision for the 3D attributes.

To address this issue, we introduce a **decoupled pseudo-label generation** (DPG) module to generate more effective pseudo-labels for both 2D and 3D attributes. Specifically, given the disparity between 2D and 3D attributes, we propose to separate the pseudo-label generation for these two types and develop a Homography-based Pseudo-label Mining (HPM) module to generate pseudo-labels specifically for 3D attributes. Leveraging the estimated 2D-3D homography transformation, HPM transforms predictions from the 2D image plane to the 3D Bird's Eye View (BEV) plane, in which the pseudo-labels with reliable 3D attributes are iteratively identified based on the localization error. However, due to the noisy nature of the depth estimation, we observe a frequent conflict between the depth supervision derived from pseudo-labels and other reliable supervision sources (ground truth of depth, ground truth & pseudo-label of the attributes except depth). As illustrated in Fig. 1 (c), the gradient conflicts between the pseudo-label depth loss and other reliable supervision loss (represented as $g_{ud}$ vs $g_{sd}$ and $g_{ud}$ vs $g_o$) is more prevalent compared to conflicts within the reliable supervision ($g_{sd}$ vs $g_o$). Such gradient conflicts potentially undermine the utilization of reliable supervision.

To mitigate this issue, we further develop a **depth gradient projection** (DGP) module. This module effectively projects the conflicting depth gradient towards the principal reliable gradient, eliminating the harmful component. This adjustment ensures that the depth supervision derived from pseudo-labels is always in harmony with reliable supervision. By incorporating both the DPG and DGP modules, our Decoupling Pseudo-Labeling (DPL) approach significantly enhances the generation and utilization of pseudo-labels for SSM3OD. We have conducted comprehensive experiments to validate the efficacy of our method on the KITTI [9] benchmark and achieved state-of-the-art results. Our contributions can be summarized as follows:

- We identify and address the quality misalignment between the predictions of 2D and 3D attributes, an issue previously overlooked in existing pseudo-labeling SSM3OD methods.
- We introduce a decoupled pseudo-label generation module featuring a homography-based depth label mining module to generate reliable pseudo-labels for both 2D and 3D attributes.
- We develop a depth gradient projection module to mitigate the adverse effects potentially caused by noisy depth pseudo-labels.
- Our extensive experimental results on the KITTI benchmark demonstrate that our approach significantly surpasses all previous state-of-the-art methods.

## 2. Related Work

**Monocular 3D Object Detection**. Monocular 3D object detection (M3OD) aims to detect objects within a three-dimensional space utilizing solely a single camera. Existing methods in M3OD can be broadly categorized into two streams: one that relies exclusively on monocular images, and another that incorporates supplementary data sources, such as CAD models [19], dense depth map [7, 21, 33, 36, 38], and LiDAR [3, 4, 11, 11, 15, 25]. Owing to their cost-effectiveness and ease of deployment, we focus on the methods that rely solely on monocular images as input. Initial efforts in the field [2, 18, 34] adapted conventional 2D object detection frameworks [26, 32, 43] to incorporate 3D object detection capabilities. Studies such as MonoDLE [22] and PGD [35] have highlighted the critical challenge in M3OD: precise depth prediction. In response, numerous studies have sought to harness the synergy between 2D-3D geometric relationships [12, 14, 20, 28, 29] or exploit spatial context [6, 10, 13, 14, 17] to enhance depth estimation accuracy. MonoFlex [41] introduces a novel depth ensemble approach, synthesizing various depth estimation techniques to elevate detection performance significantly. However, these advancements are largely contingent upon annotations with precise depth, which are labor-intensive and costly to obtain. Consequently, this research explores the potential of semi-supervised learning methodologies to alleviate the annotation burden.

**Semi-Supervised Monocular 3D Object Detection**. Semi-supervised monocular 3D object detection (SSM3OD) harnesses a wealth of unlabeled monocular imagery alongside a limited corpus of precisely annotated labeled monocular images to enhance monocular 3D object detection efficacy. Mix-Teaching [37] introduces a database-oriented pseudo-labeling strategy that pastes pseudo-instances onto the background unlabeled images, thereby generating additional training samples. It further includes a model prediction ensemble-based pseudo-label filter to isolate high-quality pseudo-labels. However, this method does not fully address the distinct characteristics between 2D and 3D attributes, resulting in sub-optimal exploitation of 3D information in pseudo-label generation, and consequently, less effective pseudo-labels. MVC-

MonoDet [16] focuses on the consistency regularization technique and designs a multi-view consistency strategy to exploit the depth clue in the unlabeled multi-view monocular images (video, stereo images). Our method is orthogonal with [16] and focuses specifically on the relatively underexplored pseudo-labeling strategies within SSM3OD. It is noteworthy that [23, 24] also propose pseudo-labeling methods for monocular 3D object detection. However, these methods derive pseudo-labels using LiDAR point clouds, which inherently provide accurate depth information for objects. In contrast, our method generates pseudo-labels exclusively from monocular images, without relying on supplementary LiDAR data, presenting a more challenging yet practical scenario.

## 3. Preliminary

**Problem Definition**. Given the labeled dataset as $\boldsymbol{D}_l = (\boldsymbol{I}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N_l}$, and unlabeled dataset $\boldsymbol{D}_u = \{(\boldsymbol{I}_i^u)\}_{i=1}^{N_u}$, where $\boldsymbol{I}_i^l$ and $\boldsymbol{I}_i^u$ denote an RGB image of labeled dataset and the unlabeled dataset respectively, and $N_l$ and $N_u$ is the corresponding data amounts. $\boldsymbol{y}_i^l = \{(c_j^l, \boldsymbol{o}_j^l)\}_{j=1}^{N_i}$ is a list of $N_i$ 3D bounding box annotations for $i$-th labeled image, where $c_j^l$ is the category label and $\boldsymbol{o}_j^l$ is the 3D box label including the orientation, dimension, and location. The target of SSM3OD is to achieve monocular 3D object detection by leveraging limited labeled images with additional abundant unlabeled images. The optimization of the SSM3OD can be formulated as:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha \mathcal{L}_{unsup}, \tag{1}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$ are the supervised loss and unsupervised loss and $\alpha$ is the loss weight and set to 1 by default.

## 4. Method

Fig. 2 presents an overview of our Decoupled Pseudo-Labeling (DPL) approach for Semi-Supervised Monocular 3D Object Detection (SSM3OD). It employs the classic teacher-student framework [31], where a teacher and student network are involved, both of which are initialized by a supervised pre-trained model. The teacher model generates pseudo boxes on unlabeled images, while the student model is trained with labeled images with ground truth annotation and unlabeled images with pseudo labels. The teacher model is iteratively updated from the student model using the exponential moving average (EMA) strategy. Our DPL method integrates two key modules: Decoupled Pseudo-label Generation (DPG) and Depth Gradient Projection (DGP), to enhance the effectiveness of pseudo-label utilization in SSM3OD.

### 4.1. Decoupled Pseudo-label Generation

Given the fact that the prediction quality of 2D and 3D attributes in SSM3OD is not aligned, we propose to decouple the pseudo-label generation process for these two types of attributes. We classify object category, 2D size, and the projected 3D center as 2D attributes, and depth, 3D size, and orientation as 3D attributes. For 2D attributes, the classification confidence threshold $\theta_c$ is used to filter pseudo-labels, following [16]. For the 3D attributes, we introduce a novel approach that leverages the homography geometric relationship for 3D attribute pseudo-label generation. This forms the basis of our Decoupled Pseudo-label Generation (DPG), the unsupervised loss can be formulated as:

$$\mathcal{L}_{unsup} = \mathcal{L}_{2D}(x_{2D}, y_{2D}) + \alpha \mathcal{L}_{3D}(x_{3D}, y_{3D}), \tag{2}$$

where $x_{2D}$, $x_{3D}$ are 2D and 3D attributes prediction of model output, and $y_{2D}$, $y_{3D}$ are the two groups of pseudo-labels for 2D and 3D attributes, respectively. The loss functions $\mathcal{L}_{2D}$ and $\mathcal{L}_{3D}$ are consistent with MonoFlex [41]. With the decoupled design, both 2D and 3D attributes can be supervised with more effective pseudo-labels.

**Homography-based Pseudo-Label Mining** Due to the perspective projection, accurately gauging the quality of 3D attribute predictions for bounding boxes on the image plane poses a challenge. In response, we have crafted a unique homography-based depth pseudo-label mining module. This module's pivotal feature is the transposition of predictions from the 2D image plane to the Bird's Eye View (BEV) plane using homography transformation. This shift significantly improves the precision of assessing 3D attribute predictions, such as depth and orientation.

*Homography Transformation*: Generally, let the coordinates of a point on the **ground surface** as $I = (u, v)$ in the 2D Image plane and $B = (x^b, y^b)$ in the BEV plane as shown in Fig. 3. The transformation between homogeneous coordinates $(u, v, 1)$ and $(x^b, y^b, 1)$ can be describe by a homography matrix $\boldsymbol{M} \in \mathbb{R}^{3\times3}$:

$$[x^b, y^b, 1]^T = \boldsymbol{M}[u, v, 1]^T. \tag{3}$$

The homography transformation [8, 42] is a geometric relationship between two coordinate systems of 2D and 3D plane. Therefore, with the flat ground assumption [10], different objects within an image will share the same homography matrix.

*Iteratively Pseudo-Label Mining*: We develop an iterative pseudo-label mining algorithm to acquire reliable 3D attribute pseudo-labels, as detailed in 4.1. This algorithm uses the deviation from a consistent homography transformation as a measure of 3D attribute prediction quality.

Specifically, it initially selects pseudo-labels with relatively accurate 3D attribute predictions to estimate the initial homography matrix reliably. To assist in this initial

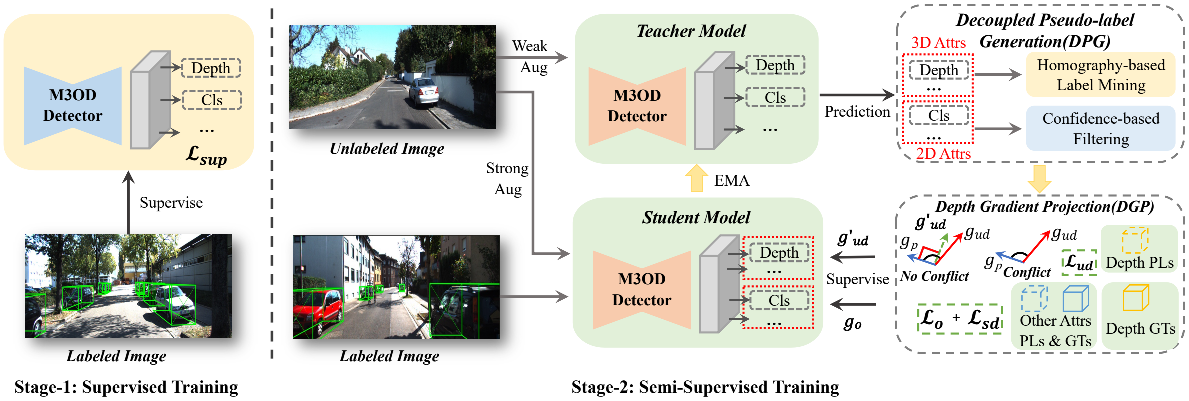


Figure 2. The overview of the **Decoupled Pseudo-Labeling** (DPL) method for SSM3OD. We conduct semi-supervised learning based on the teacher-student framework after the supervised training stage. DPL consists of *Decoupled Pseudo-label Generation* (DPG) module and *Depth Gradient Projection* (DGP) module. DPG decouple the 2D and 3D attribute and generate pseudo-labels independently, with a *Homography-based Label Mining* (HLM) algorithm designed to generate pseudo-labels 3D attributes by harnessing the homography transformation. DGP module utilizes a gradient projection operation to mitigate the potential negative impact of noisy depth supervision.

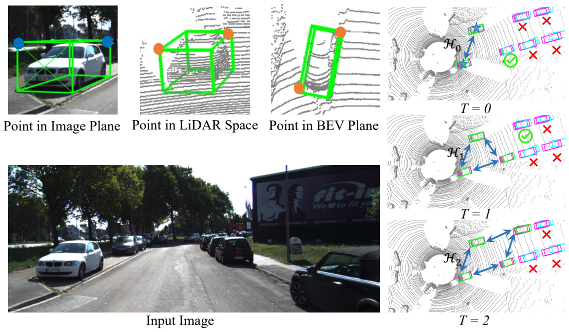


Figure 3. Illustration of HPM module. The homography transformation describes the coordinate mapping between the image plane and the BEV plane. Starting with the initial pseudo-labels, we iteratively estimate the homography matrix $H_i$ and search the reliable pseudo-labels in BEV space. Green Box: The selected Pseudo-Labels. Cyan Box: Model Prediction. Purple Box: Ground Truth.

generation of pseudo-labels, we employ the depth prediction uncertainty $\sigma$ from the Laplacian aleatoric uncertainty loss [6], as defined in Eq. 4:

$$\mathcal{L}_{depth} = \frac{\sqrt{2}}{\sigma} \left\| d_{gt} - d_{\text{pred}} \right\|_1 + \log(\sigma), \tag{4}$$

where $d_{gt}$ is the ground truth depth and $d_{\text{pred}}$ is the predicted depth of an object, $\sigma$ is known as the predicted depth uncertainty to weight the prediction. We select the pseudo-labels with $\sigma < \theta_u$ as the initial pseudo-labels.

Then, we select the bottom corner points and bottom center points (5 points) of each initial 3D pseudo-bounding box as candidate points to estimate the homography matrix. The coordinates of these points in the image plane $\tilde{I} = (\tilde{u}, \tilde{v})$ are directly predicted by the teacher model [22, 41]. We then estimate the coordinates of these points in the BEV plane. Specifically, we first get the positions of these points in the camera coordinate system $(\tilde{x}, \tilde{y}, \tilde{z})$ by the local transformations [15] which is estimated from the dimensions, orientations, and positions of the centers of the 3D boxes.

Then the coordinate in the lidar coordinate system can be obtained by projecting with the inverse of extrinsic matrices $[\mathbf{R}|\mathbf{T}]$,

$$[\tilde{x}^b, \tilde{y}^b, \tilde{z}^b]^T = [\mathbf{R}|\mathbf{T}]^{'} [\tilde{x}, \tilde{y}, \tilde{z}]^T. \tag{5}$$

Therefore, the coordinates of these points in the BEV plane are $\tilde{B} = (\tilde{x}^b, \tilde{y}^b)$. Denote the candidate points coordinates in the image plane and the BEV plane of $N$ objects, as $\tilde{C}_I \in \mathbb{R}^{2\times 5N}$ and $\tilde{C}_B \in \mathbb{R}^{2\times 5N}$, the homography transformation $\tilde{\boldsymbol{M}}$ can be derived by solving Eq.3 with Direct Linear Transform (DLT) [27].

Finally, we apply the estimated homography matrix $\tilde{\boldsymbol{M}}$ to the candidate points of the predicted bounding boxes that have not been chosen as pseudo-labels yet to get their desired coordinates in BEV space:

$$\hat{B} = [\hat{x}^b, \hat{y}^b, 1]^T = \tilde{\boldsymbol{M}} [\tilde{u}, \tilde{v}, 1]^T. \tag{6}$$

Ideally, the BEV coordinated obtained by the homography transformation $\hat{c}^b = (\hat{x}^b, \hat{y}^b)$ should be the same as $\tilde{c}^b_t = (\tilde{x}^b, \tilde{y}^b)$ estimated from the model prediction via Eq.5. However, the poorly predicted 3D attributes (e.g. depth, orientation) would result in the deviation. Therefore, this deviation can serve as the proxy for the prediction quality of these attributes, and we select the prediction satisfying $||\hat{c}^b - \tilde{c}^b_t||_2 < \theta_h$ as the eligible pseudo-labels, where $\theta_h$ is the pre-defined threshold. The newly obtained pseudo-labels are then appended to the previously acquired pseudo-labels to start a new iteration of homography matrix estimation and pseudo-label filtering as shown in Fig. 3. The iterative process continues until either no new pseudo-labels are obtained or the predefined maximum iteration limit, denoted as $t_{max}$, is reached. Please refer to the Appendix for the complete algorithm.

### 4.2. Depth Gradient Projection

Due to the inherent limitations of depth estimation from monocular images, the depth supervision from pseudo-labels will be inevitably noisy and can cause optimization conflict (i.e. conflicted gradient direction) with reliable supervision. Concretely, we split the total loss into pseudo-label depth loss $\mathcal{L}_{ud}$, ground truth depth loss $\mathcal{L}_{sd}$ and other attributes loss $\mathcal{L}_o$ (including both labeled and unlabeled loss). Their gradient are denoted as $g_{ud}(\theta)$, $g_{sd}(\theta)$ $g_o(\theta)$, respectively. Generally, since the pseudo-labels of the attributes except for depth can be estimated with reasonable accuracy than depth [35], $\mathcal{L}_{sd}$ and $\mathcal{L}_o$ can be regarded more reliable than $\mathcal{L}_{ud}$. We check the optimization conflicts between different supervision as presented in Fig. 1. It clearly shows that the gradient from depth pseudo-labels conflicts with reliable supervision more frequently.

To address this issue, we develop a simple depth gradient projection module to eliminate the possible negative impact of noisy depth supervision from the gradient perspective.

Concretely, given that $g_{sd}$ and $g_o$ conflict less frequently, we combine them together and treat them as the optimization principle gradient $g_p = g_o + g_{sd}$, which stands for optimization direction of reliable supervision. Then, we project the $g_{ud}(\theta)$ to the normal vector of gradient $g_p(\theta)$ when the conflict occurs:

$$g_{ud}^{'}(\theta) = \begin{cases} g_{ud}(\theta) - \frac{g_{ud}(\theta)g_p(\theta)}{||g_p(\theta)||_2^2} .g_p(\theta), & if\ cos(g_{ud}, g_p) < 0, \\ g_{ud}, & otherwise \end{cases} \tag{7}$$

The obtained gradient $g_{ud}^{'}(\theta)$ thus has no conflicted gradient component with $g_p(\theta)$. Equipped with this module, the noisy unsupervised depth loss is always guaranteed to share common interests with reliable supervision and deliver an equilibrium optimization target.

## 5. Experiments

### 5.1. Experimental Setup

**Dataset and Metrics**. **KITTI** dataset [9] is the standard dataset for M3OD, providing 7,481 images for training and 7,518 images for testing. Following the common practice [5], the training set is further split into 3,712 training samples and 3,769 validation samples. For the unlabeled data, we select the *completely unlabeled video sequence in the KITTI raw data that does not overlap with the video sequence of the training and validation split*. This results in approximately 35K unlabeled images, which are utilized as our unlabeled dataset. We report the evaluation results on the validation and test set based on $AP|_{R_{40}}$.

**Implementation Details**. We choose MonoFlex [41], a representative M3OD detector, to evaluate the effectiveness of our method. All experiments are conducted using the official code provided by the author. We first pre-train the model using labeled data and then perform end-to-end semi-supervised learning using both labeled and unlabeled data. Each unlabeled image has weak and strong augmented versions, which are sent to the teacher and student network respectively. The strong augmentation includes random horizontal flips, photometric distortion, random gray, and random Gaussian blur, while the weak augmentation only involves random horizontal flips. During pseudo-label generation with the outputs of the teacher network, we first filter out predictions that are background with a classification score threshold of 0.2. The $\theta_c$, $\theta_u$, and $\theta_h$ in the DPG module are set to 0.4, 0.1, and 2.0 respectively, and $t^{max}$ is set to 10. More implementation details are in the Appendix.

### 5.2. Main Results

Tab. 1 presents the results on the KITTI test set. It shows that by incorporating the proposed DPL for SSM3OD, our method significantly boosts the performance of the base detector. In particular, our method boosts the performance of MonoFlex with **+4.10** and **+4.18** on $AP_{3D}$ and $AP_{BEV}$, respectively. Moreover, based on MonoFlex, our method surpasses all existing SSM3OD methods by a large margin and achieves a new state-of-the-art performance across all fully supervised and semi-supervised methods. Specifically, integrating our method into the MonoFlex, our method outperforms Mix-Teaching by **+2.33** $AP_{3D}$ and **+2.61** $AP_{BEV}$, and exceeds the performance of MVC-MonoDet by **+2.11** and **+2.01** on $AP_{3D}$ and $AP_{BEV}$.

### 5.3. Ablation Study

**Components Effectiveness**. We ablate the effects of decoupled pseudo-label generation (DPG) and depth gradient projection (DGP) module. The results are presented in Tab.5. We start from the **SSM3OD Baseline** that utilizes the classification confidence threshold 0.6 to filter the pseudo-labels for both 2D and 3D attributes. It clearly shows that integrating the decoupled pseudo-label generation(DPG) module improves the performance by **1.21** on Car $AP_{3D}$(Mod.) without bells and whistles. It demonstrates the importance of decoupling the pseudo-labels generation process for 2D and 3D attributes. Further applying the depth gradient projection (DGP) module to eliminate potential gradient conflict leads to a **0.62** Car $AP_{3D}$(Mod.) improvement and reaches 19.85 $AP_{3D}$ (Mod.), which is **1.83** $AP_{3D}$ higher than the baseline.

**Performance on More Base Detectors**. Beyond the MonoFlex, we conducted experiments on the KITTI validation dataset with other monocular 3D object detectors, including MonoDLE[22], PGD[35], GUPNet[20], and

Table 1. Comparision with state-of-the-art (SOTA) Methods. We present the evaluation results of **'Car' category in the KITTI test set**. † denotes our reproduction results. For fair comparisons, we train Mix-Teaching with the same data volume as our method.

| Method | Extra Data | Test $AP_{3D}\|R_{40}$ | | | Test $AP_{BEV}\|R_{40}$ | | |
|---|---|---|---|---|---|---|---|
| | | Easy | Mod. | Hard | Easy | Mod. | Hard |
| PatchNet | Depth | 15.68 | 11.12 | 10.17 | 22.97 | 16.86 | 14.97 |
| D4LCN | | 16.65 | 11.72 | 9.51 | 22.51 | 16.02 | 12.55 |
| DDMP-3D | | 19.71 | 12.78 | 9.80 | 28.08 | 17.89 | 13.44 |
| Kinematic3D | Multi-frames | 19.07 | 12.72 | 9.17 | 26.69 | 17.52 | 13.10 |
| MonoRUn | LiDAR | 19.65 | 12.30 | 10.58 | 27.94 | 17.34 | 15.24 |
| CaDDN | | 19.17 | 13.41 | 11.46 | 27.94 | 18.91 | 17.19 |
| MonoDTR | | 21.99 | 15.39 | 12.73 | 28.59 | 20.38 | 17.14 |
| AutoShape | CAD | 22.47 | 14.17 | 11.36 | 30.66 | 20.08 | 15.59 |
| SMOKE | None | 14.03 | 9.76 | 7.84 | 20.83 | 14.49 | 12.75 |
| MonoPair | | 13.04 | 9.99 | 8.65 | 19.28 | 14.83 | 12.89 |
| RTM3D | | 13.61 | 10.09 | 8.18 | - | - | - |
| PGD | | 19.05 | 11.76 | 9.39 | 26.89 | 16.51 | 13.49 |
| MonoRCNN | | 18.36 | 12.65 | 10.03 | 25.48 | 18.11 | 14.10 |
| Zhang et al. DLE | | 20.25 | 14.14 | 12.42 | 28.85 | 17.72 | 17.81 |
| GUPNet | | 20.11 | 14.20 | 11.77 | - | - | - |
| HomoLoss | | 21.75 | 14.94 | 13.07 | 29.60 | 20.68 | 17.81 |
| Mix-Teaching | Unlabeled | 21.88 | 14.34 | 11.86 | 30.52 | 19.51 | 16.45 |
| MVC-MonoDet | Unlabeled | 22.13 | 14.56 | 12.09 | 31.62 | 20.11 | 17.21 |
| MonoFlex† | None | 19.23 | 12.57 | 10.73 | 26.83 | 17.94 | 15.16 |
| DPL$_{FLEX}$ | Unlabeled | **24.19** | **16.67** | **13.83** | **33.16** | **22.12** | **18.74** |
| ***Improvement*** | v.s. Baseline | **+4.96** | **+4.10** | **+3.10** | **+6.33** | **+4.18** | **+3.58** |

Table 2. Performance of **'Car' category in the KITTI validation set** based on MonoFlex under different amounts of unlabeled data. We randomly chose 5K, 15K, and 25K data from the whole 35K KITTI unlabeled set as the unlabeled data.

| Methods | Val, $AP_{3D}\|R_{40}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5K unlabel | | | 15K unlabel | | | 25K unlabel | | | 35k unlabel | | |
| | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| MonoFlex † | 22.80 | 17.51 | 14.90 | 22.80 | 17.51 | 14.90 | 22.80 | 17.51 | 14.90 | 22.80 | 17.51 | 14.90 |
| DPL$_{FLEX}$ | **24.61** | **18.76** | **16.39** | **25.23** | **18.86** | **16.47** | **26.05** | **19.22** | **16.84** | **26.51** | **19.84** | **17.13** |
| ***Improvement*** | **+1.81** | **+1.25** | **+1.49** | **+2.43** | **+1.35** | **+1.57** | **+3.25** | **+1.71** | **+1.94** | **+3.71** | **+2.33** | **+2.23** |

MonoDETR[40]. We specifically omitted FCOS3D from this study, as its performance on the KITTI dataset is suboptimal as acknowledged by the authors. The results of these experiments are detailed in Tab.4, showing consistent and substantial performance improvements across the various base detectors. These findings underscore the strong adaptability to different monocular detectors of our method.

**Effect of Labeled and Unlabeled Data Volume.** We present the impact of labeled and unlabeled data volume on the performance of DPL in Tab.3 and Tab.2. Our approach consistently enhances the performance of MonoFlex across various volumes of labeled data. Particularly, our method showcases significant benefits in scenarios where labeled data is scarce. For instance, we observe a substantial performance boost of **+4.29** in $AP_{3D}|40$ when only 10% of the labeled training images are available. These results highlight the superiority of our method in effectively leveraging limited labeled data. Furthermore, as the volume of unlabeled data increases, DPL showcases a more pronounced improvement in performance. This underscores the scalability of our method, highlighting its ability to leverage larger amounts of unlabeled data effectively.

**Analysis of Decoupled Pseudo-label Generation**. We ablate different ways for pseudo-label generation in Tab.6. It clearly shows that utilization of the classification confidence threshold(thr=0.6) for pseudo-label generation only brings limited improvement (+0.51 $AP_{3D}$). This is attributed to its poor ability to reflect the prediction quality of 3D attributes, especially depth, leading to noisy depth pseudo-labels with large depth prediction errors as presented in Fig.5. As reported by [22], the model exhibits reasonable performance in predicting objects at close range but has

Table 3. Performance of **'Car' category in the KITTI validation set** based on MonoFle under different labeled ratios. We randomly chose 10%, 50%, and 100% of KITTI train split as the labeled data.

| Methods | Val, $AP_{3D}\|R_{40}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 10% | | | 50% | | | 100% | | |
| | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| MonoFlex † | 4.77 | 3.90 | 3.29 | 18.85 | 14.81 | 12.48 | 22.80 | 17.51 | 14.90 |
| DPL$_{FLEX}$ | **10.25** | **8.19** | **7.09** | **21.33** | **16.42** | **14.57** | **26.51** | **19.84** | **17.13** |
| ***Improvement*** | **+5.48** | **+4.29** | **+3.8** | **+2.48** | **+1.61** | **+2.09** | **+3.71** | **+2.33** | **+2.23** |

Table 4. Performance of **'Car' category in the KITTI validation set** under different base detectors. * Note that the provided code of MonoDETR is only an intermediate version (not complete) which is confirmed officially by the authors. We also do not reproduce the results reported in their paper.

| Methods | Val, $AP_{3D}\|R_{40}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MonoDLE | | | PGD | | | GUPNet | | | MonoDETR* | | |
| | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| Sup Baseline | 17.25 | 13.87 | 11.83 | 19.27 | 13.22 | 10.64 | 22.76 | 16.46 | 13.72 | 26.95 | 18.87 | 15.52 |
| DPL | **19.31** | **15.67** | **13.72** | **21.34** | **15.34** | **12.49** | **24.48** | **18.51** | **14.89** | **28.12** | **20.81** | **17.37** |
| ***Improvement*** | **+2.06** | **+1.80** | **+1.50** | **+2.07** | **+2.12** | **+1.85** | **+1.72** | **+2.05** | **+1.17** | **+1.17** | **+1.94** | **+1.85** |

Table 5. Ablation of the effectiveness of proposed components. The experiments were conducted on the KITTI validation set with MonoFlex. Car category's performance $AP_{3D}/AP_{BEV}|R_{40}$ of $IoU = 0.7$ is reported

| DPG | DGP | Easy | Mod. | Hard |
|---|---|---|---|---|
| - | - | 23.13/30.14 | 18.02/23.83 | 15.24/20.18 |
| √ | - | 26.24/34.91 | 19.23/25.64 | 17.04/22.60 |
| √ | √ | **26.51/35.02** | **19.85/26.37** | **17.13/23.08** |

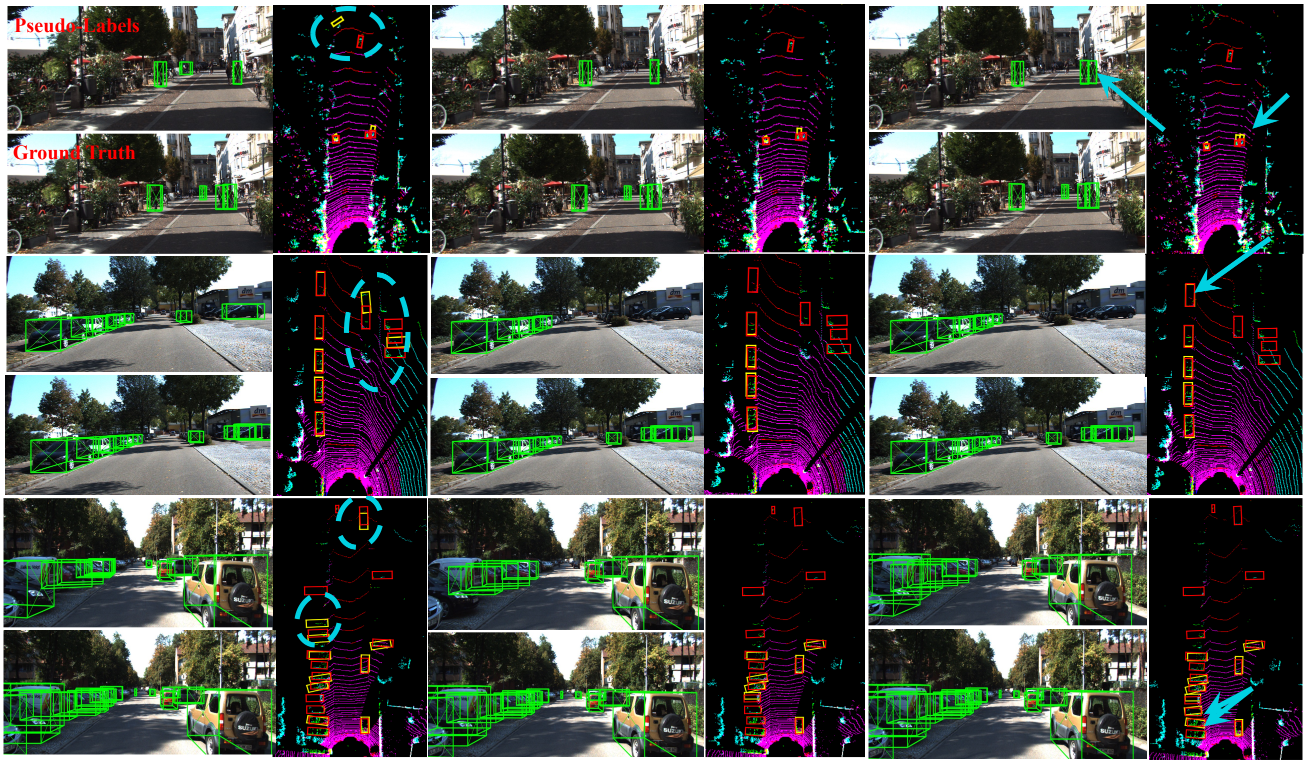
limitations in predicting objects at a distance. Therefore, we generate the pseudo-labels by only retaining the prediction with a detection distance of less than 45m as suggested by [22]. As presented in the third row of Tab.6, a notable performance boost (+1.01 $AP_{3D}$ for moderate) is observed. Nevertheless, completely disregarding pseudo-labels beyond a certain distance can impede the model's ability to detect objects that are located far away from the ego-car. This limitation is supported by the marginal improvement of 0.36 in the detection of hard category objects, compared to the use of confidence thresholding. In contrast, our DPG leverages the geometric relationship between the 2D and 3D space through homography transformation, enabling us to derive more effective pseudo-labels with more accurate depth from the more distinguishable BEV plane as proved in Fig.5. By incorporating these pseudo-labels for the supervision of both 2D and 3D attributes, we ultimately achieve significantly improved performance. By further decoupling the supervision of 2D and 3D attributes and generating pseudo-labels for 2D attributes via confidence thresholding, we are able to harness the potential of pseudo-labels with accurate 2D attribute prediction but poor 3D attribute prediction. This further enhances the performance of our approach. These results clearly highlight the significance of separately processing the 2D and 3D attributes during the pseudo-labeling process.

Table 6. Effectiveness of different strategies to generate the pseudo-labels for SSM3OD with MonoFlex. **cls confidence**: Filter the pseudo-labels with a classification confidence threshold 0.6. **det distance**: Generate the pseudo-labels by retaining the prediction with a detection distance of no more than 45m. **DPG w.o decouple**: Take the pseudo-labels generated by homography label mining for both 2D and 3D attributes supervision.

| Strategy | Val, $AP_{3D}\|R_{40}$ | | |
|---|---|---|---|
| | Easy | Mod | Hard |
| sup baseline | 22.80 | 17.51 | 14.90 |
| cls confidence | 23.13 | 18.02 | 15.24 |
| det distance | 23.47 | 18.52 | 15.60 |
| DPG w.o decouple | 25.66 | 19.04 | 16.24 |
| DPG | **26.24** | **19.23** | **17.04** |

**Visualization of Decoupled Pseudo-labeling**. We visualize the pseudo-labels generated by the classification confidence thresholding and DPG in Fig.4. Our DPG first selects the pseudo-labels via depth prediction uncertainty, which leads to the initial pseudo-labels with accurate depth. Subsequently, the homography-based pseudo-label mining further identifies additional pseudo-labels with reasonable depth and orientation predictions. In contrast, the pseudo-labels generated solely by confidence thresholding tend to be noisy, as they often include pseudo-labels with high confidence but inaccurate depth estimations.

**Analysis of Depth Gradient Projection** We conducted an analysis to examine the connection between the gradient similarity of $g_{ud}$ and $g_p$, and the quality of depth prediction.



(a) (b) (c)

Figure 4. Visual comparison of pseudo-labels generated among different pseudo-label generation strategies. (a) Pseudo labels generated with classification confidence threshold 0.6. (b) Pseudo labels generated by initial depth prediction uncertainty filtering in HPM. (c) Pseudo labels after HPM algorithm. Red Box: Ground truth. Yellow Box: Pseudo-Labels. Cyan dashed circles: The confident yet depth-deviated pseudo-labels. Cyan arrows: The pseudo-labels discovered through homography-based mining.

Our findings indicate a clear correlation between the depth error and the gradient similarity between the unsupervised depth gradient and the principal gradient. As presented in the left of Fig.5, it is evident that samples with larger deviations result in a more pronounced gradient conflict with reliable supervision. This further emphasizes the significance of our depth gradient projection module in mitigating the adverse effects of noisy pseudo-labels.

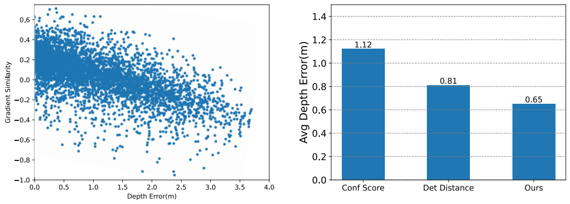


Figure 5. Left: The correlation between the gradient similarity $cos(g_{ud}, g_p)$ and the depth error. Right: The average depth error of the pseudo-labels obtained in different ways.

## 6. Conclusion

In this work, we introduced a decoupled pseudo-labeling approach for Semi-Supervised Monocular 3D Object Detection (SSM3OD), designed to optimize the use of pseudo-labels more effectively. This approach features a decoupled pseudo-label generation module, incorporating a homography-based pseudo-label mining algorithm to efficiently provide reliable pseudo-labels for both 2D and 3D attributes. Additionally, we developed a depth gradient projection module to mitigate the adverse effects of noisy depth supervision. Comprehensive evaluations on the KITTI benchmark validate the effectiveness of our proposed method, demonstrating its superior performance in SSM3OD.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NO. 62322608), in part by the CAAI-MindSpore Open Fund, developed on OpenI Community, in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU, No. 20230003).

## References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 1

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

[3] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 135–152. Springer, 2020. 2

[4] 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

[5] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015. 5

[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, 4

[7] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, pages 1000–1001, 2020. 2

[8] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5

[10] 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, 3

[11] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4012–4021, 2022. 2

[12] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 644–660. Springer, 2020. 2

[13] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. Densely constrained depth estimator for monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 718–734. Springer, 2022. 2

[14] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2800, 2022. 2

[15] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1070–1079, 2022. 2, 4

[16] Qing Lian, Yanbo Xu, Weilong Yao, Yingcong Chen, and Tong Zhang. Semi-supervised monocular 3d object detection by multi-view consistency. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 715–731. Springer, 2022. 1, 2, 3

[17] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1810–1818, 2022. 2

[18] 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

[19] 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

[20] 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. 2, 5

[21] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 311–327. Springer, 2020. 2

[22] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 2, 4, 5, 6, 7

[23] Xinzhu Ma, Yuan Meng, Yinmin Zhang, Lei Bai, Jun Hou, Shuai Yi, and Wanli Ouyang. An empirical study of pseudo-labeling for image-based 3d object detection. *arXiv preprint arXiv:2208.07137*, 2022. 3

[24] 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. In *European Confer-*

*ence on Computer Vision*, pages 123–139. Springer, 2022. 3

[25] 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

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[27] Robert Shapiro. Direct linear transformation method for three-dimensional cinematography. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, 49(2):197–205, 1978. 4

[28] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15172–15181, 2021. 2

[29] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 2

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 3

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[33] 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

[34] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 2

[35] 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, 5

[36] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 2

[37] Lei Yang, Xinyu Zhang, Li Wang, Minghan Zhu, Chuang Zhang, and Jun Li. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *arXiv preprint arXiv:2207.04448*, 2022. 1, 2

[38] 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. *arXiv preprint arXiv:1906.06310*, 2019. 2

[39] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers, 2023. 1

[40] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022. 6

[41] 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, 3, 4, 5

[42] Zhongfei Zhang and Allen R Hanson. 3d reconstruction based on homography mapping. *Proc. ARPA96*, pages 1007–1012, 1996. 3

[43] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2