Mix-Teaching: A Simple, Unified and Effective Semi-Supervised Learning Framework for Monocular 3D Object Detection

Lei Yang, Graduate Student Member, IEEE, Xinyu Zhang, Member, IEEE, Jun Li, Li Wang, Minghan Zhu, Graduate Student Member, IEEE, Chuang Zhang, and Huaping Liu, Senior Member, IEEE

Abstract—Semi-supervised learning (SSL) has promising potential for improving model performance using both labelled and unlabelled data. Since recovering 3D information from 2D images is an ill-posed problem, the current state-of-the-art methods of monocular 3D object detection (Mono3D) have relatively low precision and recall, making semi-supervised learning for Mono3D tasks challenging and understudied. In this work, we propose a unified and effective semi-supervised learning framework called Mix-Teaching that can be applied to most monocular 3D object detectors. Based on the idea of decomposition and recombination, unlabelled samples are firstly decomposed into collections of image patches with high-quality predictions and collections of background images containing no objects. The student model is then trained on the mixed images containing dense instances with high-quality pseudo-labels generated by the recombination operation. In addition, we propose an uncertainty-based filter to distinguish high-quality pseudo-labels from noisy predictions during the decomposition process. As results in KITTI and nuScenes benchmarks, Mix-Teaching consistently improves MonoFlex and GUPNet by significant margins under various labeling ratios. Our method achieves around +6.34% AP_{3D} improvement against the GUPNet on the validation set when using only 10% labelled data. Using the full training set and the additional 38K raw images from KITTI, it can further improve the MonoFlex by +4.65% absolute improvement on AP_{3D} for car detection, reaching 18.54% AP_{3D}, which ranks the 1st place among all monocular based methods on the KITTI test leaderboard.

Index Terms—Semi-supervised learning, 3D object detection, autonomous driving.

Manuscript received 26 October 2022; revised 14 March 2023 and 10 April 2023; accepted 19 April 2023. Date of publication 26 April 2023; date of current version 30 October 2023. This work was supported in part by the National High Technology Research and Development Program of China under Grant 2018YFE00204300, in part by the National Natural Science Foundation of China under Grant 62272198 and Grant 11964203, and in part by the China Post-Doctoral Science Foundation under Grant 2021M691780. This article was recommended by Associate Editor W. Wang. (Corresponding author: Xinyu Zhang.)

Lei Yang, Xinyu Zhang, Jun Li, Li Wang, and Chuang Zhang are with the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: yanglei20@mails.tsinghua.edu.cn; xzhang@tsinghua.edu.cn; lijun1958@tsinghua.edu.cn; wangli_thu@mail.tsinghua.edu.cn; zhch20@mails.tsinghua.edu.cn).

Minghan Zhu is with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: minghanz@umich.edu).

Huaping Liu is with the State Key Laboratory of Intelligent Technology and the Systems and the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: hpliu@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3270728.

Digital Object Identifier 10.1109/TCSVT.2023.3270728

I. INTRODUCTION

MONOCULAR 3D object detection (Mono3D) is the task of predicting the categories and 3D bounding boxes of surrounding objects from a single 2D image. It’s an indispensable component in autonomous driving, providing perceptual information for the downstream planning [1], [2] module. In recent years, many innovative detectors have emerged and achieved increasing accuracy. However, most of these methods are heavily dependent on labelled data. Compared with human-annotated images which are often expensive and time-consuming, raw images are easier to achieve large-scale collection. Thus, taking full advantage of both labelled and unlabelled data in model training is a promising approach to alleviate the heavy reliance on human annotations.

Semi-supervised learning (SSL) can help effectively improve the performance of fully-supervised baselines by employing both labelled and unlabelled data. In recent years, plentiful SSL methods for classification [3], [4], [5], [6], 2D object detection [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] and LiDAR-based 3D object detection [17], [18], [19], [20] have been proposed and applied. Research on semi-supervised learning for monocular 3D object detection has been preliminarily explored in [21], [22], and [23]. However, it is widely recognized that this field is still in its infancy, and there is a pressing need for further exploration of theoretical methods and continuous improvement in accuracy metrics.
Due to the challenge of recovering depth information from a single image, the accuracy of monocular 3D object detection is lagging significantly behind that of 2D object detection and LiDAR-based 3D object detection. Take KITTI [24] benchmark for example, the state-of-the-art methods for monocular 3D object detection just achieve less than 17% AP@0.7, while the candidates for the other two tasks have reached more than 85%-96% AP@0.7. This means that the pseudo-labels predicted by image-based 3D object detectors for unlabelled images possess extremely lower precision and recall. Lower precision signifies that more incorrect predictions are possible to be used as labels for unlabelled samples, which can lead to serious confirmation bias. On the other hand, lower recall explains that pseudo-labels fail to cover most objects, which will result in miss-detection. However, most existing SSL methods directly use the unlabelled images with pseudo-labels and the labelled data to train the student model as in STAC [15], which can’t effectively handle the issue of pseudo-labels with extremely low precision and recall. This is the main reason why most existing SSL frameworks are not applicable to monocular 3D object detection. To verify the above conclusion, we conducted extensive experiments following the above basic paradigm of semi-supervised learning. First, we use the teacher model to generate pseudo-labels for unlabelled images. Then, we used the raw images with pseudo-labels and the labelled data to train the student model as in STAC [15], which can’t effectively handle the issue of pseudo-labels with extremely low precision and recall. Therefore, it is indispensable to ensure high precision and meanwhile high recall pseudo-labels in semi-supervised training for monocular 3D object detection. However, following the traditional semi-supervised training technique, it is difficult to achieve pseudo-labels with high precision and recall simultaneously for a source image by only setting reasonable thresholds. To overcome these issues, we propose Mix-Teaching, a simple but effective semi-supervised learning framework for most monocular 3D object detectors. The results of our method with the same experimental setting are shown in Fig. 2. There are significant performance improvements over the baseline using different pseudo-label ratios, and the best performance is obtained when using 30% high-quality pseudo-labels.

One key challenge for semi-supervised monocular 3D object detection is the extremely low recall of pseudo-labels. In the proposed Mix-Teaching, we first predict pseudo-labels for unlabelled data by self-training. Unlabelled images are then split into image patches collection with high-quality pseudo-labels and a collection of background images containing no objects. Subsequently, the student model is trained on the mixed images that are created by merging the above instance image patches into empty backgrounds or human-labelled images through recombination operation. In this way, the generated images are full of instances with high-quality pseudo-labels while successfully avoiding the missing label cases, which is more effective for semi-supervised training. Finally, we adopt a multi-stage training scheme to progressively propagate information from the labelled to the unlabelled data.

Another key challenge for semi-supervised monocular 3D object detection is the extremely low recall of pseudo-labels. Considering the misalignment between confidence score and localization quality, it’s not exhaustive to eliminate incorrect labels using only a confidence-based filter. To this end, we further propose a model agnostic uncertainty-based filter to help remove noisy pseudo-labels. In this method, the predictions by the models with identical structures but different parameters are used to estimate the uncertainty of each object. For the prediction set belonging to the same object, the higher uncertainty, the fewer predictions in this set, and the larger localization misalignment among them. We build a formula to represent the localization uncertainty in a 3D object detection task. Based on both confidence-based and uncertainty-based filters, we manage to remove incorrect pseudo-labels more effectively in semi-supervised training and thus alleviate the confirmation bias. Because the process of removing noisy pseudo-labels is only carried out at the beginning of each training stage, the efficiency influence from uncertainty calculation is inappreciable.

We benchmark Mix-teaching with SSL setting using the full KITTI [24] object data and KITTI [24] raw data. When using MonoFlex [25] as backbone detector, Mix-Teaching achieves state-of-the-art results on the KITTI test leaderboard, which even surpasses the LPCG [26] method that directly relies on LiDAR-based 3D object detectors to generate pseudo-labels. Furthermore, we provide the SSL experiments under different labeling ratios, which can serve an important baseline for semi-supervised monocular 3D object detection.

Our contributions can be summarized as follows:

- First, we analyze the main difficulties in accomplishing semi-supervised learning for monocular 3D object...
detection and explain why most existing SSL approaches can’t handle these issues from the experimental perspective. Based on this analysis, we introduce Mix-Teaching, a simple, unified, and effective semi-supervised framework specialized for monocular 3D object detection.

- To alleviate the confirmation bias, we further propose a model agnostic uncertainty-based filter to help remove noisy pseudo-labels effectively.
- Extensive experiments on KITTI [24] datasets show that the MonoFlex [25] using our method outperforms SOTA methods by a large margin, ranking 1st on the KITTI test leaderboard at the time of first submission (car, Feb. 2022). In the nuScenes benchmark, our method achieves consistent accuracy improvements over the baseline when using 10% and 100% labelled data.

II. RELATED WORK

A. Monocular 3D Object Detection

A number of methods have been proposed for monocular 3D object detection. The core problem of these approaches is how to reconstruct spatial information more effectively. The pseudo-LiDAR-based methods [27], [28], [29], [30], [31] first transform the input image into dense artificial point clouds using existing depth estimation algorithms [32], [33], [34] and then apply LiDAR-based 3D object detectors [35], [36]. The geometry-based methods [37], [38], [39] derive depth information based on the 2D/3D geometry constraint of a specific reference. Another keypoint-based works [25], [40], [41], [42], [43], [44], [45], [46] directly estimate the 3D properties of the instance based on the high-dimensional features at the keypoint position.

B. Semi-Supervised Learning

Semi-supervised learning focuses on training models with both labelled and unlabelled data, which has achieved state-of-the-art performance on classification [3], [4], [5], [6], segmentation [47], [48], [49], [50], 2D object detection [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] and LiDAR-based 3D object detection [17], [18], [19], [20]. One popular type of SSL is consistency regularization, which constrains the outputs of different augmented inputs to be consistent. CSD [9] is a consistency-based method for 2D object detection. This approach ensures consistent predictions between input images and their flipped versions on both labelled and unlabelled data. SESS [18] is a semi-supervised learning framework for LiDAR-based 3D object detection. In order to improve the generalization ability of the model, this method applies three consistency losses to two sets of 3D proposals from teacher and student networks. The other types of SSL is pseudo-labeling, which is based on high-quality pseudo-labels and can be seen as a hard version of consistency regularisation. FixMatch [6] first generates pseudo-labels on weakly augmented unlabelled images and then trains the student model to predict the same classifications on strongly augmented data. Unbiased Teacher [51] addresses the problem of pseudo-label bias caused by the class imbalance in 2D annotations using EMA [52] training and focal loss [53]. 3DIoUMatch [17] achieves semi-supervised 3D object detection in the point cloud with a teacher-student mutual learning framework. To improve the quality of the pseudo-labels, all predictions that do not pass the thresholds for classification score, objectness confidence and 3D IoU are filtered out.

Previous studies [21], [22], [23] have explored semi-supervised learning for monocular 3D object detection, with methods such as KM3D [21] using a consistency regularization approach based on keypoint constraints, and MVC-MonoDet [23] employing both box-level and object-level regularizations to ensure consistency across unlabelled multi-view data. However, these existing methods primarily rely on consistency regularization, and there is a significant need for additional research into the use of pseudo-labeling and continuous improvement of accuracy metrics in this area.

III. METHOD

A. Problem Definition

The semi-supervised monocular 3D object detection includes \( n_l \) labelled data \( L^l = \{ (x^l_1, y^l_1), \ldots, (x^l_{n_l}, y^l_{n_l}) \} \) and \( n_u \) unlabelled data \( L^u = \{ (x^u_1, \ldots, x^u_{n_u}) \} \), where \( x \) represents the image data and \( y \) denotes the human-annotated labels that include the category and 3D bounding box labels associated with each object. We aim to significantly improve the performance of fully-supervised baselines by applying both labelled and unlabelled data in training.

B. Mix-Teaching Framework

The Mix-Teaching SSL framework is illustrated in Fig. 3. As indicated, this framework consists mainly of two stages, referred to as database-oriented pseudo-labeling and noisy student training. These components work together with the multi-stage training scheme described on the right side of Fig. 3, where the initial teacher model is trained on labelled data, followed by a pseudo-labeling process applied to the unlabelled data. A noisy student model is then trained using all the labelled and unlabelled images based on the decomposition and recombination technique. The resulting student model is then used as a new teacher model in the next round.

1) Database-Oriented Pseudo Labeling: This process seeks to make the most of sparsely distributed high-quality pseudo-labels in semi-supervised training by combining all the labels and background images together. Here, pseudo-labels are generated by applying a test-time inference of the teacher model with unlabelled images. The instance database, consisting of instance-level image patches and their corresponding high-quality pseudo-labels, is created by applying confidence-based and uncertainty-based filters. The background database is composed of unlabelled images without any predictions from the teacher model.

2) Noisy Student Training: The image patch and background databases are used in the noisy student with image recombination component along with labelled data to create mixed images with more intensive and precise labels in semi-supervised training. These mixed images are assembled based on two strategies. The first one pastes the image patches from the instance database onto the labelled images. The second
Mix-Teaching follows a multi-stage training scheme. There are two crucial processes in each round of training: Database-Oriented Pseudo-Labeling and Noisy Student Training.

**Database-Oriented Pseudo-Labeling:** Based on the pseudo-labels generated by applying the teacher model to unlabelled images, we create two databases: one consists of images without objects, and the other consists of image patches with high-quality pseudo-labels.

**Noisy Student Training:** The student model is trained on the mixed images, which have much more intense and precise labels, by merging image patches from the above instance database with labelled images or those from the background database.

The student model is trained using a batch of mixed images with the corresponding human-annotated labels and pseudo-labels by jointly minimizing the total loss \( L \) formed of the supervised loss \( L_s \) and unsupervised loss \( L_u \) as follows:

\[
L = L_s + \lambda \times L_u,
\]

where a hyper-parameter \( \lambda \) with a value in the range of 0-1 is applied to balance the extent of \( L_u \) relative to that of \( L_s \) considered in the total loss.

The supervised loss \( L_s \) consists of a classification loss \( L_{cls} \) and a regression loss \( L_{reg} \). It can be calculated as:

\[
L_s = \sum_{L}^1 \frac{1}{N_l} \sum_{i}^1 (L_{cls}(b_i^l) + L_{reg}(b_i^l)),
\]

where \( L \) denotes the index of labelled images in a batch, \( N_l \) represents the number of human annotations for each image, \( b_i^l \) is the \( i \)-th label in the \( L \)-th labelled image.

The unsupervised loss \( L_u \) is computed on pseudo-labels and can be written as:

\[
L_u = \sum_{L}^1 \frac{1}{N_u} \sum_{i}^1 (L_{cls}(b_i^u) + L_{reg}(b_i^u))
\]

\[
+ \sum_{B}^1 \frac{1}{N_u} \sum_{i}^1 (L_{cls}(b_i^u) + L_{reg}(b_i^u)),
\]

where \( B \) indicates the index of background images, \( N_u \) is the number of pseudo-labels on each image, \( b_i^u \) represents the \( i \)-th pseudo label on a labelled or a background image.

The decomposition and recombination methodology discussed above creates new mixed images composed of all positive instances that have been collected and merged into background images. Accordingly, the positive instances are object-level image patches with high-quality pseudo-labels, while the backgrounds denote empty unlabelled images or labelled data. Therefore, the new merged images will possess both high recall and relatively few false positives, effectively addressing the problems of extremely low recall and confirmation bias in semi-supervised mono3D operations.
models; (2) the spatial discrepancy between the predictions. Visualization of box-level strong augmentations. Fig. 4.

6836 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 33, NO. 11, NOVEMBER 2023

C. Uncertainty-Based Filter

We plot the relationship between the 3D IoU of object bounding boxes predicted by MonoFlex [25] with ground truth boxes and the classification score in Fig. 5(a), while the relationship between those 3D IoU values and the localization uncertainty defined below is plotted in Fig. 5(b). As can be seen in Fig. 5(a), the localization precision of bounding boxes predicted by MonoFlex [25] with ground truth boxes is strongly misaligned with the classification score, where a considerable proportion of predictions have a high confidence classification score but low 3D IoU with the ground truth boxes. As discussed above, this leads to the strong confirmation bias that is addressed herein by removing imprecise labels in the semi-supervised training process using a high confidence classification score but low 3D IoU with the ground truth boxes. As discussed above, this leads to the strong confirmation bias that is addressed herein by removing imprecise labels in the semi-supervised training process using both a standard confidence-based filter and an uncertainty-based filter.

We define the localization uncertainty mainly based on the M predictions obtained for a given object in an image given to N isomorphic teacher models with different parameters. Accordingly, the localization uncertainty can be defined from two perspectives: (1) the value of M associated with an object, since M reflects the level of missed detection among N models; (2) the spatial discrepancy between the M predicted bounding boxes, which reveals the randomness in the model predictions.

As depicted in algorithm 1, the localization uncertainty is calculated according to the following steps:

1) All predictions from N teacher models are stored in list B. list S contains their corresponding confidence scores.
2) Declare three lists G, H and U. G is used to store box clusters. Each cluster represents the predictions for a certain object from teacher models. H is for the box with the highest confidence score in each cluster. U saves the localization uncertainty for each box in list H.
3) Iterate through all the boxes in list B to find the matching box that belongs to the current cluster C. The matching condition is defined as a box with a large overlap with the initial box \( b_m \) of cluster C under the condition \( IoU3D > thr \). All matching boxes will be moved from list B to cluster C. And then update the current cluster C to list G.
4) If there are still unprocessed boxes in list B, select the box \( b_m \) that has the maximum score in list B and move it to list H. Initialize a new cluster C with box \( b_m \) and proceed to step 3.
5) When all boxes in B are processed, calculate the uncertainty \( u \) for each box cluster C in list G with the

Fig. 4. Visualization of box-level strong augmentations. From left to right: original image patch, border cut, color padding, mixup and the fusion of the previous three methods.

Algorithm 1 Calculate Uncertainty

\begin{verbatim}
Input:
B = \{b_1, \ldots, b_N\}, S = \{s_1, \ldots, s_N\}
B is the list of all box candidates.
S contains the corresponding classification scores.
thr is the 3D IoU threshold of matching condition.

Begin:
1: H \leftarrow \emptyset; G \leftarrow \emptyset; U \leftarrow \emptyset
2: while B \neq empty do
3: \quad m \leftarrow \text{argmax}(S); H \leftarrow H \cup \{b_m\};
4: \quad B \leftarrow B - b_m; S \leftarrow S - s_m
5: \quad for b_i in B do
6: \quad \quad if max(IoU3D(b_i, b_m)) > thr then
7: \quad \quad \quad C \leftarrow C \cup b_i; B \leftarrow B - b_i; S \leftarrow S - s_i
8: \quad \quad end if
9: \quad G \leftarrow G \cup C
10: \quad end while
11: for C in G do
12: \quad u = \text{uncertain}(C)
13: \quad U \leftarrow U \cup u
14: \quad end for
15: return H, U
End
\end{verbatim}
Fig. 5. The comparison of localization uncertainty and confidence score in evaluating localization precision. (a) the relationship between the confidence of detector’s prediction and the 3D IoU with ground truth box. (b) the relationship between the 3D IoU and the localization uncertainty confidence of detector’s prediction and the 3D IoU with ground truth box.

The uncertainty $u$ ranges from 0 to 1. When the value is 0, it indicates that there exists no miss-detection in $N$ models ($M = N$), and all $M$ box candidates are perfectly consistent. When the value is 1, it means that all models fail to detect this object.

The relationship between the 3D IoU of object bounding boxes predicted by MonoFlex with ground truth boxes and the localization uncertainty defined herein is presented in Fig. 5(b). Compared with the classification score (Fig. 5(a)), the obtained uncertainty is clearly a better measure of localization accuracy. From the perspective of each object, as shown in Fig. 6, both miss-detection and the poor consistency of box candidates will lead to high uncertainty. In these cases, predictions from teacher models always possess lower localization accuracy.

Compared with 3D confidence [54] or 3D IoU [17], [21], which require a corresponding branch design for specific detectors, our uncertainty-based filter is model-independent and can be applied to many types of image-based 3D object detectors, which is more appropriate for the proposed general semi-supervised learning framework.

IV. EXPERIMENTS

A. Dataset and Metrics

1) KITTI: [24] dataset provides 15K frames of labelled image data and 38K unlabelled image data, which is appropriate for semi-supervised learning research that relies on limited labelled data and massive unlabelled data. Therefore, we evaluate our Mix-teaching on the challenging KITTI [24] dataset. KITTI contains 7,481 images for training and 7,518 images for testing. However, since we have no access to the manual annotations of the testing set, the training set is further divided into 3,712 training samples and 3,769 validation samples for local evaluation as mentioned in a previous study [55]. Referring to the data split in [26], we further collect the KITTI raw scenes consisting of 156 sequences from the website (https://www.cvlibs.net/datasets/kitti/raw_data.php). Then, the samples that appear in KITTI 3D validation and test set in continuous sequences are excluded to avoid data leakage. The remaining 38K samples are used as unlabelled data for semi-supervised training. We use the average precisions(AP) for 3D and bird’s eye view object detection as the metrics. All evaluation results obtained for the validation and testing datasets are based on a 40-point interpolated average precision (AP40) rather than the original 11-point interpolated average precision, according to a previously published scheme [56]. We report the detection results with three level difficulties, i.e. easy, moderate and hard, in which the moderate scores are normally for ranking. We conduct the experimental evaluation for Car category unless otherwise indicated.

2) nuScenes: [57] dataset is designed to evaluate autonomous driving algorithms and contains data from 6 cameras, 1 LiDAR, and 5 radars. It includes 1000 scenarios, with 700 for training, 150 for validation, and 150 for testing. The dataset includes 1.4 million annotated 3D bounding boxes for 10 different object classes, and the evaluation of the detection task is based on mean average precision (mAP) across four different thresholds that are based on center distance on the ground plane. The performance of the algorithm is evaluated using five true-positive metrics, including ATE for measuring translation error, ASE for scale error, AOE for orientation error, AVE for velocity error, and AAEP for attribute error. In addition to these metrics, the nuScenes detection score (NDS) is defined, which combines detection accuracy (mAP) with the five true-positive metrics. The nuScenes dataset has a camera frame rate of 12Hz. Image keyframes synchronized with LiDAR and RADAR are sent to annotation partners for annotation, which is used as labelled data. The remaining 10Hz raw images, which are five times the labelled data, are used as unlabelled data for semi-supervised learning.

B. Implement Details

We adopt the MonoFlex [25] and GUPNet [42] as two baseline detectors. The proposed Mix-Teaching framework is
implemented with $N = 5$ and a minimum $t_{3D_{IoU}}$ threshold of 0.7, and only predictions with a confidence score greater than a threshold of 0.7 and localization uncertainty values less than a threshold of 0.25 are added to the instance database during the pseudo-labeling process. The images with no detections are collected to build a background database. During the training phase of the student model, the student is initialized with the previous teacher model. The images from the background database are selected with a chance of 40% apart from the labelled images. For strong data augmentation at the instance level, we apply mixup augmentations on each instance, patch, border cut, and color padding augmentations are applied randomly with a probability of 50%. We set the hyperparameter $\lambda = 1.0$, $\beta = 1.0$. Following the multi-training scheme, we run 4 rounds of semi-supervised training for all experiments unless otherwise specified, where the student model at the end of one round becomes the teacher model in the next round.

C. Quantitative Results

1) Results on KITTI Validation Set: We make a detailed comparison with the supervised baselines, including GUPNet [42] and MonoFlex [25], under different training set ratios. All 38K raw images of KITTI are used as unlabelled data for semi-supervised training. As shown in Table I, the $\text{AP}_{3D}/\text{AP}_{BEV} (\text{IoU} = 0.7)/R_{90}$ results obtained for the car category in the KITTI validation dataset, the Mix-Teaching framework uniformly improves the performance of the 3D object detection algorithms significantly under all conditions. When using only 10% labelled data, our approach achieves about +6.34% and +5.98% $\text{AP}_{3D}$ improvements at a moderate level over the MonoFlex [25] and GUPNet [42] baselines. The results indicate that the proposed framework is able to learn from unlabelled data, which is particularly evident when the number of labelled data is small. Furthermore, it is worth noting that when using the full training set, our Mix-Teaching is still able to significantly outperform the upper-bound performance of the two baselines.

2) Results on KITTI Test Set: We evaluate the proposed Mix-Teaching on the KITTI test set using MonoFlex [25] and GUPNet [42] as two base monocular detectors. Table II shows the quantitative results of our method and other top-performing detectors from the official KITTI leaderboard, with the highest performance in bold and the next highest in underlined. As can be seen, applying Mix-Teaching with the GUPNet [42] and MonoFlex [25] 3D object detectors achieves superior $\text{AP}_{3D}/\text{AP}_{BEV}$ values in nearly all cases. For instance, the proposed method improves the $\text{AP}_{3D}$ of GUPNet [42] by $+7.44+/3.62+/2.95$ absolute improvements under easy/moderate/hard setting. Meanwhile, our approach increases the same metric of MonoFlex [25] from 19.94/13.89/12.07 to 26.89/18.54/15.79, which is absolutely remarkable. Moreover, the performance of the proposed Mix-Teaching SSL framework even surpasses that of the LPCG SSL framework [26] that directly relies on LiDAR-based 3D object detectors to help generate pseudo-labels using the same MonoFlex [25] baseline. We rank the 1st place according to $\text{AP}_{3D}$ on a moderate setting (same as KITTI leaderboard).

D. Ablation Studies

In this section, we perform ablation studies to investigate the effects of each element. We use MonoFlex [25] as the base detector. The training of ablation experiments is conducted on the full KITTI training set. The results for car category are evaluated on the corresponding validation set.

1) The Scale of Unlabelled Data: The effects of the scale of unlabelled data usage on the detection performance of the Mix-Teaching implementation are presented in Table V, where 0% corresponds to MonoFlex [25] without Mix-teaching, 50% or 100% means that 19K or 38K unlabelled image data is used in the Mix-Teaching process. Compared with the results...
### Table II

**Performance of the Car Category on KITTI Test Set.** We use **bold** to highlight the highest results and **underlined** for the second-highest ones. † represents the baseline we employed. * means using the extra point cloud. All methods are ranked by AP$_{3D}$ on a moderate setting (same as KITTI leaderboard). Our method outperforms the baseline by a large margin and achieves the best performance.

| Method          | Reference  | GPU       | Runtime (ms) | AP$_{3D}$ (IoU=0.7)/$R_{40}$ Easy | AP$_{BEV}$ (IoU=0.7)/$R_{40}$ Easy |
|-----------------|------------|-----------|--------------|-----------------------------------|-----------------------------------|
| MonoGRNet [58]  | TPAMI 2021 | Tesla P40 | 60           | 9.61 / 5.74 / 4.25               | 18.19 / 11.17 / 8.73             |
| MonoPair [43]   | CVPR 2020  | Tesla V100| 60           | 13.04 / 9.99 / 8.65              | 19.28 / 14.83 / 12.89            |
| RTM3D [59]      | ECCV 2020  | 1080Ti    | 50           | 14.41 / 10.34 / 8.77             | 19.17 / 14.20 / 11.99            |
| KM3D [21]       | RAL 2021   | 1080Ti    | 50           | 16.73 / 11.45 / 9.92             | 23.44 / 16.20 / 14.47            |
| D2LCN [31]      | CVPR 2020  | 1080Ti    | 200          | 16.65 / 11.72 / 9.51             | 22.51 / 16.02 / 12.55            |
| Monolde [44]    | CVPR 2021  | 1080Ti    | 40           | 17.23 / 12.26 / 10.29            | 24.79 / 18.89 / 16.00            |
| Monorun [45]    | CVPR 2021  | 1080Ti    | 70           | 19.65 / 12.30 / 10.58            | 27.94 / 17.34 / 15.24            |
| GroMeD-NMS [60] | CVPR 2021  | Tikn X    | 120          | 18.10 / 12.32 / 9.65             | 26.19 / 18.27 / 14.05            |
| MonorCNN [61]   | ICCV 2021  | -         | 70           | 18.36 / 12.65 / 10.03            | 25.48 / 18.11 / 14.10            |
| DDMP-3D [62]*   | CVPR 2021  | -         | 180          | 19.71 / 12.78 / 9.80             | 28.08 / 17.89 / 13.44            |
| Ground-Aware [63]| RAL 2021   | 1080Ti    | 50           | 21.65 / 13.25 / 9.91             | 29.81 / 17.98 / 13.08            |
| PCT [64]        | NIPS 2021  | -         | 487          | 21.00 / 13.37 / 11.31            | 29.65 / 19.03 / 15.92            |
| CaDDN [65]      | CVPR2021   | 2080Ti    | 485          | 19.17 / 13.41 / 11.46            | 27.94 / 18.91 / 17.19            |
| DFR-Net [66]*   | ICCV 2021  | -         | 180          | 19.40 / 13.63 / 10.35            | 28.17 / 19.17 / 14.84            |
| MonoBIF [67]    | TPAMI 2021 | -         | 30           | 21.29 / 13.87 / 11.71            | 29.03 / 19.70 / 17.26            |
| AutoShape [68]  | ICCV 2021  | 2080Ti    | 52           | 22.47 / 14.17 / 11.36            | 30.66 / 20.08 / 15.95            |
| Monadetr [69]   | CVPR 2022  | -         | 37           | 21.99 / 15.39 / 12.73            | 28.59 / 20.38 / 17.14            |
| Monodetr [70]   | Arxiv 2022 | -         | 40           | 23.65 / 15.92 / 12.99            | 32.08 / 21.44 / 17.85            |
| Monodistill [71]*| ICLR 2022  | 1080 Ti   | 40           | 22.97 / 16.03 / 13.60            | 31.87 / 22.59 / 19.72            |
| MonoISG [72]    | CVPR 2022  | -         | 42           | 24.69 / 16.14 / 13.64            | 32.59 / 21.26 / 18.18            |
| DDS3D [73]      | ECCV 2022  | 2080Ti    | 60           | 23.19 / 16.87 / 14.36            | 32.35 / 23.41 / 20.42            |
| MVC-MonoDet [73]| ECCV 2022  | 2080Ti    | 33           | 25.03 / 16.89 / 14.83            | -                                 | -                                 |
| LPCG [56]*      | ECCV 2022  | 2080Ti    | 30           | 25.56 / 17.80 / 15.38            | 35.96 / 24.81 / 21.86            |

### Table III

**Quantitative Results for Pedestrian and Cyclist on KITTI Test Set.** "Rel. Imp." represents relative improvements.

| Method          | Cat. | AP$_{3D}$/$R_{40}$ | AP$_{BEV}$/$R_{40}$ |
|-----------------|------|---------------------|---------------------|
|                | Easy | Mod. | Hard | Easy | Mod. | Hard |
| MonoFlex        | Ped. | 9.43 | 6.31 | 5.26 | 10.36 | 7.36 | 6.29 |
| Our.            |      | 11.67 | 7.47 | 6.81 | 12.34 | 8.40 | 7.06 |
| Rel. Imp. (%)   |      | 23.75↑ | 18.85↑ | 25.67↑ | 23.75↑ | 18.85↑ | 25.67↑ |
| MonoFlex        | Cycl. | 4.17 | 2.35 | 2.04 | 4.41 | 2.67 | 2.50 |
| Our.            |      | 8.04 | 4.91 | 4.15 | 8.56 | 5.36 | 4.62 |
| Rel. Imp. (%)   |      | 92.81↑ | 108.94↑ | 103.43↑ | 94.10↑ | 100.75↑ | 84.80↑ |

### Table IV

**Quantitative Results of GUPNet and Its Improved Models With Our Methods on the nuScenes Validation Set**

| Method          | mAP↑ | NDS↑ |
|-----------------|------|------|
| GUPNet          | 0.143| 0.164|
| GUPNet + Ours   | 0.184| 0.213|
| Abs. Imp.       | +4.1%| +4.9%|

of 19K KITTI raw data, the experiment when using the whole 38K unlabelled data can further improve the AP$_{3D}$ from 20.61% to 22.27%, 8.63% to 9.46% and 3.21% to 4.94% for car, pedestrian, and cyclist respectively. This means that the values of AP$_{3D}$/$R_{40}$ and AP$_{BEV}$/$R_{40}$ uniformly increase monotonically as the scale of unlabelled data increases.

2) **Background Database:** Next, we investigate whether the backbone database is necessary during the student model training period. As shown in Table V, when using half of the raw KITTI data, we gain +2.32%, +1.17% and +0.90% absolute improvement over the version without background database on car, pedestrian, and cyclist respectively. And when it comes to all the unlabelled data conditions, the background database brings the additional +1.66%, +0.83% and +1.73% improvements on the same three categories as well. The background database is indispensable in semi-supervised training.

3) **Box-Level Data Augmentations:** We ablate the effects of box-level data augmentations. As shown in Table VI, M (mixup) is applied to each instance patch, B (border cut) and C (color padding) are applied randomly with a probability of 50%, it turns out that all these three strategies are helpful in improving performance. The combination of the above three data augmentations can further improve the performance of Mix-Teaching.

4) **Confidence-Based and Uncertainty-Based Filters:** As shown in Table VII, applying only the localization uncertainty
as a filter can improve prediction performance better than only applying the confidence score as a filter, which explains that our proposed uncertainty-based filter is much more effective. The best results are obtained by using both filters.

5) The Thresholds for Confidence-Based and Uncertainty-Based Filters: We investigated the effect of different confidence and uncertainty thresholds on the ability to discriminate high-quality pseudo-labels. As shown in Table VIII, the results clearly show that the best prediction performance is achieved by using a confidence threshold of 0.7 and a localization uncertainty threshold of 0.25.

6) The Existing Depth Variance-Based Filter: We conduct comparative experiments on the effect of our uncertainty estimation technique and the existing variance of the estimated depth in MonoFlex and GUPNet. As shown in Table IX, under the condition that the confidence-based filter is enabled, the existing depth variance-based filter achieves only 0.37% improvement in accuracy for MonoFlex and 0.17% for GUPNet in terms of accuracy, which is relatively insignificant. This is due to the fact that the variance of the estimated depth is not a reliable measure of the confidence of the prediction.

**Fig. 7.** Ablation study on the number of semi-supervised training rounds. ‘Rnd.’ means Round.

6) The Existing Depth Variance-Based Filter: We conduct comparative experiments on the effect of our uncertainty estimation technique and the existing variance of the estimated depth in MonoFlex and GUPNet. As shown in Table IX, under the condition that the confidence-based filter is enabled, the existing depth variance-based filter achieves only 0.37% improvement in accuracy for MonoFlex and 0.17% for GUPNet in terms of accuracy, which is relatively insignificant. This is due to the fact that the variance of the estimated depth is not a reliable measure of the confidence of the prediction.
Fig. 8. Qualitative results on the KITTI val set. We present four pairs of comparisons marked with capital letters from A to D. Each pair consists of four pictures, the upper left displays the predictions of MonoFlex [25] baseline (blue), the lower left is its representation in the bird’s-eye view. The upper right shows the results of our Mix-Teaching (green), and the lower right is its bird’s-eye view display. The red boxes in the bird’s eye view represent ground truths. We use dashed ovals to highlight the pronounced difference in the predictions.
depth and the classification score is already included in the confidence-based filter for the MonoFlex and GUPNet baselines. In particular, when our proposed uncertainty-based filter is enabled, significant accuracy gains of up to 2% are achieved, indicating the strength of our designs. In addition, other small accuracy improvements are achieved when we apply both our uncertainty-based filter and the depth variance-based filter.

7) The Number of Semi-Supervised Training Rounds: Following the multi-stage training scheme, the number of semi-supervised training rounds is an important hyperparameter. We perform ablation studies on this parameter, as shown in Fig. 7, there are significant accuracy improvements in the first four training rounds, the average precision gradually grows slowly in the later 5 or 6 rounds. The detailed results of the experiments using 100% labelled data can be seen in Table X.

E. Qualitative Analysis

The 3D bounding box predictions obtained by MonoFlex [25] alone and Mix-Teaching are illustrated qualitatively according to the images A-D given in Fig. 8, where the MonoFlex [25] predictions are given in blue, the Mix-Teaching predictions are given in green, and the corresponding BEV representations are given below each image along with the ground truth objects given in red. Here, pronounced differences between the predicted and ground truth objects are highlighted by the dashed ovals applied within the BEV representations. These results very clearly demonstrate that the proposed Mix-Teaching framework can significantly improve the prediction performance of MonoFlex [25] in a variety of street scenes, including those with pedestrian and cyclist objects (image A), over-occluded objects (image C), and car objects at close range (images B and C) and those at long range (images D).

V. CONCLUSION AND FUTURE WORK

In this paper, we propose Mix-Teaching, a simple, unified, and effective semi-supervised learning framework for monocular 3D object detection. Our method first generates pseudo-labels for unlabelled data through self-training. Then, according to the decomposition and recombination technique, we break the limitation of the original images and create new diverse and label-rich mixed images for semi-supervised training, which can effectively handle the problems caused by the extremely low precision and recall of the initial pseudo-labels. With the proposed uncertainty-based filter, we are able to effectively filter out poorly positioned pseudo-labels, resulting in less noise to mitigate confirmation bias. Experiments on the KITTI and nuScenes datasets show that Mix-Teaching improves the baseline model by a large margin under different labeling ratios. More importantly, using the 100% training set and MonoFlex as the baseline, we successfully rank first among all monocular 3D object detectors on the KITTI test leaderboard. In this way, we can continuously improve monocular 3D object detectors by collecting more unlabelled images, which has great economic significance in autonomous driving. In addition, the proposed Mix-Teaching follows the multi-stage training scheme. It will be left to future work to adopt the end-to-end training mode.

REFERENCES

[1] H. Wang, Y. Huang, A. Khajepour, D. Cao, and C. Ly, “Ethical decision-making platform in autonomous vehicles with lexigraphic optimization based model predictive controller,” IEEE Trans. Veh. Technol., vol. 69, no. 8, pp. 8164–8175, Aug. 2020.
[2] H. Wang, Y. Huang, A. Soltani, A. Khajepour, and D. Cao, “Cyber-physical predictive energy management for through-the-road hybrid vehicles,” IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 3246–3256, Apr. 2019.
[3] B. Zhang et al., “FlexMatch: Boosting semi-supervised learning with curriculum pseudo labels,” in Proc. NeurIPS, 2021, pp. 18408–18419.
[4] Q. Xie, M. T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves ImageNet classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10684–10695.
[5] D. Berthelot et al., “RemixMatch: Semi-supervised learning with distribution matching and augmentation anchoring,” in Proc. ICLR, 2020, pp. 1-13.
[6] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in Proc. NeurIPS, 2020, pp. 596-608.
[7] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, “Instant-teaching: An end-to-end semi-supervised object detection framework,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4081-4090.
[8] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, “Humble teachers teach better students for semi-supervised object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3132-3141.
[9] J. Jeong, S. Lee, J. Kim, and N. Kwak, “Consistency-based semi-supervised learning for object detection,” in Proc. NeurIPS, 2019, pp. 32-44.
[10] J. Jeong, V. Verma, M. Hyun, J. Kannala, and N. Kwak, “Interpolation-based semi-supervised learning for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11602–11611.
[11] Z. Wang, Y. Li, Y. Guo, L. Fang, and S. Wang, “Data-uncertainty guided multi-phase learning for semi-supervised object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4568–4577.
[12] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang, “Interactive self-training with mean teachers for semi-supervised object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5941–5950.
[13] M. Xu et al., “End-to-end semi-supervised object detection with soft teacher,” 2021, arXiv:2106.09018.
[14] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, “Rethinking pseudo labels for semi-supervised object detection,” 2021, arXiv:2106.00168.
[15] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, “A simple semi-supervised learning framework for object detection,” 2020, arXiv:2005.04757.
[16] F. Zhang, T. Pan, and B. Wang, “Semi-supervised object detection with adaptive class-rebalancing training,” 2021, arXiv:2107.05011.
[17] H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas, “3DIoU-Match: Leveraging IoU prediction for semi-supervised 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14615–14624.
[18] N. Zhao, T.-S. Chua, and G. H. Lee, “SESS: Self-ensembling semi-supervised 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11076–11084.
[19] J. Yin et al., “Semi-supervised 3D object detection with proficient teachers,” in Proc. ECCV, 2022, pp. 727–743.
[20] J. Yin et al., “ProposalContrast: Unsupervised pre-training for LiDAR-based 3D object detection,” in Proc. ECCV, 2022, pp. 17–33.
[21] P. Li and H. Zhao, “Monocular 3D detection with geometric constraint embedding and semi-supervised training,” IEEE Robot. Autom. Lett., vol. 6, no. 3, pp. 5565–5572, Jul. 2021.
[22] C. Lian, B. Ye, R. Xu, W. Yao, and T. Zhang, “Exploring geometric consistency for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 1685–1694.
[23] Q. Lian, Y. Xu, W. Yao, Y. Chen, and T. Zhang, “Semi-supervised monocular 3D object detection by multi-view consistency,” in Proc. ECCV, 2022, pp. 715–731.
[24] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 3354–3361.
[25] Y. Zhang, J. Lu, and J. Zhou, “Objects are different: Flexible monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3288–3297.
[26] L. Peng et al., “LiDAR point cloud guided monocular 3D object detection,” 2021, arXiv:2104.09035.

[27] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, “Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving,” in Proc. CVPR, 2019, pp. 8445–8453.

[28] Y. You et al., “Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving,” in Proc. ICLR, 2020.

[29] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, “Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6850–6859.

[30] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, “Rethinking pseudo-LiDAR representation,” in Proc. ECCV, 2020, pp. 311–327.

[31] M. Ding et al., “Learning depth-guided convolutions for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4306–4315.

[32] R. Díaz and A. Marathe, “Soft labels for ordinal regression,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4733–4742.

[33] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, “Vip-DeepLab: Learning visual perception with depth-aware video panoptic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3997–4008.

[34] S. Chen, Z. Pu, X. Fan, and B. Zou, “Fixing defect of photometric loss for self-supervised monocular depth estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1328–1338, Mar. 2022.

[35] L. Zhao, J. Guo, D. Xu, and L. Sheng, “Transformer3D-Det: Improving 3D object detection by vote refinement,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 12, pp. 4735–4746, Dec. 2021.

[36] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, “Temporal-channel transformer for 3D LiDAR-based video object detection for autonomous driving,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 4, pp. 2068–2078, Apr. 2022.

[37] A. Mousavilian, D. Anguelov, J. Flynn, and J. Kosecka, “3D bounding box estimation using deep learning and geometry,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 4306–4315.

[38] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, “GSSD: An efficient 3D object detection framework for autonomous driving,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 12, pp. 4747–4758, Dec. 2021.

[39] X. Zhang, Z. Li, X. Gao, D. Jin, and J. Li, “Channel attention in LiDAR-camera fusion for lane line detection,” Pattern Recognit., vol. 118, Oct. 2021, Art. no. 108020.

[40] Z. Zou, X. Zhang, H. Liu, Z. Li, A. Hussain, and J. Li, “A novel multimodal fusion network based on a joint-coding model for lane line segmentation,” Inf. Fusion, vol. 80, pp. 167–178, Apr. 2022.

[41] Y.-C. Liu et al., “Unbiased teacher for semi-supervised object detection,” in Proc. ICLR, 2021.

[42] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. NeurIPS, 2017.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

[44] A. Simonelli, S. R. Bulo, L. Porzi, P. Kontschieder, and E. Ricci, “Are we missing confidence in pseudo-LiDAR methods for monocular 3D object detection?” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 3205–3215.

[45] X. Chen et al., “3D object proposals for accurate object class detection,” in Proc. NeurIPS, 2015, pp. 29–42.

[46] A. Simonelli, S. R. Bulo, L. Porzi, M. Lopez-Antequera, and P. Kontschieder, “Disentangling monocular 3D object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1999–2009.

[47] H. Caeser et al., “NuScenes: A multimodal dataset for autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11621–11631.

[48] Z. Qin, J. Wang, and Y. Lu, “MonoGRNet: A general framework for monocular 3D object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5170–5184, Sep. 2022.

[49] P. Li, H. Zhao, P. Liu, and F. Cao, “RTM3D: Real-time monocular 3D detection from object keypoints for autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 644–650.

[50] A. Kumar, G. Brazil, and X. Liu, “GruoMeD-NMS: Grouped mathematically differentiable NMS for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8969–8979.

[51] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, “Geometry-based distance decomposition for monocular 3D object detection,” 2021, arXiv:2109.13005.

[52] L. Wang et al., “Depth-conditioned dynamic message propagation for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 454–463.

[53] Y. Liu, Y. Yixuan, and M. Liu, “Ground-aware monocular 3D object detection for autonomous driving,” IEEE Robot. Autom. Lett., vol. 6, no. 2, pp. 919–926, Apr. 2021.

[54] L. Wang et al., “Progressive coordinate transforms for monocular 3D object detection,” in Proc. NeurIPS, 2021, pp. 13364–13377.

[55] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, “Categorical depth distribution network for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5170–5180.

[56] Z. Zou et al., “The devil is in the task: Exploiting reciprocal geometric cost volume for monocular 3D object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 2693–2702.

[57] Z. Zhou, Y. He, H. Zhu, C. Wang, H. Li, and Q. Jiang, “Monocular 3D object detection: An extrinsic parameter free approach,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 7352–7362.

[58] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, “AutoShape: Real-time shape-aware monocular 3D object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15641–15650.

[59] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, “MonoDTR: Monocular 3D object detection with depth-aware transformer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 4012–4021.

[60] R. Zhang et al., “Monocr: Depth-aware transformer for monocular 3D object detection,” 2022, arXiv:2203.13310.

[61] Z. Chong et al., “MonoDistill: Learning spatial features for monocular 3D object detection,” in Proc. ICLR, 2022.

[62] Q. Lian, P. Li, and X. Chen, “MonoJS: Joint semantic and geometric cost volume for monocular 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 4012–4021.

[63] D. Park, R. Ambrus, V. Guizilini, L. Li, and A. Gaidon, “Is pseudo-LiDAR needed for monocular 3D object detection?” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 3142–3152.
Lei Yang (Graduate Student Member, IEEE) was born in Datong, Shanxi, China, in 1993. He received the B.E. degree from the Taiyuan University of Technology, Taiyuan, China, and the M.S. degree from the Robotics Institute, Beihang University, in 2018. He is currently pursuing the Ph.D. degree with the School of Vehicle and Mobility, Tsinghua University.

Then, he joined the Autonomous Driving Research and Development Department, JD.COM, as an Algorithm Researcher, from 2018 to 2020. His current research interests include computer vision, 3D scene understanding, and autonomous driving.

Xinyu Zhang (Member, IEEE) was born in Huining, Gansu. He received the B.E. degree from the School of Vehicle and Mobility, Tsinghua University, in 2001.

He was a Visiting Scholar with the University of Cambridge. He is currently a Researcher with the School of Vehicle and Mobility and the Head of the Mengshi Intelligent Vehicle Team, Tsinghua University. He is the author of more than 30 SCI/EI articles. His research interests include intelligent driving and multimodal information fusion.

Jun Li was born in Jilin, China, in 1958. He received the Ph.D. degree in internal-combustion engineering from the Jilin University of Technology in 1989.

He has joined the China FAW Group Corporation in 1989 and a Professor with the School of Vehicle and Mobility, Tsinghua University. He is currently the Chairperson of the China Society of Automotive Engineers (SAE). In these years, he has presided over the product development and technological innovation of large-scale automobile companies in China. He has many scientific research achievements in the fields of automotive powertrains, new energy vehicles, and intelligent connected vehicles. He is the author of more than 98 papers.

Dr. Li was awarded as an Academician of Chinese Academy of Engineering (CAE) for contributions to vehicle engineering in 2013.

Li Wang was born in Shangqiu, Henan, China, in 1990. He received the Ph.D. degree in mechatronic engineering from the State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, in 2020.

He was a Visiting Scholar with the Nanyang Technology of University for two years. Currently, he is a Post-Doctoral Fellow with the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University. He is the author of more than 15 SCI/EI articles. His research interests include autonomous driving perception, 3D robot vision, and multi-modal fusion.

Chuang Zhang received the B.E. degree from Southeast University, Nanjing, China, in 2017, and the M.S. degree from Beihang University, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree in mechanical engineering with the School of Vehicle and Mobility, Tsinghua University. His research interests include object detection, multi-object tracking, and multi-sensor fusion.

Minghan Zhu (Graduate Student Member, IEEE) received the B.E. degree in automotive engineering from the Department of Automotive Engineering, Tsinghua University, in 2016. He is currently pursuing the Ph.D. degree in mechanical engineering with the University of Michigan, Ann Arbor. His current research interests include vision-based 3D scene understanding and SLAM with the emphasis on the application to automated driving.

Huaping Liu (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2004. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. His research interests include robot perception and learning. He was a Senior Program Committee Member of the International Joint Conference on Artificial Intelligence in 2018. He was a recipient of the Andy Chi Best Paper Award in 2017. He was the Area Chair for Robotics Science and Systems in 2018.