

Semi-Supervised Stereo-based 3D Object Detection via Cross-View Consensus

Wenhao Wu¹, Hau-San Wong¹, and Si Wu²

¹Department of Computer Science, City University of Hong Kong

²School of Computer Science and Engineering, South China University of Technology

wenhaowu5-c@my.cityu.edu.hk, cshswong@cityu.edu.hk, cswusi@scut.edu.cn

Abstract

Stereo-based 3D object detection, which aims at detecting 3D objects with stereo cameras, shows great potential in low-cost deployment compared to LiDAR-based methods and excellent performance compared to monocular-based algorithms. However, the impressive performance of stereo-based 3D object detection is at the huge cost of high-quality manual annotations, which are hardly attainable for any given scene. Semi-supervised learning, in which limited annotated data and numerous unannotated data are required to achieve a satisfactory model, is a promising method to address the problem of data deficiency. In this work, we propose to achieve semi-supervised learning for stereo-based 3D object detection through pseudo annotation generation from a temporal-aggregated teacher model, which temporally accumulates knowledge from a student model. To facilitate a more stable and accurate depth estimation, we introduce Temporal-Aggregation-Guided (TAG) disparity consistency, a cross-view disparity consistency constraint between the teacher model and the student model for robust and improved depth estimation. To mitigate noise in pseudo annotation generation, we propose a cross-view agreement strategy, in which pseudo annotations should attain high degree of agreements between 3D and 2D views, as well as between binocular views. We perform extensive experiments on the KITTI 3D dataset to demonstrate our proposed method's capability in leveraging a huge amount of unannotated stereo images to attain significantly improved detection results.

1. Introduction

3D object detection, as one of the most significant perception tasks in the computer vision community, has witnessed great progress in recent years, especially after the advent of neural-network-based deep learning. Most state-of-the-art works which focus on 3D object detection mainly

Figure 1. Illustration of the proposed Temporal-Aggregation-Guided (TAG) disparity consistency constraint, in which the teacher model, temporally collecting knowledge from the student model, leads the disparity estimation of the student model across the views.

rely on LiDAR data [17, 32, 33, 43, 51] to extract accurate 3D information, such as 3D structure and depth of the points. However, LiDAR data are costly to harvest and annotate, and have limited sensing ranges in some cases. Instead, vision-based methods, which detect 3D objects based on images only, have drawn more attention in recent years. While it is convenient to collect image data, there are significant challenges in applying image-based methods to depth sensing, which is an ill-posed problem when localizing objects with only images. This problem is further exacerbated in monocular-based 3D object detection [24, 25, 48]. Stereo images, in which image pairs took at different viewpoints are available, can be used to reconstruct depth information through pixel-to-pixel correspondence or stereo geometry, making them more useful for detecting 3D objects without the introduction of expensive LiDAR data.

With more attention focusing on stereo-based 3D object detection, the performance of these methods gets continuously improved, and the gap between LiDAR-based methods and stereo-based methods becomes progressively narrower. However, the improving performance is built on large-scale manual annotation, which is costly in terms of both time and human resources. When the amount of annotation is limited, the performance of stereo-based methods deteriorates rapidly. Semi-supervised learning, which

* Corresponding author.

makes use of a limited set of annotated data and plenty of unannotated data, is a promising method for solving the data deficiency problem. In this work, we propose an effective and efficient semi-supervised stereo-based 3D object detection method to alleviate this limited annotation problem.

With limited annotated stereo images, depth estimation, a key stage for accurately localizing 3D objects, becomes increasingly unstable, which in turn leads to poor object localization. However, the inherent constraint between the left and right image in each stereo pair can be adopted to promote the model's performance in depth estimation. Inspired by the left-right disparity consistency constraint proposed by Godard et al. [11], we propose a Temporal-Aggregation-Guided (TAG) disparity consistency constraint, as shown in Fig. 1. In this method, disparity estimation of the base model, as the student model, should pursue that of the teacher model, with continuous knowledge accumulation from the base model, across the views. With the proposed TAG method, the base model can enhance its capability in depth estimation when provided with extra data without annotations. In addition, to generate more accurate pseudo annotations for the unannotated data, we propose a cross-view agreement strategy, which comprises a 3D-2D agreement constraint, in which pseudo annotations with high localization consistency in both 2D view and 3D view are kept, and a left-right agreement constraint, in which pseudo annotations with high similarities in both the left and right view at the feature space are retained for pseudo-supervision. With the proposed cross-view agreement strategy, the remaining pseudo annotations can effectively encompass the objects and provide high-quality supervision on the unannotated data, further improving the detection performance of the base model.

We evaluate the proposed method on the KITTI 3D dataset [10]. With our proposed method, the base model achieves relative performance gains of up to 18 percentage points under the evaluation metric AP_{3D} and 20 percentage points under the evaluation metric AP_{BEV} with the IoU threshold of 0.7 in the car category when only 5% of training data are annotated, thus verifying the effectiveness of our proposed method in making full use of data without annotations.

We summarize our main contributions as follows:

- 1) We propose a semi-supervised method for stereo-based 3D object detection, in which plentiful and easy-to-access unannotated images are fully utilized to enhance the base model.
- 2) We propose a Temporal-Aggregation-Guided (TAG) disparity consistency constraint to direct the disparity estimation of the base model through the teacher model, with cumulative knowledge from the base model, across the views.

3) We introduce a cross-view agreement strategy to refine the generated pseudo annotations through enforcing agreements between the 3D view and 2D view, and between the left view and right view.

- 4) Our proposed semi-supervised method leads to significant performance gains without requiring extensive annotations on the KITTI 3D dataset.

2. Related Works

2.1. Stereo-based 3D Object Detection

In 3D object detection, most of the attention is focused on LiDAR-based 3D object detection and monocular-based 3D object detection, and few works have been proposed to address 3D object detection based on stereo images, in which image pairs captured at different viewpoints are utilized to estimate the depth of each object in 3D space. Li et al. [18] proposed Stereo R-CNN, which is extended from Faster R-CNN [31], to predict and aggregate candidate 3D boxes across different views for estimating dense 3D boxes. Qin et al. [29] focused on object-level triangulation to locate 3D objects by introducing the Triangulation Learning Network. Wang et al. [39] proposed to lift pixels in stereo images to 3D points and perform 3D object detection on the pseudo point cloud. Chen et al. [6, 7] proposed to simultaneously achieve depth estimation, based on cost volume, and 3D object detection in the 3D world space. Xu et al. [42] reduced the depth estimation error of distant objects by zooming in corresponding image patches, followed by instance-level depth estimation and 3D object detection. Similarly, Sun et al. [35] implemented disparity estimation at the instance-level supervised by disparity maps generated from a shape prior model. Peng et al. [27] proposed to focus more on depth estimation by a cost reweighting strategy through the correlation between left and right feature maps. Guo et al. [12] enhanced the quality of depth estimation by distilling depth information from a LiDAR-based detection network trained on LiDAR points. Gao et al. [9] proposed to distill geometry-related information from a LiDAR-based model to a stereo-based model. Li et al. [21] focused on achieving fast 3D object detection, and proposed an efficient single-stage framework to achieve disparity estimation and 3D object detection at the same time.

2.2. Semi-supervised Object Detection

The impressive performance on 2D/3D object detection comes at the huge cost of time-consuming manual annotations for the training images. To reduce the annotation burden, there is an increasing focus on semi-supervised object detection, in which limited annotated data and numerous unannotated data are used to achieve acceptable performance on challenging scenes.

For 2D object detection, Jeon et al. [13] adopted consistency constraints on classification and localization between original images and their corresponding flipped versions. Sothet al. [34] adopted a fixed teacher model, trained on accessible annotated data only, to generate pseudo annotations, which will be used to optimize the student model. Liu et al. [22] utilized a teacher model, updated from a student model, to provide pseudo supervision for classification, optimized by the focal loss [20] to reduce imbalances between foreground-background and across classes. Tang et al. [36] adopted soft labels, instead of one-hot hard labels, from the teacher model to supervise the training of the student model. Zhou et al. [50] proposed a novel Instant-Teaching strategy, in which pseudo annotations are generated from two models that rectify each other under different initialization and data augmentation conditions. Wang et al. [40] proposed multi-phase learning for progressively generating and utilizing pseudo annotations from easy to difficult images. Xu et al. [41] proposed a Soft Teacher with reliability scores as weight factors to reduce the error due to the false-negatives. Li et al. [23] implemented semi-supervised learning on an anchor-free model, FCOS [37], and proposed the Listen2Student strategy to improve the regression performance on unannotated data. Chen [2] also extended FCOS by proposing an adaptive threshold method to reduce noise in pseudo annotations. Kral et al. [14] proposed a general mixing-unmixing strategy as strong augmentation to preserve more semantic information in the image space. Chen et al. [1] solved the label-mismatching problem at both distribution-level and instance-level. Chen et al. [3] and Zhan et al. [47] utilized unannotated data with weak annotations to facilitate semi-supervised object detection.

Although there is increasing attention on semi-supervised 2D object detection, not many works aim at utilizing a large amount of unannotated data on the 3D object detection task. Zhao et al. [49] first adopted the teacher-student framework to implement semi-supervised learning on 3D object detection with several proposed consistency constraints. Wang et al. [38] adopted a pseudo-labeling strategy to assign pseudo annotations for unannotated images after careful filtering and selection. Ye et al. [45] proposed to generate pseudo annotations by aggregating over-voted predicted boxes from the teacher model on unannotated data under different views. Li et al. [19] implemented semi-supervised learning on monocular 3D object detection with the proposed Geometric Reasoning Module. Park et al. [26] introduced a framework to implement semi-supervised learning on multi-modality 3D object detection, in which unannotated data in images and point clouds are not fully utilized to improve the performances in both 2D object detection and 3D object detection with the help of 2D-3D matching and consistency. Different from previous works,

we propose to achieve semi-supervised learning on 3D object detection with only stereo images through our proposed Temporal-Aggregation-Guided (TAG) disparity consistency constraint to promote the accuracy of the predicted disparity maps, and cross-view agreement strategy to improve the quality of generated pseudo annotations.

3. Method

3.1. Problem Definition

The objective of stereo-based object detection is to locate and discriminate 3D objects, represented by centers $(x; y; z)$, dimensions $(l; w; h)$ and orientation angles, based on paired left and right images. In the semi-supervised setting, limited images $X_a = \{I_{l_i}^a; I_{r_i}^a; y_i^a\}_{i=1}^{N_a}$ are manually annotated, where $I_{l_i}^a$ and $I_{r_i}^a$ are the left image and right image, respectively, y_i^a denotes the ground truth and N_a is the number of annotated images. The remaining large amount of images $X_u = \{I_{l_i}^u; I_{r_i}^u\}_{i=1}^{N_u}$ are without annotations, where N_u denotes the number of unannotated images and $N_u \gg N_a$. Our goal is to develop an effective 3D object detector based on limited stereo images with annotations and a large amount of annotation-free stereo images.

3.2. Overview

An overview of our proposed method is shown in Fig. 2. We adopt the teacher-student framework to make use of unannotated images, where a student model and a teacher model mutually learn from each other. The student model and the teacher model are initialized with a pre-trained model on X_a . The teacher model is utilized to generate pseudo annotations for unannotated images, which will be used to progressively enhance the student model updated through standard gradient descent. In return, the teacher model is temporally updated from the student model through the exponential moving average (EMA) as follows:

$$T_{t+1} = T_t + (1 - \alpha) S_t; \quad (1)$$

where T and S denote parameters from the teacher model and the student model respectively, and α is the smoothing factor to prevent the teacher model from over-fitting to the student model when updating. We set $\alpha = 0.999$ in all experiments.

To encourage more effective self-ensembling learning, we adopt data augmentation to enhance the diversity of annotated images and unannotated images. Since the environmental factors, such as illumination, are slightly different between left and right images in different scenes, the model should adjust itself to various patterns between pixel correspondences through asymmetric chromatic augmentation [44], in which the left and right images are enhanced

Figure 2. An overview of the proposed framework. In the framework, the student model is trained on limited annotated stereo images by the supervision loss L_a and numerous unannotated stereo images with pseudo annotations, generated from the teacher model with cumulative knowledge of the student model, by the pseudo-supervision loss L_u . The Temporal-Aggregation-Guided (TAG) disparity consistency loss L_{TAG} is proposed to guide the disparity estimation of the student model through the stable and accurate disparity estimation of the teacher model. To reduce noise in the pseudo annotations, the cross-view agreement strategy is proposed to generate high-quality pseudo annotations with high degrees of agreements between different views, including 3D/2D views and left/right views.

with different chromatic augmentations, like contrast, saturation and brightness. The introduction of asymmetric chromatic augmentation improves the model's generalization of pixel variations between binocular images. Besides, challenging objects, the pseudo-supervision loss will not be applied on the background of unannotated images to prevent the problem of false-negatives.

We adopt random horizontal flipping as the weak augmentation for the unannotated images, which will be fed into the teacher model, and both symmetric geometric augmentation and asymmetric chromatic augmentation as the strong augmentation for the annotated and unannotated images, which will be sent to the student model.

The student model is trained on limited annotated images and numerous unannotated images with pseudo annotations from the teacher model, and the objective to be optimized is as follows:

$$L = L_a(X_a) + \omega_u L_u(X_u) + \omega_{cons} L_{TAG}(X_u); \quad (2)$$

where $L_a(X_a)$ represents the supervision loss applied on annotated images, including classification, localization and disparity estimation losses, as follows:

$$L_a(X_a) = L_{cls}(X_a) + L_{reg}(X_a) + L_{disp}(X_a); \quad (3)$$

ω_u and ω_{cons} are weighting factors to control the contributions of pseudo-supervision loss $L_u(X_u)$, with a form similar to $L_a(X_a)$, on unannotated images with pseudo annotations and Temporal-Aggregation-Guided (TAG) disparity consistency loss L_{TAG} between the student model and the teacher model under different views, which will be detailed in Sec. 3.3. We set ω_u and ω_{cons} as 0.1 and 1.0 in all experiments. The pseudo annotations will be redefined

3.3. Temporal-Aggregation-Guided Disparity Consistency

Depth estimation plays a key role in vision-based 3D object detection. When given limited annotated images, depth estimation becomes more important for accurately locating objects in 3D space. To facilitate more stable and precise depth estimation, Godard et al. [11] proposed the left-right disparity consistency constraint through imposing consistency between the output disparity maps of one view and the output disparity maps translated from the opposite view as follows:

$$L^{r \rightarrow l} = d_{ij}^r + d_{ij}^l; \quad (4)$$

where d_{ij}^l and d_{ij}^r represent disparity value at location $(i; j)$ for left and right views, respectively. However, annotation insufficiency drives poor disparity estimation of the base model, leading to unreliability of the cross-view disparity consistency. We resort to the output disparity maps from the teacher model, temporally aggregating knowledge from the student model, to direct the disparity estimation of the student model, which is referred to as Temporal-Aggregation-Guided (TAG) disparity consistency. Specifically, we denote the output disparity maps of the left and right images as d^l and d^r for the teacher model and d^s for the student model. The photometric consistency

between the teacher and student model after exchanging views' disparity maps is as follows:

$$L_{\text{cross } 1}^{r \rightarrow l} = d_{ij}^{r \rightarrow s} + d_{ij}^{l \rightarrow t} : \quad (5)$$

Right disparity maps of the student model projected by left disparity maps of the teacher model should be consistent with right disparity maps of the teacher model projected by left disparity maps of the student model, since they are directed towards the same target. The corresponding loss on the opposite view can be obtained after mirroring. To further enforce disparity consistency between the student model and the teacher model, we also introduce photometric consistency between the teacher model and student model without exchanging views' disparity maps as follows:

$$L_{\text{cross } 2}^{r \rightarrow l} = d_{ij}^{r \rightarrow s} + d_{ij}^{l \rightarrow t} : \quad (6)$$

As before, we can mirror the direction to obtain the corresponding loss on the opposite view. Finally, disparity maps of the student model should also be aligned to those of the teacher model in each view as follows:

$$L_{t \rightarrow s}^{l \rightarrow r} = d_{ij}^{l \rightarrow t} + d_{ij}^{r \rightarrow s} : \quad (7)$$

Taking all of these into account, we can obtain the final disparity consistency loss applied on unannotated images as follows:

$$L_{\text{TAG}}(X_u) = \frac{1}{W \cdot H} \sum_{i=1}^W \sum_{j=1}^H [(L_{\text{cross } 1}^{r \rightarrow l} + L_{\text{cross } 1}^{l \rightarrow r}) + (L_{\text{cross } 2}^{r \rightarrow l} + L_{\text{cross } 2}^{l \rightarrow r}) + (L_{t \rightarrow s}^{l \rightarrow r} + L_{t \rightarrow s}^{r \rightarrow l})]; \quad (8)$$

where W and H are the width and height of the disparity maps after downsampling.

3.4. Pseudo Annotation Refinement via Cross-View Agreement

While the quality of depth estimation improves with the proposed TAG method, there may exist noise in the pseudo annotations, in which the predicted boxes are unable to completely enclose the target objects in the unannotated images, or the predicted categories do not correspond to the target objects. To address these mismatching problems on localization and classification, we propose to filter pseudo annotations through cross-view agreement, from 3D view to 2D view and from left view to right view, to improve the quality of generated pseudo annotations.

Dimensional Agreement A high-quality pseudo annotation requires high degrees of agreement on localization in both 3D and 2D views. Therefore, with the camera intrinsic parameters, predicted 2D boxes should have significant overlaps with regions re-projected from predicted 3D

boxes. Taking the Intersection-over-Union (IoU) between predicted 2D boxes and re-projected 3D boxes as the signal, we filter out predicted boxes with IoU less than the threshold iou , and the remaining predicted boxes are considered as pseudo annotation candidates. In this work, we set $\text{iou} = 0.95$ in all experiments.

Viewpoints Agreement Objects in the low-level pixel space may appear slightly different in the two views due to environmental factors such as illumination and occlusion, leading to different decisions in either classification or localization between views. However, in the high-level feature space, the same objects should maintain similar representations in the feature maps of different views. Based on this assumption, we propose to filter pseudo annotations based on the cosine similarity scores between feature vectors of the corresponding boxes in the feature maps of different views as follows:

$$s_i^{l \rightarrow r} = \frac{(f_i^l)^T f_{i+D_i^l}^r}{\|f_i^l\| \|f_{i+D_i^l}^r\|}; \quad (9)$$

where f_i^l is the feature vector of each predicted box in the left view, and $f_{i+D_i^l}^r$ is the feature vector of the corresponding predicted box, after translated by the downsampled left view's disparity map D_i^l , in the right view. The cosine similarity score can measure the agreement of the model on different views for the same predicted box, and serves as a reliability measure of the generated pseudo annotations. We further select the pseudo annotations with $s_i^{l \rightarrow r}$ greater than a threshold sim , which is set as 0.7 in all experiments.

4. Experiments

4.1. Dataset and Evaluation Metric

Dataset We adopt the KITTI 3D dataset [10], an informative dataset with diverse stereo images, for evaluation. KITTI 3D dataset includes 7481 stereo images with manual annotations for training and 7518 stereo images for testing. Since there is no annotation provided for the test set, we follow [4] to divide training images into a training split with 3712 samples and a validation split with 3769 samples. Following [8, 30], we first project the sparse LiDAR points on the image plane to construct the sparse depth maps, followed by interpolating with the efficient depth completion method [16] to generate the corresponding dense depth maps for annotated data. With the refined dense depth maps, we can obtain the dense disparity maps for initial supervision of the disparity estimation on annotated data. The experiments are mainly performed on the car category, and the experimental results of other classes can be found in the supplementary material.

Evaluation Metric We follow the official evaluation metric [10] of Average Precision (AP) on three modes, easy,

moderate and hard, considering occlusion, truncation and the size of objects in the images. We sample 40 recall points, as the KITTI benchmark, for AP calculation on the test set, which is reported in the supplementary material, and 11 recall points on the validation set for a fair comparison with previous competing methods. We also report the AP calculation with 40 recall points on the validation set in the ablation study.

4.2. Implementation Details

Baseline Model We adopt the light-weight and efficient YoloStereo3D [21], which is built on a light-weight cost volume with the correlation between feature maps of binocular images for 3D object detection and disparity estimation, as the base detector for the student model and the teacher model.

Training The base detector is initially trained on limited annotated data for pre-training with the same hyperparameters as in YoloStereo3D [21], the resulting model of which is used to initiate both the student model and the teacher model. We train the model using Adam [15] optimizer with a batch size of 4 images, half of which are from the annotated subset and another half of which are from the unannotated subset. We set the initial learning rate as 0.0001 with the cosine annealing strategy as the learning rate decay, and stop the model's training at 36 epochs.

We adopt random horizontal flipping as the weak augmentation for unannotated data to be input into the teacher model for stable and accurate pseudo annotation generation and both asymmetric chromatic augmentation and symmetric geometric augmentation as the strong augmentation for annotated data and unannotated data to be fed into the student model. Asymmetric chromatic augmentation randomly applies different brightness, contrast, saturation, and hue on binocular images. Symmetric geometric augmentation randomly applies horizontal flipping, translation, and scaling on stereo images to the same extent.

Inference Images with the top 100 pixels cropped are resized to 288×1280 for inference and fed into the trained teacher model with an inference speed of 160 ms per frame on one GTX 2080Ti. During inference, images are not augmented by either asymmetric chromatic augmentation or symmetric geometric augmentation.

4.3. 3D Object Detection on KITTI

We first conduct a performance comparison between the competing stereo-based 3D detection methods and the fully-supervised base detector, YoloStereo3D, with our proposed TAG method and cross-view agreement strategy on the validation set. The results are shown in Tab. 1. The proposed TAG method and cross-view agreement strategy outperform the base model, YoloStereo3D, exhibits its outstanding performance compared to all depth-independent methods, such as the SIDE [28], and even some of the sophisticated multi-stage

(a) 10% Annotated

(b) 20% Annotated

Figure 3. Error comparison of depth estimation between the base model with and without the proposed TAG method under different annotation ratios.

pseudo-LiDAR methods. The base model also demonstrates its high efficiency on inference compared to high-cost depth-dependent methods, like DSGN [7]. Building on this powerful framework, our proposed TAG method further improves the performance of the base model while just requiring comparable inference time. To fully explore the effectiveness of our proposed cross-view agreement strategy, we exploit additional 3k easy-to-access stereo images from the KITTI-raw dataset without annotations to enhance the base detector, which leads to further performance improvement compared to the model trained on annotated data only, thus highlighting the effectiveness of our proposed framework on utilizing readily available unannotated binocular images for achieving enhanced performance. We also conduct an experiment on comparing performance between the base detector with our proposed method and a number of stereo-based 3D detection methods, trained on only 5% of full annotations, on the validation set. The result is shown in Tab. 2. Both the performance of depth-independent and depth-based methods deteriorate significantly when provided with only limited annotated stereo images. On the other hand, our proposed method demonstrates its advantages even with only hundreds of stereo images with annotations. Specifically, the base model with our proposed method, trained on partially annotated stereo images, achieves performance gains of up to 9% on both AP_{BEV} and AP_{3D} under the setting of $\alpha = 0.5$ and $\beta = 0.7$, a significant improvement compared to the base model trained on only 5% of fully-annotated data, demonstrating that our proposed method can effectively exploit unannotated stereo images to enhance the performance of the base detector.

4.4. Ablation Study

We further conduct a number of experiments to highlight the contributions of different strategies in our proposed semi-supervised learning framework. Unless otherwise stated, we randomly sample 5% of the training split as the annotated subset and the remaining images are unannotated for the ablation experiments. All reported results are

Methods	Depth	AP _{BEV} / AP _{3D} (IoU=0.5)			AP _{BEV} / AP _{3D} (IoU=0.7)			Time (ms)
		Easy	Mod.	Hard	Easy	Mod.	Hard	
3DOP [5]		55.04/46.04	41.25/34.63	34.55/30.09	12.63/6.55	9.49/5.07	7.59/4.10	-
TLNet [29]		62.46/59.51	45.99/43.71	41.92/37.99	29.22/18.15	21.88/14.26	18.83/13.72	-
IDA-3D [27]		88.05/87.08	76.69/74.57	67.29/60.01	70.68/54.97	50.21/37.45	42.93/32.23	-
S-RCNN [18]		87.13/85.84	74.11/66.28	58.93/57.24	68.50/54.11	48.30/36.69	41.47/31.07	280
SIDE [28]		88.35/87.70	76.01/69.13	67.46/60.05	72.75/61.22	53.71/44.46	46.16/37.15	260
PL+FP [39]	X	89.80/89.50	77.60/75.50	68.20/66.30	72.80/59.40	51.80/39.80	44.00/33.50	670
PL+AVOD [39]	X	89.00/88.50	77.50/76.40	68.70/61.20	74.90/61.90	56.80/45.30	49.00/39.00	510
PL++ [46]	X	89.80/89.70	83.80/78.60	77.50/75.10	82.00/67.90	64.00/50.10	57.30/45.30	510
DSGN [7]	X	-	-	-	83.24/72.31	63.91/54.27	57.83/47.71	682
YoloStereo3D [21]	X	90.63/90.49	78.85/70.77	61.38/60.76	78.37/73.85	57.58/48.62	42.35/39.33	160
Ours wo/ unannot.	X	90.98/90.67	79.52/73.11	62.01/61.21	79.08/74.36	58.00/49.01	43.65/40.99	160
Ours w/ unannot.	X	91.60/91.48	81.49/76.82	63.08/62.53	80.66/75.19	59.23/51.49	45.56/42.69	160

Table 1. Performance comparison of average precision on bird's eye view (AP_{BEV}) and 3D boxes (AP_{3D}) between our proposed method and competing methods, trained on fully-annotated data, on the KITTI validation set. "Time" means inference time on the validation set.

Methods	Depth	AP _{BEV} / AP _{3D} (IoU=0.5)			AP _{BEV} / AP _{3D} (IoU=0.7)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
S-RCNN [18]		36.99/30.05	33.25/26.95	28.06/22.42	13.32/10.35	12.47/10.06	11.09/9.83
IDA-3D [27]		68.53/62.65	52.25/48.42	44.23/39.97	30.69/18.78	23.38/12.46	19.53/11.18
SIDE [28]		52.26/48.34	45.16/40.64	37.48/34.42	28.43/16.72	21.68/13.79	18.29/12.03
YoloStereo3D [21]	X	81.91/73.04	57.34/48.89	41.66/40.37	45.77/32.43	33.14/23.47	26.76/18.94
Ours	X	85.98/83.91	66.80/58.65	50.72/49.34	54.76/38.42	40.21/26.90	29.26/22.57

Table 2. Performance comparison of average precision on bird's eye view (AP_{BEV}) and 3D boxes (AP_{3D}) between our proposed method and competing methods, trained on 5% of fully-annotated data, on the KITTI validation set.

evaluated with 40 recall points on the validation set.

Disparity Consistency We show the effectiveness of the proposed TAG method by comparing the depth estimation error between the models with and without the TAG method when training with different ratios of fully-annotated data, which is shown in Fig. 3. We observe that the model with our proposed TAG method achieves satisfactory depth estimation with less error in different depth ranges, verifying that the teacher model can serve as a good supervisor to improve the disparity estimation of the student model.

Pseudo Supervision We conduct an ablation study to explore the contribution of each training strategy in our proposed method, as shown in Tab. 3. The performance of the

base model trained on 5% of fully-annotated data, as the baseline model, is far from satisfactory on the validation set. We then introduce unannotated images into the training process, leading to further improvement in the detection performance of the base model. To reduce noise in the pseudo annotations, we enforce the cross-view agreement between predicted boxes without high degrees of agreements between different views, including 3D/2D views and left/right views, which provide effective pseudo supervision for the

Con gs	AP _{BEV} / AP _{3D} (IoU=0.7)		
	Easy	Mod.	Hard
Baseline	45.25/28.36	27.93/18.09	20.99/13.82
+ pseudo anno.	46.11/29.99	29.98/19.61	21.88/14.01
+ data aug.	48.11/31.60	32.01/21.47	24.10/16.20
+ TAG	49.38/32.71	34.08/22.30	26.46/17.54
+ agreement	53.89/35.91	36.30/24.50	27.99/18.81

Table 3. Ablation study of different strategies in our proposed semi-supervised learning framework on the KITTI validation set.

Chrom.	Geo.	AP _{BEV} / AP _{3D} (IoU=0.7)		
		Easy	Mod.	Hard
X	X	46.85/30.44	31.15/20.23	23.86/14.64
		47.62/31.31	32.00/21.12	24.64/15.61
X	X	49.61/33.13	34.15/22.90	26.01/17.07
		53.89/35.91	36.30/24.50	27.99/18.81

Table 4. Performance comparison of the base model trained on different data augmentation settings.

Dim.	View.	AP _{BEV} / AP _{3D} (IoU=0.7)		
		Easy	Mod.	Hard
X	X	49.38/32.71	34.08/22.30	26.46/17.54
		50.15/33.40	34.98/23.12	26.87/17.90
X	X	52.02/34.21	35.18/23.93	27.49/18.43
		53.89/35.91	36.30/24.50	27.99/18.81

Table 5. Performance comparison of the base model trained on different agreement strategies.

Methods	Ratios of Annotated Stereo Images				
	5%	10%	20%	50%	100%
Baseline	18.09	24.32	38.13	43.29	46.96
Ours	24.50	30.84	41.25	44.89	47.67

Table 6. Performance comparison of our proposed method trained on different ratios of annotated stereo images AP_{BEV} under the difficulty of moderate and the setting of IoU=0.7.

trained with our proposed method consistently surpasses that trained on annotated data only. With the increase of annotated data, the relative improvements between both models become smaller since the proportion of unannotated data decreases in the training set. The consistently high performances verify the effectiveness of our proposed TAG method and cross-view agreement strategy in utilizing unannotated data to enhance the resulting model. More details about detection performances under different evaluation metrics can be found in the supplementary material.

unannotated images and further enhance the base model.

As a result, the contributions from all the strategies in our proposed method are required to achieve significant performance gains of up to 8% on the detection results.

Data Augmentation We next adopt different augmentation combinations to verify the effectiveness of the asymmetric chromatic augmentation and symmetric geometric augmentation, the results of which are shown in Tab. 4. Asymmetric chromatic augmentation can adapt the base model to different environment variations across different views. Symmetric geometric augmentation can promote the diversity of training data and the resulting robustness of the base model. Both types of augmentations are important for improving the performance of the base model.

Cross-View Agreement We also explore the contribution of each stage in the cross-view agreement strategy in Tab. 5. Filtering based on the agreement between 3D and 2D views guarantees that the selected pseudo annotations satisfy high localization consistency between the 3D space and 2D space, a necessary but not sufficient requirement for high-quality pseudo annotations. To further alleviate the problem of noisy pseudo annotations, only predicted boxes with a high degree of consensus from different viewpoints in the high-level feature space are retained for pseudo supervision on the unannotated data. The model trained on unannotated stereo images with high agreements on both 3D/2D and left/right views achieves significant performance gains on both AP_{BEV} and AP_{3D} under the three modes.

Annotation Ratio We also quantitatively analyze the cases when different ratios of stereo images are annotated in Tab. 6. Under different annotation ratios, the base model

5. Conclusion

In this work, we propose a semi-supervised learning framework to address the problem of data deficiency in stereo-based 3D object detection through leveraging information in unannotated stereo images. In the framework, we adopt a teacher model, which temporally accumulates knowledge from the student model, to generate pseudo annotations for supervising the training of a student model on unannotated data. With limited annotated stereo images, the student model performs poorly on depth estimation. To address this problem, we propose the Temporal-Aggregation-Guided (TAG) disparity consistency constraint, in which the teacher model with cumulative knowledge directs the disparity estimation of the student model across the views. To mitigate noise in the pseudo annotations, we propose a cross-view agreement strategy to preserve high quality pseudo annotations with high degrees of agreements between different views, including 3D/2D views and left/right views. Incorporating our proposed TAG method and cross-view agreement strategy, the resulting model attains significant performance improvement on the KITTI 3D dataset through the more reliable pseudo annotations and stable depth estimation process.

Acknowledgement This work was supported in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), in part by the National Natural Science Foundation of China (Project No. 62072189), and in part by the Guangdong Basic and Applied Basic Research Foundation (Project No. 2022A1515011160).

References

- [1] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 14381–14390, 2022. **3**
- [2] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 4815–4824, 2022. **3**
- [3] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8823–8832, 2021. **3**
- [4] 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**
- [5] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence* 40(5):1259–1272, 2017. **7**
- [6] Yilun Chen, Shijia Huang, Shu Liu, Bei Yu, and Jiaya Jia. Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2022. **2**
- [7] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 12536–12545, 2020. **2, 6, 7**
- [8] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022. **5**
- [9] Aqi Gao, Yanwei Pang, Jing Nie, Zhuang Shao, Jiale Cao, Yishun Guo, and Xuelong Li. Esgn: Efficient stereo geometry network for fast 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 2022. **2**
- [10] 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**
- [11] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 270–279, 2017. **2, 4**
- [12] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 3153–3163, 2021. **2**
- [13] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems* 32, 2019. **3**
- [14] Jongmok Kim, Jooyoung Jang, Seunghyeon Seo, Jisoo Jeong, Jongkeun Na, and Nojun Kwak. Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 14512–14521, 2022. **3**
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014. **6**
- [16] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *2018 15th Conference on Computer and Robot Vision (CRV)* pages 16–22. IEEE, 2018. **5**
- [17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 12697–12705, 2019. **1**
- [18] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 7644–7652, 2019. **2, 7**
- [19] Peixuan Li and Huaici Zhao. Monocular 3d detection with geometric constraint embedding and semi-supervised training. *IEEE Robotics and Automation Letters* 6(3):5565–5572, 2021. **3**
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision* pages 2980–2988, 2017. **3**
- [21] Yuxuan Liu, Lujia Wang, and Ming Liu. Yolostereo3d: A step back to 2d for efficient stereo 3d detection. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13018–13024. IEEE, 2021. **2, 6, 7**
- [22] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. **3**
- [23] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9819–9828, 2022. **3**
- [24] Zechen Liu, Zizhang Wu, and Roland Sattler. 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. **1**
- [25] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization

- errors for monocular 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 4721–4730, 2021. 1
- [26] Jinhyung Park, Chenfeng Xu, Yiyang Zhou, Masayoshi Tomizuka, and Wei Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. *arXiv preprint arXiv:2203.09510*, 2022. 3
- [27] Wanli Peng, Hao Pan, He Liu, and Yi Sun. Ida-3d: Instance-depth-aware 3d object detection from stereo vision for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 13015–13024, 2020. 2, 7
- [28] Xidong Peng, Xinge Zhu, Tai Wang, and Yuexin Ma. Side: Center-based stereo 3d detector with structure-aware instance depth estimation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* pages 119–128, 2022. 6, 7
- [29] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 7615–7623, 2019. 2, 7
- [30] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8555–8564, 2021. 5
- [31] 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
- [32] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 10529–10538, 2020. 1
- [33] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Point-rcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 770–779, 2019. 1
- [34] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3
- [35] Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qin-hong Jiang, Xiaowei Zhou, and Hujun Bao. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 10548–10557, 2020. 2
- [36] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 3132–3141, 2021. 3
- [37] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision* pages 9627–9636, 2019. 3
- [38] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 14615–14624, 2021. 3
- [39] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8445–8453, 2019. 2, 7
- [40] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 4568–4577, 2021. 3
- [41] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 3060–3069, 2021. 3
- [42] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12557–12564, 2020. 2
- [43] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors* 18(10):3337, 2018. 1
- [44] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 5515–5524, 2019. 3
- [45] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teacher. *arXiv preprint arXiv:2207.12655*, 2022. 3
- [46] 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. *International Conference on Learning Representations*, 2020. 7
- [47] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9417–9426, 2022. 3
- [48] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 3289–3298, 2021. 1
- [49] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 11079–11087, 2020. 3

- [50] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 3
- [51] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1