# COCO-TEACH: A CONTRASTIVE CO-TEACHING NETWORK FOR INCREMENTAL 3D OBJECT DETECTION

*Zhongyao Cheng, Cen Chen, Ziyuan Zhao, Peisheng Qian, Xiaoli Li, Xulei Yang*†

Institute for Infocomm Research ($I^2R$), A*STAR, Singapore

## ABSTRACT

Deep learning (DL) models for 3D object detection from point clouds have shown remarkable progress in various autonomous perception scenarios. However, the issue of catastrophic forgetting seriously hinders the deployment of these models in real-world applications where new classes are encountered over time. In order to address this issue, we present the Contrastive Co-Teaching Network (COCO-TEACH) framework for class-incremental 3D object detection. Our proposed framework consists of two teacher networks: a primary teacher network that detects old class objects in new data and provides them with pseudo-labels and an auxiliary teacher network that leverages the unlabelled objects in new data. The two teacher models transfer their learned knowledge to the target student model through a class-aware consistency loss. To enhance this transfer, a supervised contrastive loss is further incorporated into the loss function. We evaluate the performance of our proposed method against baseline methods through extensive experiments on two benchmark datasets. The results show that our proposed framework achieves state-of-the-art performance on incremental 3D object detection.

***Index Terms***— Deep learning, Incremental learning, Knowledge distillation, Point cloud data, 3D object detection

## 1. INTRODUCTION

Deep learning has shown impressive progress in object detection with various real-world applications. Compared to 2D image data, point cloud provides a more robust 3D sparse representation containing rich geometry and layout information of the environment. However, existing deep learning-based approaches require a large amount of labeled data which is difficult to collect over time, hence suffering from "catastrophic forgetting" issue when new annotations/classes are incrementally accumulated where a deep learning model forgets its previous knowledge while learning from new classes, ultimately leading to poor performance on the old classes.

†Corresponding author. This research work is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-027).

To address this issue, incremental learning is proposed to enable the model to accommodate new classes without forgetting the knowledge learned with old classes. Incremental learning has been widely investigated for 2D vision tasks, especially on classification [1]. Incremental learning has also been gradually investigated for 2D object detection, such as [2, 3, 4, 5], in which knowledge distillation [6] is utilized to prevent forgetting. However, incremental learning for 3D object detection remains underexplored. There is only a few research works [7, 8, 9] studying this task.

The previous works [7] use pseudo label approach to guide the target model via supervised learning without revisiting the old samples. However, this approach can sometimes misguide the student through the absent and incorrect labels because of "inaccurate" and "incomplete" detection results. Zhao et al. [8] tackle potential issues related to the efficiency of the pseudo label by applying knowledge distillation for old classes. However, the hard target term in knowledge distillation loss still relies on these "incorrect" labels. Additionally, the regions identified by the student model may only contain objects of new classes due to its interest in all classes, which limits the efficiency of knowledge transfer. Therefore, efficient transfer of previous knowledge to the student model is required for effective knowledge distillation.

In this study, we propose a Contrastive Co-Teaching Network framework, termed COCO-TEACH, to investigate class-incremental learning for 3D object detection. Our proposed framework consists of two teachers. The primary teacher provides pseudo labels for old classes to teach the student without storing and revisiting the previous samples of old classes. Moreover, it also transfers knowledge previously learned to the student via knowledge distillation. The auxiliary teacher uses a self-ensemble approach inspired by Mean Teacher [10] to further exploit the underlay knowledge. To overcome the issues from the pseudo label, we incorporate supervised contrastive learning into the loss function, which summarizes the feature well when the high-quality label is unavailable. Besides, we exploit the loss function used for knowledge distillation and propose to use the class-aware consistency loss in [11], which does not rely on the hard target item from the naive pseudo label. Furthermore, we use the primary teacher to identify regions containing abundant knowledge of old classes, preventing the student from

only focusing on new class objects, as the primary teacher is only interested in old classes. To demonstrate the effectiveness of our proposed COCO-TEACH, we conduct extensive experiments on the ScanNet and SUN RGB-D datasets.

## 2. RELATED WORK

### 2.1. 3D Object Detection

The modern 3D object detection techniques can be classified into three groups, namely voting-based, transformer-based, and 3D convolutional-based. The voting-based methods were the first to be introduced, with VoteNet [12] being the pioneer that presented point voting for detecting 3D objects. VoteNet extracts features from 3D points, groups a set of points for each object candidate based on their voted center, and computes object features from each point group. Subsequent works, such as BRNet [13], H3DNet [14], and RBGNet [15], have improved the generation of point groups by considering local geometric primitives to refine the voting results. Transformer-based methods utilize end-to-end learning and perform a forward pass on inference, which makes them less domain-specific. GroupFree [16] incorporates a transformer module to iteratively update object query locations and aggregate intermediate outcomes. 3DETR [17] was the first to use a transformer model to solve the 3D object detection task. Voxel-based 3D object detection approaches [18, 19] transform points into voxels and handle them using 3D convolutional networks. We choose VoteNet as our backbone for 3D object detection.

### 2.2. Incremental Learning

The term "incremental learning" refers to the ability to continuously learn new knowledge while retaining previously learned knowledge. Most studies on incremental learning have focused on classification problems, which can be broadly divided into three categories [1]. The first category is regularization-based, which includes data regularization, weight regularization, and other regularization terms to mitigate the catastrophic forgetting. The second category prevents catastrophic forgetting by either growing a sub-network or learning a mask for each task. The third category, rehearsal-based approaches, stores a small number of training samples from previous tasks or uses a generative model to sample synthetic data from previously learned distributions. Object detection through incremental learning is an area of research that has received limited attention. Shmelkov et al [2] were the first to propose a method for learning an object detector incrementally by adding one or a few classes at a time within a single domain. Subsequent works, such as Hao et al [3], Peng et al [4] and Li et al [5], utilize knowledge distillation, employing either feature or response knowledge or both. However, incremental learning for 3D object detection remains less studied. There is only a few research works [7, 8, 9] on this task. Our proposed COCO-TEACH method focuses on the same goal and aims to achieve state-of-the-art performance on class-incremental 3D object detection.

## 3. METHODOLOGY

### 3.1. Problem Definition

In class-incremental learning tasks, we denote the base data as $D_{base}$, only containing the annotation of base classes $C_{base}$, which is used to train the base model. The novel data $D_{novel}$, which is used to train the target model, only contains the annotation of novel classes $C_{novel}$. The $C_{base}$ and $C_{novel}$ do not overlap. However, the $D_{base}$ and $D_{novel}$ can contain some common samples when these samples contain both $C_{base}$ and $C_{novel}$ objects, but these samples have different annotation in the two datasets.

### 3.2. Contrastive Co-Teaching Framework

To alleviate the aforementioned issue, we propose a novel Contrastive Co-Teaching framework consisting of three networks as shown in Fig. 1: a primary teacher, an auxiliary teacher, and a student. All three networks use Votenet [12] as a 3D object detection model backbone, given its efficiency and simplicity in point cloud-based 3D object detection.

The co-teaching framework is inspired by previous research works [8, 20], with two major differences in our framework: the first one is that knowledge distillation uses the $C_{base}$ scores of the regions proposed by the primary teacher, rather than the student. The second one is that we modify the loss function by introducing contrastive learning to overcome the challenges because of the "inaccurate" pseudo label.

Given a point cloud in $D_{novel}$, the primary teacher, the model is well-trained by $D_{base}$, detects the $C_{base}$ objects. The detection result is used to preserve previous knowledge in two aspects. Firstly, it provides the pseudo label for the $C_{base}$ objects in $D_{novel}$ to avoid the student is misguided because of the absence of the annotation. To guarantee the quality of the pseudo label, we set two thresholds for both of proposing region score and classification score. Only the candidates who get high scores in both of the two steps can be selected to build the pseudo label. Secondly, the detection result is also used in knowledge distillation. The paired outputs of the student and the primary for the same regions are utilized to compute the difference by the proposed loss function for knowledge distillation.

The auxiliary teacher is inspired by the mean teacher [10] a self-ensembling technique that is originally proposed to exploit unlabeled data effectively. Its weight $\theta'_t$ is updated by the EMA [21] from the weight of student $\theta_s$ while $\alpha$ governs the effects of the current parameter in the student model

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_s \tag{1}$$

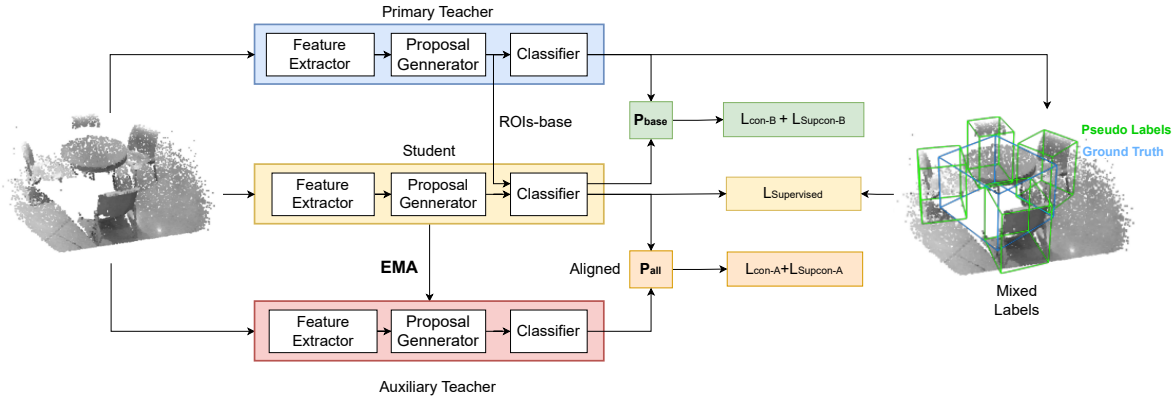


**Fig. 1**. Overview of our proposed contrastive co-teaching framework. The primary teacher detects old class objects and provides them with pseudo-labels. The auxiliary teacher network leverages the unlabelled objects in new data via a Mean-teacher manner. Both teachers transfer their learned knowledge to the target student through a class-aware consistency loss and a supervised contrastive loss.

### 3.3. Proposed Regions for Knowledge Distillation

In order to transfer knowledge efficiently, we use the primary teacher to extract the regions with abundant knowledge of $C_{base}$. The previous research work [8] compares the pair of $C_{base}$ scores of the same regions extracted by student. However, the student's target is all classes while the knowledge distillation step only pays attention to the $C_{base}$. In some scenarios, these regions extracted by the student only contain the $C_{novel}$ objects, lacking wealthy knowledge for $C_{base}$. Compared to the student, which has interest to all classes, the primary teacher focuses on the $C_{base}$. All of the proposed regions contain abundant knowledge for $C_{base}$, improving the efficiency of knowledge distillation.

### 3.4. Loss Functions

To optimize the loss function and train the network parameters, we compare the paired scores for $C_{base}$ from the same ROIs and introduce knowledge distillation with the proposed loss function to transfer the previous knowledge. In addition, the point cloud with perturbations is simultaneously input into the auxiliary teacher, and the corresponding detection results are compared with those of the student.

The cross-entropy in knowledge distillation loss function [6] relies on pseudo labels to compute the score difference between the paired models. However, pseudo labels are not as strong as real annotations because the primary teacher may fail to detect all $D_{base}$ objects due to factors such as limited detection capabilities or different data distribution. Consequently, incomplete or inaccurate detection results cause the primary teacher to mislead the student via the knowledge distillation loss function.

To address the issue mentioned above, we introduce supervised contrastive learning [10] which is widely used in self-supervised learning. It compares the pair of input labeled objects and brings similar objects closer, and maximizes the difference while the objects are dissimilar. The loss function is described as followed:

$$\mathcal{L}_{\text{supcon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)},$$

Where $z_i$ denotes the feature representation of the $i$-th sample, $\tau$ is the temperature parameter, $z_a$ is the feature representation for the samples $a$ belonging to class $A$. $P(i)$ is the set of indices of all positives in the batch distinct from $i$, and $|P(i)|$ is the corresponding cardinality

Compared to cross-entropy, the above loss function has two advantages. 1) It can effectively learn from noisy and incomplete labels, as it learns to group similar instances together regardless of their exact class labels. This makes it particularly useful in scenarios the well-annotated data is unavailable. 2) It reduces dependence on large amounts of labeled data because it can leverage the structure of the data to improve generalization.

The primary teacher use loss function $L_{\text{pri}}$, consists of class-aware consistency loss for $C_{base}$, $\mathcal{L}_{\text{con-B}}$ and supervised contrastive loss $L_{\text{Supcon-B}}$ as the final loss function.

$$\mathcal{L}_{\text{pri}} = \mathcal{L}_{\text{con-B}} + \mathcal{L}_{\text{Supcon-B}}.$$

We also import the supervised contrastive loss for all classes $L_{\text{Supcon-A}}$ in class-aware consistency loss for the auxiliary teacher because it also needs to face the challenges of the incomplete label. The loss function $L_{\text{adv}}$ can be describe as:

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{con-A}} + \mathcal{L}_{\text{Supcon-A}}\,.$$

Finally, the overall loss function for the target student model is formulated as:

$$\mathcal{L} = \lambda_s \mathcal{L}_{sup} + \lambda_p \mathcal{L}_{\text{pri}} + \lambda_c \mathcal{L}_{\text{adv}},$$

where $\mathcal{L}_{sup}$ is the standard supervised loss for 3D object detection. The $\lambda_s$, $\lambda_p$ and $\lambda_c$ are the weight of the losses.

## 4. EXPERIMENTS

### 4.1. Dataset and Implementation Details

This section evaluates the proposed COCO-TEACH architecture on the ScanNet and Sun RGB-D datasets [22, 23]. ScanNet has 18 categories from 1513 samples with point level semantic segmentation. We build bounding boxes base on the semantic segmentation annotation. We select 10 common categories from hundreds of classes, the same as prior works (VoteNet) to evaluate our model better. To customize the class-incremental learning task dataset, We divide the classes into two sets: base classes $C_{base}$ and novel classes $C_{novel}$ according to the alphabetical order. We split the training data into $D_{base}$ and $D_{novel}$. $D_{base}$ is the subset that contains any $C_{base}$ object with only the annotation of $C_{base}$, the annotation of $C_{novel}$ will be removed if it contains the object belongs to $C_{novel}$, $D_{novel}$ is created by the same way.

We use VoteNet with pointnet2 backbone as the object detection model on Tesla V100. The primary teacher is trained from scratch, and the backbone of the student and the auxiliary teacher is initialized by the well-trained primary teacher. For both datasets, the learning rate is 0.001, and the batch size is 8. We use the mAP@0.25 as the evaluation metric.

### 4.2. Baseline Methods

Compared baselines are shown as follows: 1) **Freeze and add**: we freeze the weight of the primary teacher and only add the classifier for $C_{novel}$es for the novel data. 2) **Fine tuning**: fine-tune the parameter of the primary teacher and a novel classifier for the novel data, but keep the same weight for the base classifier. 3) **SDCOT** [8]: The distillation loss is used for the student net to learn from the primary teacher. 4) **COCO-TEACH**: The class-aware consistency loss and supervised-contrastive loss are used to learn from the primary teacher and the auxiliary teacher. The knowledge distillation step uses the regions proposed by the primary teacher.

### 4.3. Experimental Results

Table 1 displays the comparison results for two datasets. The "Freeze and add" method exhibits poor performance on $C_{novel}$ for both datasets, while the "fine-tuning" method experiences catastrophic forgetting on $C_{base}$. Overall, mAP shows improvement in both benchmarks of our proposed COCO-TEACH over the SOTA method SDCoT. Specifically, in Scannet, mAP improves from 53.96 to 55.28, and in Sun RGB-D, the result improves from 56.37 to 57.19. Notably, there is a significant improvement in $C_{base}$ for both datasets, improving from 53.65 to 55.63 in Scannet and from 52.10 to 53.28 in Sun RGB-D. Additionally, the results for $C_{novel}$ also show improvement in both datasets. Please take note the visualization results are not shown here due to the page limitations.

**Table 1**. Overall performance of different methods

| Methods | Base | Novel | All |
|---|---|---|---|
| SCANNET Dataset | | | |
| Base training | 59.60 | - | - |
| Freeze and add | 58.80 | 4.11 | 31.46 |
| Fine-tuning | 1.85 | 52.35 | 27.10 |
| SDCOT | 53.65 | 54.26 | 53.96 |
| COCO-TEACH | **55.63** | **54.89** | **55.28** |
| Sun RGB-D Dataset | | | |
| Base training | 56.82 | - | - |
| Freeze and add | 54.14 | 10.32 | 32.23 |
| Fine-tuning | 3.45 | 54.05 | 28.75 |
| SDCOT | 52.10 | 60.65 | 56.37 |
| COCO-TEACH | **53.28** | **61.40** | **57.19** |

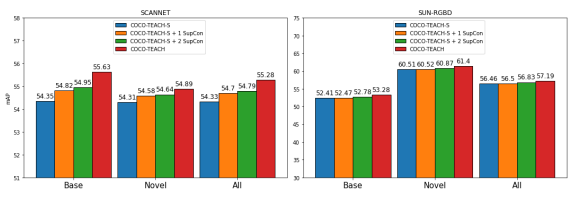


**Fig. 2**. Ablation Results for different Components

### 4.4. Ablation Study

Fig. 2 demonstrates the effectiveness of different components in the proposed models. **COCO-TEACH-S** uses class-aware consistency loss in knowledge distillation for the regions proposed by the student. **COCO-TEACH-S + 1 SupCon** adds supervised-contrastive loss for the primary teacher, while **COCO-TEACH-S + 2 SupCon** further adds supervised contrastive loss for the auxiliary teachers. Compared to COCO-TEACH-S, COCO-TEACH-S + 1 SupCon displays a noticeable improvement in all classes, particularly in $C_{base}$, indicating the benefits of supervised contrastive loss for the primary teacher. With the addition of the supervised contrastive loss in the auxiliary teacher, the performance for all classes shows some further improvement. Comparing the results of COCO-TEACH-S + 2 SupCon and COCO-TEACH model reveals our model gains a noticeable improvement from the extracted regions with ample information for $C_{base}$. Furthermore, it brings some improvement for $C_{novel}$ as well.

## 5. CONCLUSION

This study presents COCO-TEACH, a method that efficiently transfers acquired knowledge in class-incremental 3D object detection tasks. We accomplish this by incorporating two loss functions, supervised contrastive loss and class-aware consistency loss, in the knowledge distillation process. Moreover, we use primary teachers who specialize in old classes to propose regions of interest, thereby enhancing the efficacy of knowledge distillation. We demonstrate the effectiveness of our proposed COCO-TEACH on two benchmark datasets.

## 6. REFERENCES


[1] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[3] Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian, "An end-to-end architecture for class-incremental object detection with knowledge distillation," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1–6.

[4] Can Peng, Kun Zhao, and Brian C Lovell, "Faster ilod: Incremental learning for object detectors based on faster rcnn," *Pattern Recognition Letters*, vol. 140, 2020.

[5] Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck, "Rilod: near real-time incremental learning for object detection at the edge," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 113–126.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] Jiahua Dong, Yang Cong, Gan Sun, Bingtao Ma, and Lichen Wang, "I3DOL: incremental 3d object learning without catastrophic forgetting," *CoRR*, vol. abs/2012.09014, 2020.

[8] Na Zhao and Gim Hee Lee, "Static-dynamic co-teaching for class-incremental 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 3436–3445.

[9] Ziyuan Zhao, Mingxi Xu, Peisheng Qian, Ramanpreet Singh Pahwa, and Richard Chang, "Da-cil: Towards domain adaptive class-incremental 3d object detection," in *Proceedings of the British Machine Vision Conference*, 2022.

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

[11] Na Zhao, Tat-Seng Chua, and Gim Hee Lee, "Sess: Self-ensembling semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[12] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[13] Bowen Cheng, Sheng Lu, Shaoshuai Shi, Ming Yang, and Dong Xu, "Back-tracing representative points for voting-based 3d object detection in point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8963–8972.

[14] Zaiwen Zhang, Bo Sun, Haitao Yang, and Qixing Huang, "H3dnet: 3d object detection using hybrid geometric primitives," in *IEEE/CVF European Conference on Computer Vision (ECCV)*, 2020, pp. 311–329.

[15] Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng LI, Bernt Schiele, and Liwei Wang, "Rbgnet: Ray-based grouping for 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1110–1119.

[16] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong, "Group-free 3d object detection via transformers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2949–2958.

[17] Ishan Misra, Rohit Girdhar, and Armand Joulin, "An end-to-end transformer model for 3d object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2906–2917.

[18] Ji Hou, Angela Dai, and Matthias Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4421–4430.

[19] JunYoung Gwak, Christopher Choy, and Silvio Savarese, "Generative sparse detection networks for 3d single-shot object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 297–313.

[20] Ziyuan Zhao, Andong Zhu, Zeng Zeng, Bharadwaj Veeravalli, and Cuntai Guan, "Act-net: Asymmetric co-teacher network for semi-supervised memory-efficient medical image segmentation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1426–1430.

[21] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto, "Exponential moving average normalization for self-supervised and semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 194–203.

[22] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[23] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.