Prompt3D: Random Prompt Assisted Weakly-Supervised 3D Object Detection

Xiaohong Zhang*, Huisheng Ye*, Jingwen Li, Qinyu Tang, Yuanqi Li, Yanwen Guo† and Jie Guo
Nanjing University
{zhangxiaohong,huishengye,jingwenli,tangqinyu,yuanqili}@smail.nju.edu.cn,
{ywguo,guojie}@nju.edu.cn

Abstract

The prohibitive cost of annotations for fully supervised 3D indoor object detection limits its practicality. In this work, we propose Random Prompt Assisted Weakly-supervised 3D Object Detection, termed as ***Prompt3D****, a weakly-supervised approach that leverages position-level labels to overcome this challenge. Explicitly, our method focuses on enhancing labeling using synthetic scenes crafted from 3D shapes generated via random prompts. First, a Synthetic Scene Generation (SSG) module is introduced to assemble synthetic scenes with a curated collection of 3D shapes, created via random prompts for each category. These scenes are enriched with automatically generated point-level annotations, providing a robust supervisory framework for training the detection algorithm. To enhance the transfer of knowledge from virtual to real datasets, we then introduce a Prototypical Proposal Feature Alignment (PPFA) module. This module effectively alleviates the domain gap by directly minimizing the distance between feature prototypes of the same class proposals across two domains. Compared with sota BR, our method improves by 5.4% and 8.7% on mAP with VoteNet and GroupFree3D serving as detectors respectively, demonstrating the effectiveness of our proposed method. Code is available at: https://github.com/huishengye/prompt3d.*

1. Introduction

3D object detection, crucial for scene understanding, involves identifying and classifying objects within sensory data like images and point clouds. Advances in 3D sensing and deep learning have fueled interest in detecting 3D objects from unstructured point clouds. State-of-the-art algorithms [26, 33, 35, 36, 48, 51] treat the task of 3D object detection as a supervised learning problem. However, strong supervision using detailed annotations such as point-level labels or 3D bounding boxes is resource-intensive, hindering the growth of 3D detection applications.

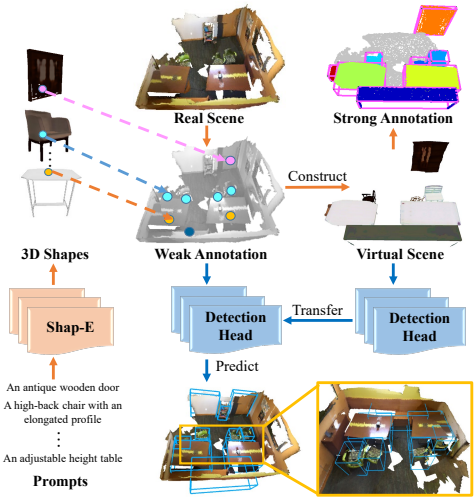

Figure 1. Illustration of the pipeline of the proposed Prompt3D. Our method first converts random textual prompts into 3D shapes using Shap-E [13]. These shapes are then matched with real scenes using a coarse layout inferred from position-level annotations. The resulting synthetic scenes, with pixel labels, enhance the training of the detection head through domain adaptation.

To address this limitation, recent studies [24, 37, 43] explored model training with weak annotations. This has led to the development of two primary methods: using scene-level class tags [43] and employing detailed position-level labels [23, 24, 54], such as object centroids and class tags. Position-level weak supervision, especially for outdoor detection [47], offers better time-accuracy balance and has gained prominence. Some weakly supervised methods [20, 30] in outdoor scenarios have achieved results on par with fully supervised techniques. However, adapting position-level weak supervision to complex indoor scenes is still challenging, with few precedents. To enhance weak labeling in indoor settings, recent advancements [39, 54] use virtual scenes with freely available annotations from stan-

*Equal Contribution.
†Corresponding author.

This work was supported by the National Natural Science Foundation of China (62032011) and the Natural Science Foundation of Jiangsu Province (No. BK20211147).

dard 3D shape datasets like ModelNet40 [53]. Nevertheless, the constrained scope of standard 3D shape datasets leads to synthetic scenes that feature only a limited number of object types in indoor environments, and these types are often represented with limited coverage. This limitation in both scale and diversity weakens the effectiveness of the weakly supervised approach.

Recently, the ascendancy of 3D generative models has provided an avant-garde avenue for the creation of diverse 3D shapes, unlocking new potentials in exploiting category information to its fullest extent. In light of this progression, our work introduces a novel weakly-supervised approach for 3D indoor object detection named Random Prompt Assisted Weakly Supervised 3D Object Detection (Prompt3D). In our work, the use of position-level labels, including the center and class tag of each object, serves as efficient supervision for annotation purposes. Unlike existing weakly-supervised methods that depend on predefined 3D shapes, our approach utilizes a plethora of 3D shapes generated by 3D text-driven generative models to provide additional supervision signals. The description of the scheme is shown in Fig. 1.

Firstly, our Synthetic Scene Generation (SSG) module is designed to create virtual scenes using textual prompts. We start by building a diverse library of prompt-based 3D shapes. To ensure broad coverage of object diversity, we prepare a comprehensive set of prompts for each category. By employing the advanced 3D generative model Shap-E [13], we convert textual prompts into corresponding 3D geometries. Our method leverages textual diversity and scalability, creating a 3D shape repository that surpasses current datasets in breadth and diversity. These shapes, integrated with position-level annotations, form the foundation for constructing novel and realistic synthetic scenes. The initial layout of virtual scenes is formed by systematically selecting the most compatible 3D shapes from our generated library to replace objects in real-world scenes. Our shape library's diversity expands the range of options for selecting lifelike shapes, which in turn enhances the diversity and realism of virtual scenes.

Furthermore, to reduce the disparity between synthetic and real scenes, we introduce the Prototypical Proposal Feature Alignment (PPFA) module. Specifically, our objective is to minimize the distance between prototypical features of the same class proposals in both real and synthetic scenes. This alignment strategy aims to harmonize the feature distributions of the two scenes. In contrast to prior work [54] utilizing adversarial feature alignment (AFA) to bridge the gap by aligning overall features in an adversarial manner, which can introduce scene-level bias and instability, our PPFA module focuses on transforming detailed knowledge into class-specific features. The proposed transfer method enhances the robustness and accuracy of knowledge transfer from virtual to real scenarios, thereby fully leveraging the powerful supervisory signals generated by synthetic scenes.

Guided by two essential modules, Prompt3D boosts weakly supervised object detection by incorporating supervisory information from virtual scenes created with text-generated shapes. Experiments on ScanNet affirm our approach's superiority. Compared to weakly supervised methods using standard shape repositories, leveraging prompt-based synthetic scenes notably improves detector discriminative abilities in weakly supervised settings. Specifically, when utilizing VoteNet as the detector in the weakly supervised setting, the performance difference from the fully supervised setting is trimmed to just 7.7% mAP@0.25. It is worth noting that when GroupFree3D is used as the detector, this gap further shrinks to only 3%.

To summarize, our main contributions are:

- We devise a novel approach for 3D indoor object detection under weak supervision, generating synthesized scenes derived from random textual prompts corresponding to various categories to enrich the supervisory signal.
- Our introduced transfer module is grounded in prototype learning, markedly boosting the generalization ability of detectors and effectively leveraging the supervisory information from the virtual scene.
- Experiments on the popular indoor dataset ScanNet validate the effectiveness of the proposed approach. Compared with sota BR, Prompt3D improves by 5.4% and 8.7% on mAP with VoteNet and GroupFree3D serving as detectors respectively. Our method achieves comparable detection performance with the fully-supervised setting.

2. Related Work

2.1. 3D Object Detection

3D object detection aims to identify and localize objects within their encompassing amodal 3D bounding boxes. Contemporary deep learning-based 3D object detection frameworks [6, 11, 33, 36, 40–42, 55, 59] presuppose comprehensive supervision and hinge on the premise of precise 3D annotations within datasets. In Qi et al. [33], 3D object detection from RGB-D data is performed in both indoor and outdoor scenes. MonoGRNet [35] is proposed for the amodal 3D object localization from a monocular RGB image. TLNet [36] triangulates the objects with stereo images.

Recently, weakly supervised 3D object detection has gained popularity as labeling precise 3D boxes is extremely laborious. As explored in [3, 7, 10, 14, 18, 45, 49, 50], it is assumed that the instance-level bounding box annotations are not provided by the training set, and the supervision can come from weak annotations. Explicitly, it encompass scene-level [43] and position-level methods [23, 24, 54]. The former annotates objects with only class tags, whereas the latter provides annotations for both the object's center

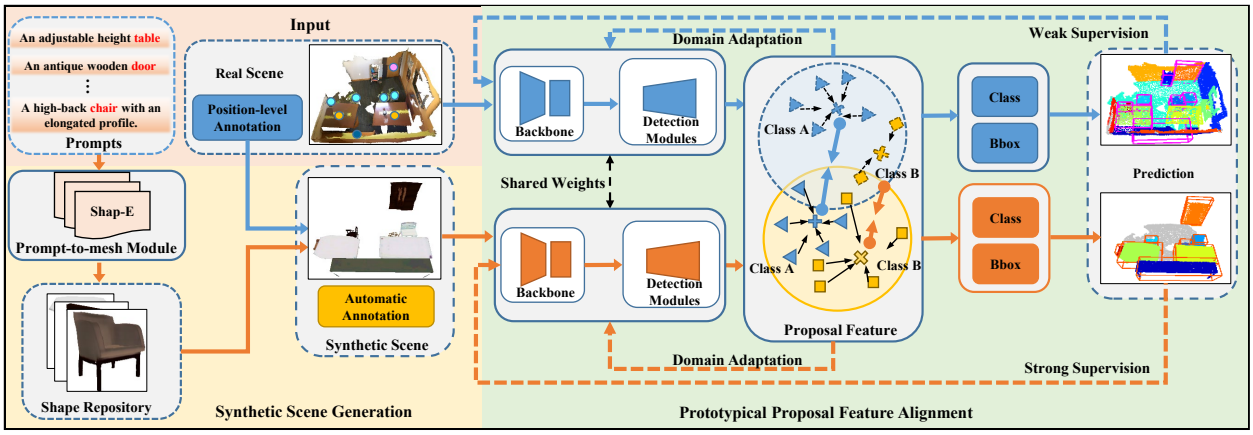

Figure 2. The framework of the proposed Prompt3D method. 1) The **Synthetic Scene Generation (SSG)** module uses weak annotations to select and assemble 3D shapes from a prompt-based generated repository, then recreating a new scene mirroring the original layout. 2) The **Prototypical Proposal Feature Alignment (PPFA)** module enhances object detection by aligning prototype box proposal features between synthetic and real scenes, effectively bridging the domain gap without relying on annotations.

and its class. While scene-level methods offer a reduction in manual annotation efforts, their efficacy is typically compromised by the absence of precise positional data, a limitation not faced by their position-level counterparts.

2.2. 3D generative models

The field of 3D generation [12, 27, 28, 31] seeks to create 3D assets from various modalities like text and images. The developments in 3D generation have reduced costs for expanding 3D datasets, resulting in larger and more realistic virtual datasets, essential for applications such as 3D detection. The achievements in neural volume rendering [1, 17, 25] have increased interest in rendering-aware 3D synthesis [4, 5, 46] that learns directly from images. The success of text-to-image models [38] has also sparked innovation in text-to-3D synthesis techniques [12, 27, 31, 44]. Technologies like CLIP-forge [44] employ normalizing flow models to create shape embeddings directly from text. DreamFusion [31] has shown impressive text-to-3D capabilities using a pre-trained text-image model. Magic3d [19] has led to more efficient mesh representations, streamlining the process further. The innovative Shap-E [13], which is trained on extensive 3D datasets, stands out for its ability to quickly and efficiently generate 3D structures from textual prompts. In our research, we utilize Shap-E to generate 3D objects from textual prompts, improving 3D detection by extracting more information from weakly supervised data.

2.3. Domain Adaptation

Domain adaptation involves transferring knowledge from one domain (the source domain) to another (the target domain), which is especially valuable when the training dataset is limited. Domain adaptation is commonly used in 2D visual tasks such as classification [52, 60] and semantic segmentation [2, 16, 56–58]. Typical methods include image-level bias correction, distribution-aware alignment, and centroid-aware alignment. Aligning distribution centroids has proven effective in reducing the domain gap. In more developments, domain adaptation has extended to 3D visual tasks, particularly in scenarios involving interactions across domains. For example, approaches like BR [54] and Randomrooms [39] have introduced global alignment modules to improve feature discriminability in 3D detection models operating in synthetic environments. In contrast, our approach focuses on aligning anchor distributions at the proposal level, presenting a tailored solution for 3D domain adaptation challenges.

3. Approach

3.1. Overview

The architecture of our weakly supervised object detection method is depicted in Fig. 2. Central to our approach is the compensation for information loss associated with weak supervision via the use of synthetic scenes, from which full annotations are automatically derived. These synthetic environments are crafted to approximate the real-world layout and are constituted using a 3D shape repository generated by the text-driven 3D generative model.

To be specific, the proposed method in our paper consists of two key stages: 1) Synthetic Scene Generation (SSG). During this phase, we meticulously curate a synthetic scene under the guidance of the weak supervision of real scenes. 2) Prototypical Proposal Feature Alignment (PPFA). The PPFA module is crafted to ameliorate the divergence between synthetic scenes and actual environments.

3.2. Synthetic Scene Generation

In the SSG module, we create synthetic scenes with 3D shapes under weak supervision. This process enables the

automatic generation of ground truth point annotations in synthetic scenes. The generated scenes can offer strong supervision. They fully harness the information within the limited annotations of real scenes, thereby enhancing the detection performance. Explicitly, the construction of synthetic scenes can be divided into two steps: prompt-to-mesh translation and 3D shape assembly.

Prompt-to-mesh translation. Considering the specific categories of objects present in real scenes, such as 'sofa' and 'table', we can easily generate corresponding descriptive text prompts for each category. We have compiled an extensive collection of prompt sets for each category, guaranteeing a thorough representation of the diverse range of objects within each category. These prompts are translated into 3D forms via the Prompt-to-Mesh (P2M) module. Here, we do not introduce a novel 3D shape generation modality but rather employ a mature 3D generative model Shap-E to process textual prompts into 3D meshes. For architecture details, please refer to Jun and Nichol [13]. Benefiting from the diversity of prompts, a vast and varied 3D shape library is successfully synthesized. The text's inherent diversity and expandability circumvent the constraints typically associated with finite 3D shape datasets. For instance, while ModelNet40 [53] includes only 109 variations in the "door" category, we can expand this category to any desired quantity in our own 3D shape library by increasing the number of prompts, as depicted in Fig. 5.

3D shape assembly. After generating the 3D shape repository, we construct a virtual scene to replicate the layout of a real scene using position-level ground truth labels. This process involves four steps:

1) Segmentation of the real scene. To achieve a reasonable and realistic layout, we first use a normal-based graph cut method [9] to segment the mesh version of the real scene. This segmentation breaks down the point cloud in the real scene into geometric plane components. Each segmented set of points represents a *seg*.

2) Initial generation of the virtual scene. We form the initial layout of virtual scenes by selecting the most matching 3D shapes to replace objects in real-world scenes. In detail, to approximate the shape of an object, we analyze the shapes of all components within a certain distance from the object's center point. This helps us find similar shapes to replace them. We obtain weak labeling information from real scenes for each object in the form of [(x, y, z), semantic_id]. We also calculate the 3D average size $(dx_{avg}, dy_{avg}, dz_{avg})$ for objects of each category, and $radius_{avg} = (dx_{avg} + dy_{avg})/2$. Points within a sphere of radius $\alpha \cdot radius_{avg}$ around the object's center point are collected from the scene point cloud. We increment the sphere's radius gradually until a sufficient number of points are gathered. With these points, we then determine the *seg* they belong to. We calculate the center and the minimum-maximum values in three coordinate dimensions for each *seg*. Segments with excessively large spans in the z-dimension or that are too far from the object's center in the xy-plane are removed. Next, we collect all relevant point clouds based on all *segs* of an object. We project the points onto the xy-plane and find the Minimum Enclosing Rectangle (MER), represented as (l, s, θ), where l, s, θ denote the length, width, and the rotation angle of the rectangle, respectively. We search for the 3D shape in our 3D library that is closest in terms of the ls_ratio.

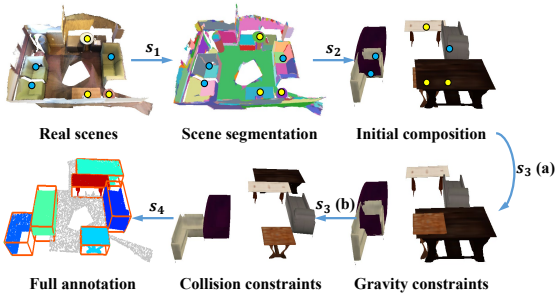

Figure 3. Illustration of the 3D Shape Assembly Process in SSG. $\mathbf{s_1}$**: Scene Segmentation**, where the structure of the real scene is identified; $\mathbf{s_2}$**: Preliminary Composition**, showcasing the initial arrangement of 3D shapes; $\mathbf{s_3}$ $(\mathbf{a})$ and $\mathbf{s_3}$ $(\mathbf{b})$**: Fine-tuning** with gravity and collision constraints to closely mimic real-world scene layouts; $\mathbf{s_4}$**: Computing Bounding Boxes** for each instance.

3) Adjustment of the virtual scene layout. To achieve the final scene layout, we apply gravity constraints and collision constraints based on location information and support relationships between objects. Compared to the point-version shapes, our mesh-version shapes enable more efficient and accurate adjustment of object positions.

4) Generation of ground-truth bounding box. Due to the explicit knowledge of the generated 3D shapes, the full annotation of the synthetic scenes can be easily generated. First, individual instance objects within the point cloud data are identified by analyzing the instance labels, with each label corresponding to a different object. Following identification, all points comprising an instance object's point cloud are extracted. For each object's point cloud, an axis-aligned bounding box is calculated. Each bounding box is assigned a semantic label indicating the object's category, such as "chair" or "table".

The process of 3D shape assembly is visually presented in Fig. 3. See more details in the supplementary.

3.3. Prototypical Proposal Feature Alignment

To facilitate weakly supervised object detection, our approach introduces the Prototypical Proposal Feature Alignment (PPFA) module, designed to leverage the robust supervision of synthetic scenes. By exploiting the structural consistency between synthetic and actual scenes, the PPFA module minimizes the domain distribution distance within

the feature space of prototype proposals. This enables the detection model to acquire discriminative capabilities from merely box-level annotations in real-world datasets.

Formally, denote the synthetic scene composed of m sampled 3D shapes by P_s=$\{o_1^s, o_2^s, \ldots, o_m^s\}$. Similarly, the real scene composed of n objects is denoted as P_r=$\{o_1^r, o_2^r, \ldots, o_n^r\}$. Both P_s and P_r are forwarded to the proposal feature extractor F to generate per-proposal features of two scenes $I_s = F(P_s) \in \mathbb{R}^{N\times d}$ and $I_r = F(P_r) \in \mathbb{R}^{N\times d}$. N is the predicted proposal number and d is denoted by the feature dimension.

For the synthetic scene, the prototype of the c^{th} class for the objects can be calculated as :

$$\mu_s'^c = \frac{1}{|\Lambda^c|} \sum_{i\in\{1,2,\cdots,N\}} \mathbb{I}(M_{s,i} = c) I_{s,i}, \tag{1}$$

where $I_{s,i} \in \mathbb{R}^d$ is the i_{th} proposal feature and $\mathbb{I}$ is an indicator function. $|\Lambda^c|$ denotes the number of all the proposal features belonging to the c^{th} class. $M_{s,i}$ represents the category of the i_{th} proposal in the synthetic scene. In reality, we regard the category of 3D object closest to the predicted box center coordinate as the class label of the proposal.

Similarly, denote the class label of the i_{th} proposal by $M_{r,i}$, prototype of the c_{th} class for the real scene can be defined as:

$$\mu_r'^c = \frac{1}{|\Lambda^c|} \sum_{i\in\{1,2,\cdots,N\}} \mathbb{I}(M_{r,i} = c) I_{r,i}, \tag{2}$$

In addition, in order to avoid introducing invalid proposal features, we introduce a foreground proposal mask to filter out features that are more likely to belong to foreground proposals to participate in feature alignment. Specifically, to delineate the likelihood of each predicted object center being a valid detection, the mask m_i of the i_{th} proposal is defined as:

$$m_i = \mathbb{I}(\min_j d(\mathbf{v}_i, \mathbf{c}_j) < \tau) \tag{3}$$

where $d(\mathbf{v}_i, \mathbf{c}_j)$ is the Euclidean distance between the predicted object center $\mathbf{v}_i$ and the ground truth center $\mathbf{c}_j$. Besides, τ is a predefined proximity threshold. In our work, τ is set to 0.3.

Therefore, the masked prototype of the c_{th} class $\mu_{mr}'^c$ for the real scene can be defined as:

$$\mu_{mr}'^c = \frac{1}{|\Lambda^c|} \sum_{i\in\{1,2,\cdots,N\}} m_i \mathbb{I}(M_{r,i} = c) I_{r,i}, \tag{4}$$

Similarly, for the target domain, the mask prototype of the synthetic scene $\mu_{ms}'^c$ is also operated in the same way.

To align the feature distribution across domains, we minimize the L_1 distance between each prototype of the same class of the synthetic and real scenes. The prototypical domain adaptation objective can be defined as:

$$\mathcal{L}_D = \sum_{c=1}^{C} \|\mu_{mr}'^c - \mu_{ms}'^c\|_1 . \tag{5}$$

In implementation, such calculation is increasingly computationally intensive as the training batches accumulate. To alleviate the issues, we estimate the prototypes with an online estimation algorithm following the same strategy as in [15, 21]. Specifically, the updated prototype of the synthetic scene after the t_{th} training iteration is calculated as:

$$\mu_{(s,t)}^c = \frac{p_{(s,t-1)}^c \mu_{m,(s,t-1)}^c + q_{(s,t)}^c \mu_{m,(s,t)}'^c}{p_{(s,t-1)}^c + q_{(s,t)}^c}, \tag{6}$$

where $p_{(s,t-1)}^c$ denotes the total number of the masked proposals belonging to c^{th} category during previous $t-1$ iterations and $q_{(s,t)}^c$ represents the number of the masked proposals belonging to k^{th} category in the current training batch.

Naturally, the updated prototype of the synthetic scene at the t_{th} mini-batch can be written as:

$$\mu_{(r,t)}^c = \frac{p_{(r,t-1)}^c \mu_{m,(r,t-1)}^c + q_{(r,t)}^c \mu_{m,(r,t)}'^c}{p_{(r,t-1)}^c + q_{(r,t)}^c}, \tag{7}$$

where $p_{(r,t-1)}^c$ represents the number of the masked proposal belonging to c^{th} category accumulated in history and $q_{(s,t)}^c$ is the proposal number of c^{th} category during the current iteration.

According to Eq. 6 and 7, calculating the updated prototype only requires features of the current batch and the prototype of the last step, thus avoiding the heavy computational and memory overhead.

Therefore, the final migration loss can be written as:

$$\mathcal{L}_D = \sum_{c=1}^{C} \|\mu_r^c - \mu_s^c\|_1 . \tag{8}$$

3.4. Training Objective

Our object detection framework is trained using both synthetic and real-world scenes, each with its own supervisory nuances. The loss function for synthetic scenes, $\mathcal{L}_S$, incorporates objectness, bounding box, and semantic classification losses, paralleling fully-supervised methods [34].

For real scenes with box-level annotations, the adapted loss function $\mathcal{L}_R$ retains objectness and semantic classification components but simplifies bounding box regression due to incomplete geometric annotations.

Additionally, a domain adaptation loss $\mathcal{L}_{DA}$, introduced through the PPFA module, augments the training process to improve feature discriminability across domains. This

Detector	Setting	Method	batht.	bed	bench	bsf.	bot.	chair	cup	curt.	desk	door	dres.	keyb.	lamp	lapt.	monit.	n.s.	plant	sofa	stool	table	toil.	ward.	mAP@0.25
VoteNet	*full sup.*	FSB	66.8	86.2	24.4	55.6	0.0	88.3	0.0	48.5	62.8	45.8	24.1	0.1	47.2	5.2	62.1	73.2	13.4	88.7	35.1	62.6	94.6	7.8	45.1
	weak sup.	WSB	21.9	46.9	0.3	2.3	0.0	53.7	0.0	0.9	32.1	1.0	6.6	0.1	0.2	0.1	1.8	53.6	0.1	57.0	4.6	6.4	19.7	0.0	14.1
		WSBP	38.5	83.1	16.7	33.2	**0.1**	75.5	0	9.0	53.2	14.7	**22.1**	0.0	26.2	**8.5**	**64.9**	**82.1**	14.1	82.5	**37.3**	41.6	86.1	4.0	36.1
		BR_P	50.8	70.4	10.7	29.1	0.0	65.6	0.0	**14.6**	53.9	9.8	8.5	0.3	34.5	4.2	34.9	68.9	5.7	70.9	22.5	35.3	88.2	**9.2**	31.3
		BR_M	57.1	80.4	14.3	31.7	0.0	**77.4**	0.0	13.2	49.7	11.3	14.8	**1.0**	**43.5**	6.0	56.5	65.0	10.6	80.2	26.9	**44.2**	91.4	6.5	35.5
		Ours	**59.7**	**83.7**	**35.9**	**40.5**	**0.1**	74.8	0.0	9.1	**60.1**	**21.7**	19.7	0.0	34.5	5.4	58.2	59.5	**17.2**	**84.0**	22.4	39.4	**95.4**	2.5	**37.4**
GroupFree3D	*full sup.*	FSB	86.2	87.5	16.3	49.6	0.6	92.5	0.0	70.9	78.5	53.5	56.0	6.4	68.2	11.5	81.5	88.5	15.2	88.2	45.6	65.0	99.7	31.2	54.2
	weak sup.	WSB	75.0	75.7	4.3	17.2	0.0	81.4	0.0	3.5	34.0	4.7	3.2	2.1	46.6	3.3	45.8	52.8	8.3	71.0	15.7	18.1	90.8	0.7	29.7
		WSBP	75.0	91.3	21.1	26.9	**7.9**	82.9	1.2	5.3	56.0	6.8	45.4	26.9	56.3	31.3	58.4	60.7	4.0	86.8	39.1	33.5	92.4	3.0	41.5
		BR_P	72.3	73.5	**45.8**	27.7	0.0	77.2	**8.2**	30.8	35.0	17.8	51.7	0.3	64.2	25.0	63.5	66.6	23.8	86.7	33.9	37.6	98.3	5.2	43.0
		BR_M	**85.3**	90.9	8.8	34.3	1.9	80.0	7.7	24.7	**58.0**	20.8	45.4	**31.3**	**64.4**	25.8	67.5	76.7	**27.3**	91.4	43.3	**46.7**	94.8	8.3	47.1
		Ours	76.2	**96.1**	5.8	**40.5**	0.0	**83.5**	7.7	**50.2**	55.4	**29.2**	**52.4**	0.0	57.5	**42.7**	**80.2**	**82.9**	21.8	**92.3**	**63.5**	45.5	**100.0**	**42.3**	**51.2**

Table 1. 3D object detection results on ScanNet validation set, evaluated using mAP@0.25. Experiments were conducted with VoteNet and GroupFree3D detectors. Following Xu et al. [54], we use 22 categories to assess performance on ScanNet. Best scores in **bold**.

combined loss framework effectively supports learning in scenarios with limited annotation detail.

The overall loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_R + \lambda_S \mathcal{L}_S + \lambda_D \mathcal{L}_D, \tag{9}$$

where λ_S and λ_D are trade-off parameters that weigh the importance of the corresponding loss terms.

4. Experiments

In this section, we perform experiments to validate the effectiveness of the proposed weakly supervised detection method. Firstly, we outline the experimental details. Then we provide examples of the generated synthetic scenes and present the detection results of our method. Finally, we conduct an analysis of our method via ablation studies.

4.1. Experimental Setup

Datasets. We conducted a series of experiments utilizing the ScanNet dataset [8]. The meticulously annotated indoor scene dataset includes 1201 training and 312 validation scenes but lacks manually labeled bounding boxes, necessitating the computation of axis-aligned bounding boxes. We meticulously evaluated these bounding box predictions on the validation set. Our model training was based solely on supervision from the object's center coordinates and category data. Following the framework of the weakly supervised method [54], we assessed our method's detection performance on 22 indoor object categories.

Implementation details. In our study, we utilize VoteNet [34] with a PointNet++ backbone [32] and GroupFree3D [22] as foundational object detectors, implemented using PyTorch. For VoteNet, the box proposal features I with $d = 128$ are derived from MLP_1 layer outputs in the proposal module, setting proposal count (N) to 256. During the training phase, we set $\lambda_D = 0.1$ and $\lambda_S = 0.05$. For GroupFree3D which has several decoders and each one outputs a stage of proposal features, we choose features of the last stage as the box proposal features I and $d = 288$. During the training phase, we set $\lambda_D = 10$ and $\lambda_S = 0.5$. Details on ablation experiments regarding loss weights are provided in the supplementary material.

In the Synthetic Scene Generation stage, We first generated a prompt set around 22 types of indoor objects. The prompt set is randomly generated using the large language model ChatGPT [29]. To get the 3D shape repository, the latest 3D generator Shap-E [13] is used to transfer the prompts to the 3D meshes. Furthermore, we have performed several rounds of scene enhancement in our virtual dataset. Initially, we simulated real-world scenes to create a foundational virtual dataset, and then randomly inserted indoor objects to increase complexity. Each enhancement round augmented a fixed set of 233 scenes, with a total of 8 rounds conducted.

4.2. Results and Analysis

Compared methods. To evaluate the performance of our weakly-supervised object detection method, we conduct comparisons across different settings and with other weakly-supervised approaches. We chose VoteNet [34] and GroupFree3D [22] as baseline detectors.

The weakly-supervised settings include: (i) WSB: Weakly Supervised Baseline which is the baseline detector trained on ScanNet with position-level annotations. (ii) WSBP: Weakly Supervised Baseline pre-trained on the synthetic scenes based on ModelNet40. We also give the result of the fully-supervised setting FSB, which is the Fully Supervised Baseline which is the baseline detector trained with strong annotations. Besides, the weakly supervised networks compared are BR_P and BR_M, which are trained using the point cloud version and mesh version of BR respectively. They use virtual scenes generated by ModelNet40 as additional supervision.

Weakly-supervised object detection. The quantitative experimental results can be seen in Tab. 1.

On ScanNet, we evaluate detection performance across 22 classes, including bathtub, bed, bench, and chair, etc., using mAP@0.25 as the evaluation metric. The results for mAP@0.5 can be found in the supplementary material.

On both baseline object detectors, our method effectively shortens their performance differences in weakly

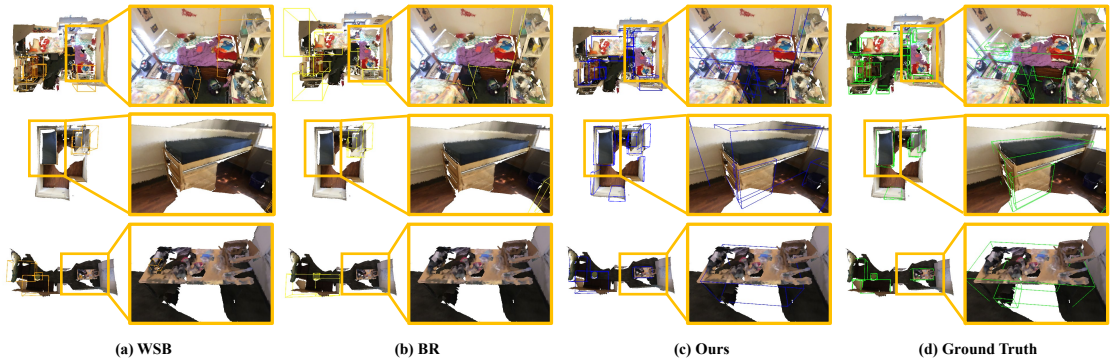

Figure 4. Visualization of detection results, with zoomed-in details on the right for clarity. Comparing (a) WSB, (b) BR, (c) Our Prompt3D, and (d) ground truth labels on the ScanNet validation dataset.

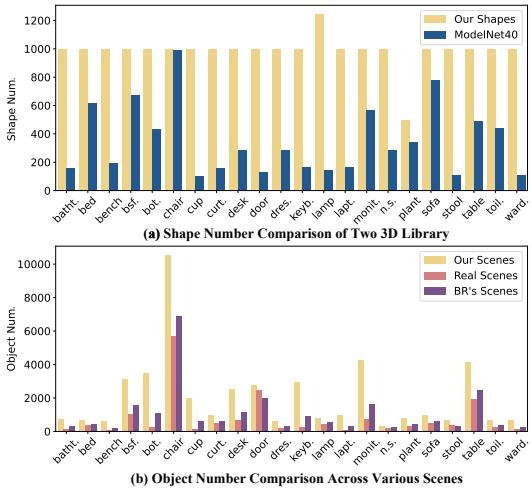

Figure 5. Scene Synthesis Details. Aligning with [54], we chose 22 categories for data generation. **1)** Our 3D library surpasses ModelNet40 [53] in shape count and boasts scalability through extra prompts, thus offering superior scene variety. **2)** Compared to ScanNet (real scene) [8] and BR [54], our virtual scenes feature a greater quantity and diversity of objects.

supervised settings and fully supervised settings. When using VoteNet as the detector, the performance disparity with fully supervised settings is trimmed to just 7.7% mAP@0.25. Compared with WSB, our method has more than 23.0% improvement. Compared with the previous sota BR [54], our method still gets nearly 2% improvement. In most categories, our method performs the best among all the weakly supervised methods.

When using GroupFree3D as the detector, the performance difference with the fully supervised setting is further reduced to only 3% mAP@0.25, reaching 51.2%. Compared with WSB, our method has more than 23.0% improvement. Our method obtains up to 4.1% improvement over BR. In most categories, our method performs best compared to other WSB settings, and in some categories (door, laptop, stool, toilet) it even far outperforms the FSB setting. This further validates the effectiveness

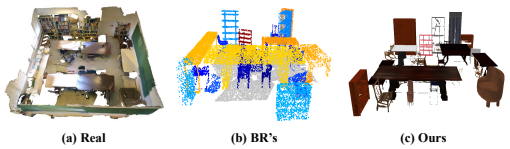

Figure 6. Synthetic scenes visualization: **(a)** Real scene; **(b)** The scene generated by BR via ModelNet40; **(c)** Our virtual scene constructed using a set of 3D shapes that are generated from prompts.

of our weakly supervised method. The migration effect is markedly better with GroupFree3D as the detector compared to using VoteNet. This superiority is attributed to GroupFree3D's attention mechanism, which renders box features of the same category during migration more reliable than those obtained with VoteNet.

Furthermore, both the detectors pretrained with our synthetic scenes improve the ability to learn discriminative features under the weakly supervised setting. This could be attributed to the fact that our synthesized scenarios based on textual prompts have introduced diversity, thereby promoting the learning of network parameters during pre-training.

4.3. Visualization

Qualitative detection results. In Fig. 4, we visualize the detection results of WSB, BR, and our method, with GroupFree3D as the detector. Despite the complexity of certain cases that WSB and BR fail to localize correctly, our method consistently achieves successful detection. These results highlight the benefits of label enhancement through virtual scenes created with prompts.

Virtual scene evaluation. We construct synthetic scenes by generating 3D shapes from textual prompts. As shown in Fig. 5, our approach has the ability to generate a greater variety of 3D shapes per class, surpassing existing real or virtual datasets in terms of shape quantity. This achievement is facilitated by the cost-effective method of generating diverse 3D shapes from abundant prompts.

Fig. 6 showcases that our synthesized scenes offer a

more realistic visual appearance compared to BR's scenes created with ModelNet40. Additionally, our scenes include more training objects than both real scenes and those of BR, as quantitatively compared in Fig. 5.

4.4. Ablation Study

Network setting. Our ablation studies evaluate the impact of each module on the network, confirming the effectiveness of our two proposed modules, as presented in Tab. 2.

(i) SSG module. We explore the Synthetic Scene Generation module (SSG) from two angles: 3D shape repository selection and virtual scene assembly.

Firstly, we scrutinize the 3D shape repository. Our comparison of settings 1 and 3, or 2 and 4 in Tab. 2, reveals that our prompt-based 3D shape repository greatly enhances weakly-supervised performance of the detector compared to ModelNet40, thanks to its textual diversity.

Secondly, we analyze the methodology of constructing virtual scenes. The former method BR [54] bases the initial layout of virtual scenes on horizontal segments with significant area and height and primarily uses point cloud versions of 3D shapes for construction, which may miss finer details. In contrast, our approach starts the scene layout around the centroids of real scene objects, facilitating the use of 3D mesh shapes that better resemble real objects. Moreover, we opt for assembling using the mesh versions of shapes, resulting in virtual scenes with a more realistic appearance. When comparing similar settings (1 and 2, or 3 and 4) within the same 3D shape repository, our assembly method outperforms BR in improving the performance of weakly-supervised networks.

(ii) Domain adaptation. After establishing an effective virtual scene generation technique, we then focus on optimally transferring the knowledge from virtual scenes to real-world scenarios.

Common transfer methods, like Adversarial Feature Alignment (AFA), align features adversarially, with a discriminator predicting feature domains to produce indistinguishable features. Conversely, our method aligns feature prototypes of the same category proposals to transfer knowledge from virtual scenes to real datasets. As shown in settings 4 and 5 in Tab. 2, our transfer algorithm, when using a constant virtual dataset, enhances object detection model transferability more effectively than AFA. PPFA leverages similarities between virtual and real scenes and utilizes category-specific features, leading to more robust and accurate real-world object detection.

Feature selection for PPFA. To enhance the effective transfer of valuable knowledge from synthetic to real scenes, our PPFA module is designed to align the distribution of bounding box proposal features. However, determining which box proposal features to use as the alignment features is challenging without thorough ablation studies. To this end, we conduct a feature-level ablation study on two different object detection networks. The corresponding results are displayed in Tab. 3 and Tab. 4.

Setting	SSG				DA		mAP@0.25
	3D Shape		Scene Assembly		AFA	PPFA	
	ModelNet40	Ours	BR	Ours			
1	✔		✔			✔	30.6
2	✔			✔		✔	35.0
3		✔	✔			✔	31.1
4		✔		✔		✔	**37.4**
5		✔		✔	✔		35.5

Table 2. Ablation study focusing on network settings, using VoteNet as the detector.

Layer	Seed Feature	Vote Feature	Aggregate Vote Feature
best map@0.25	33.2	35.3	**37.4**

Table 3. Selection of features for migration on VoteNet.

Layer	0_head	1_head	2_head	3_head	4_head	5_head
best map@0.25	44.3	47.4	48.7	48.3	49.4	**51.2**

Table 4. Selection of features for migration on GroupFree3D.

For VoteNet, the features aligned within the bounding boxes are chosen from the indirect features generated by VoteNet. There are three options: seed features, voting features, and aggregated voting features, corresponding to the outputs of the backbone module, voting module, and proposal module, respectively. Under the same experimental setup, the network aligning the distribution of aggregated voting features achieves the best detection performance. Hence, we select aggregated voting features as the alignment proposal features.

In the case of GroupFree3D, which possesses six decoder stages, each stage outputs a kond of proposal feature. To distinguish, we sequentially name the six detection heads as: 0_head, 1_head, 2_head, 3_head, 4_head, and 5_head. As illustrated in Tab. 4, under identical experimental settings, aligning features from the final stage 5_head yields the best weakly-supervised performance. Therefore, we choose the features generated by 5_head as our bounding box proposal features.

5. Conclusion

In this paper, we propose a novel weakly-supervised object detection approach, named Random Prompt Assisted Weakly-Supervised 3D Object Detection (Prompt3D). To mitigate the impact of weak labeling on network training, we propose an SSG module, which generates synthesized scenes with the 3D shapes generated by Shap-E using random prompts, offering strong supervision with a domain gap from real scenes. To bridge this gap, the PPFA module is designed to transfer knowledge via aligning prototypical features across the synthesized scenes and real scenes. Extensive experiments on the widely used indoor dataset ScanNet validate the effectiveness of the proposed method.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 3

[2] Xiwen Bi, Xiaohong Zhang, Shidong Wang, and Haofeng Zhang. Entropy-weighted reconstruction adversary and curriculum pseudo labeling for domain adaptation in semantic segmentation. *Neurocomputing*, 506:277–289, 2022. 3

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 2

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 3

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 3

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, pages 1907–1915, 2017. 2

[7] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, pages 1201–1210, 2015. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *ICCV*, pages 5828–5839, 2017. 6, 7

[9] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:167–181, 2004. 4

[10] Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and Jinchang Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2014. 2

[11] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *ICCV*, pages 5390–5399, 2019. 2

[12] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 3

[13] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 2, 3, 4, 6

[14] Kamran Kowsari and Manal H Alassaf. Weighted unsupervised learning for 3d object detection. *International Journal of Advanced Computer Science and Applications*, 2016. 2

[15] Shuang Li, Binhui Xie, Bin Zang, Chi Harold Liu, Xinjing Cheng, Ruigang Yang, and Guoren Wang. Semantic distribution-aware contrastive adaptation for semantic segmentation. *arXiv preprint arXiv:2105.05013*, 2021. 5

[16] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019. 3

[17] Yuanqi Li, Shun Liu, Xinran Yang, Jianwei Guo, Jie Guo, and Yanwen Guo. Surface and edge detection for primitive fitting of point clouds. In *ACM SIGGRAPH*, pages 1–10, 2023. 3

[18] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object instance mining for weakly supervised object detection. In *AAAI*, pages 11482–11489, 2020. 2

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pages 300–309, 2023. 3

[20] Kangcheng Liu, Yuzhi Zhao, Zhi Gao, and Ben M Chen. Weaklabel3d-net: A complete framework for real-scene lidar point clouds weakly supervised multi-tasks understanding. In *ICRA*, pages 5108–5115. IEEE, 2022. 1

[21] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *ICCV*, pages 8801–8811, 2021. 5

[22] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. 2021 ieee. In *ICCV*, pages 2929–2938, 2021. 6

[23] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Luc Van Gool, and Dengxin Dai. Weakly supervised 3d object detection from lidar point cloud. In *ECCV*, 2020. 1, 2

[24] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Yunde Jia, and Luc Van Gool. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *PAMI*, 2021. 1, 2

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[26] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *ICCV*, pages 2906–2917, 2021. 1

[27] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia*, pages 1–8, 2022. 3

[28] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[29] OpenAI. Gpt-4 technical report, 2023. 6

[30] Liang Peng, Senbo Yan, Boxi Wu, Zheng Yang, Xiaofei He, and Deng Cai. Weakm3d: Towards weakly supervised monocular 3d object detection. *arXiv preprint arXiv:2203.08332*, 2022. 1

[31] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 6

[33] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, pages 918–927, 2018. 1, 2

[34] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 5, 6

[35] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*, pages 8851–8858, 2019. 1, 2

[36] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. In *CVPR*, pages 7615–7623, 2019. 1, 2

[37] Zengyi Qin, Jinglu Wang, and Yan Lu. Weakly supervised 3d object detection from point clouds. In *ACM*, pages 4144–4152, 2020. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[39] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *ICCV*, pages 3283–3292, 2021. 1, 3

[40] Dayong Ren, Zhengyi Wu, Jiawei Li, Piaopiao Yu, Jie Guo, Mingqiang Wei, and Yanwen Guo. Point attention network for point cloud semantic segmentation. *Science China Information Sciences*, 65(9):1–14, 2022. 2

[41] Dayong Ren, Jiawei Li, Zhengyi Wu, Jie Guo, Mingqiang Wei, and Yanwen Guo. Mffnet: multimodal feature fusion network for point cloud semantic segmentation. *The Visual Computer*, pages 1–13, 2023.

[42] Dayong Ren, Zhe Ma, Yuanpei Chen, Weihang Peng, Xiaode Liu, Yuhan Zhang, and Yufei Guo. Spiking pointnet: Spiking neural networks for point clouds. *arXiv preprint arXiv:2310.06232*, 2023. 2

[43] Zhongzheng Ren, Ishan Misra, Alexander G Schwing, and Rohit Girdhar. 3d spatial recognition without spatially labeled 3d. In *CVPR*, pages 13204–13213, 2021. 1, 2

[44] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, pages 18603–18613, 2022. 3

[45] Enver Sangineto, Moin Nabi, Dubravko Culibrk, and Nicu Sebe. Self paced deep learning for weakly supervised object detection. *PAMI*, 41(3):712–725, 2018. 2

[46] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *NeurIPS*, 35:33999–34011, 2022. 3

[47] Feifei Shao, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao. Deep learning for weakly-supervised object detection and localization: A survey. *Neurocomputing*, 2022. 1

[48] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 1

[49] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, pages 2843–2851, 2017. 2

[50] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *PAMI*, 42(1):176–191, 2018. 2

[51] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *IROS*, pages 1742–1749. IEEE, 2019. 1

[52] Guanqun Wei, Zhiqiang Wei, Lei Huang, Jie Nie, and Xiaojing Li. Center-aligned domain adaptation network for image classification. *Expert Systems with Applications*, 168: 114381, 2021. 3

[53] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 2, 4, 7

[54] Xiuwei Xu, Yifan Wang, Yu Zheng, Yongming Rao, Jie Zhou, and Jiwen Lu. Back to reality: Weakly-supervised 3d object detection with shape-guided label enhancement. In *CVPR*, pages 8438–8447, 2022. 1, 2, 3, 6, 7, 8

[55] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, pages 1951–1960, 2019. 2

[56] Xiaohong Zhang, Haofeng Zhang, Jianfeng Lu, Ling Shao, and Jingyu Yang. Target-targeted domain adaptation for unsupervised semantic segmentation. In *ICRA*, pages 13560–13566, 2021. 3

[57] Xiaohong Zhang, Yi Chen, Ziyi Shen, Yuming Shen, Haofeng Zhang, and Yudong Zhang. Confidence-and-refinement adaptation model for cross-domain semantic segmentation. *IEEE TITS*, 23(7):9529–9542, 2022.

[58] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, pages 2020–2030, 2017. 3

[59] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018. 2

[60] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems*, 32(4):1713–1722, 2020. 3