IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

CMG3D: Compensation Towards Modality Gap for Open-vocabulary Indoor 3D Object Detection

Sheng Zhang, Lian Huai, Yuyu Liu, Xingqun Jiang

***Abstract*— For open-vocabulary indoor three-dimensional (3D) object detection (OVI3DOD), there is a gap between the image and the point cloud for indoor scenes, especially on distant objects. However, existing algorithms ignore this problem, which weakens the detection performance. Therefore, we propose Compensation towards the Modal Gap for open-vocabulary indoor 3D object detection (CMG3D). CMG3D consists of three modules: multimodal compensation (MC), object proposal filtering (OPF) and pseudo label refinement and generation (PLRG). In the MC, features from images are converted into the pseudo voxel space and then summed with the voxel space of the point cloud, which is used to compensate for the modality gap, while the OPF filters the object proposals to avoid confusion between the foreground and background. Finally, in the PLRG, the predictions from the two-dimensional (2D) detector are refined by the multimodal large language model (LLM) SigLIP and then transformed into 3D pseudo labels for the training process. Finally, we evaluate CMG3D on two indoor datasets, SUN RGB-D and ScanNet, and achieve state-of-the-art results.**

I. INTRODUCTION

Three-dimensional (3D) object detection has been widely applied in intelligent robotics and autonomous driving [1] and in both indoor and outdoor scenes [2]. Indoor 3D object detection is based mainly on point clouds captured by RGB-D cameras. However, the high cost of acquisition for 3D data means that the existing algorithms have limited detection categories, as new categories require relabeling and retraining. The open-vocabulary learning method can recognize unseen categories beyond the base categories by pretraining on data from images and captions and then retraining with the base category [3]. To solve the problem of limited categories in existing indoor 3D object detection methods, open-vocabulary indoor 3D object detection (OVI3DOD) has attracted much attention from both academia and industry.

OVI3DOD methods can be categorized into single-modality methods and multimodal methods [4]. Single-modality methods include image-based methods [5] and point cloud-based methods [6]. The multimodal method uses a point cloud and an image as the inputs [7]. For the image-based method, the 2D detection results are first obtained via the open-vocabulary two-dimensional (2D) detector, and then the 3D detection results are obtained via 2D-3D conversion. For the point cloud-based method, the learning of the semantic space of both the point cloud and text is realized via cross-domain contrastive learning. For the multimodal method, more accurate detection results are obtained by combining 2D semantic prior knowledge and 3D spatial information. Most of the current studies focus on multimodal methods, where the features of point clouds and images come from different modalities and there is an obvious gap in multimodal features.

OVI3DOD is trained with an open-vocabulary 2D detector to obtain 3D pseudo labels from 2D-3D conversion. The 3D predictions need to match with the 3D pseudo labels, and the matching influences the detection performance. Near objects can be effectively recognized from both the point cloud and the image. While far objects can be effectively recognized from the image, it is more difficult to recognize them from the point cloud. Therefore, there is a modality gap between the point cloud and the image, especially on far objects. This gap affects the matching between 3D predictions and pseudo labels. However, existing algorithms ignore this problem. For this reason, we propose Compensation towards the Modality Gap for open-vocabulary indoor 3D object detection (CMG3D).

The CMG3D module primarily includes multimodal compensation (MC), object proposal filtering (OPF) and pseudo-label refinement and generation (PLRG) modules. For the MC, there is a gap between the image and the point cloud, especially on the far objects. Therefore, the pseudo 3D data converted from the image are used to compensate for the far objects. For the OPF, noise in point clouds can weaken the performance of 3D detection. To avoid confusion between the foreground and background, we filter the proposals with a threshold. Finally, for the PLRG, high-quality pseudo labels can increase the detection performance. We refine the detection results of open-vocabulary 2D object detection (OV2DOD) with the multimodal large language model (LLM) SigLIP, thus eliminating the influence of erroneous pseudo labels on the detection results. The innovations by this work are summarized as follows:

(1) To eliminate the modality gap between the point cloud and image, we propose the MC, in which features from different modalities are converted into a unified voxel space and then summed, thus enriching the features of distant objects.

(2) To obtain high-quality pseudo labels, we propose the PLRG, which first refines the predictions from OV2DOD with the multimodal LLM SigLIP, and then converts the refined results to pseudo labels, thus eliminating the influence of erroneous 3D pseudo labels on the detection results.

Corresponding author: Sheng Zhang.

Sheng Zhang, Lian Huai, Yuyu Liu, Xingqun Jiang are with AIoT CTO, BOE Technology Group Co., Ltd., Beijing, China. (zhangsheng-cto@boe.com.cn)

II. RELATED WORKS

A. Open-vocabulary 2D Object Detection

For OV2DOD, the existing methods include the image-text pair-based method [8] and the LLM-based method [9]. For the image-text pair-based method, the representative algorithm is the OVR-CNN [8]. The LLM-based method uses the image-text embedding space of the LLM to improve the detection performance, where the representative algorithms include Region CLIP [9], GLIP [10], Grounding DINO [11] and YOLO-World [12]. For Region-CLIP, the algorithm first obtains proposals from the image, and then matches the proposals with text embedding via CLIP [13]. For GLIP, the algorithm extends the comparison learning method of CLIP from the image level to the object level, so that the condition in which the text corresponds to the whole image changes to the condition in which the text corresponds to a certain object or several objects in the image. Grounding DINO is based on DINO and expands DINO from closed-set detection to open-set detection. For example, given a textual cue, it automatically locates the object with the bounding box in the image. For YOLO-World, the algorithm applies YOLOv8 as the detection backbone, and employs the image-text modeling capability of CLIP to identify any specified object by text.

OV2DOD is the basis of OVI3DOD and OVI3DOD can be realized by migrating the semantic knowledge from OV2DOD to OVI3DOD due to the relatively limited 3D data.

B. Open-vocabulary 3D Object Detection

With the development of OV2DOD, some studies have begun to apply open-vocabulary technology to indoor 3D object detection. The OVI3DOD methods include the both the single-modality method and the multimodal method. The single-modality methods include image-based methods and point cloud-based methods. While, the multimodal method uses point clouds and images for training. The representative image-based and point cloud-based algorithms are OpenNav [5] and Object2Scene [6], respectively. The multimodal method applies point clouds and images as the inputs for training. Most existing studies on OVI3DOD have focused on multimodal methods, and representative algorithms include FM-OV3D [7] and CoDA [14]. For the FM-OV3D, the algorithm processes the features for point clouds and images by fusing multiple LLMs to improve the localization and recognition ability of OVI3DOD. For the CoDA, the algorithm simultaneously solves the two fundamental problems of OVI3DOD, which are the localization and categorization of new categories. CoDA can locate new 3D objects with limited base categories under a unified framework.

In point cloud and image-based methods, there is a large modality gap between point clouds and images, especially with respect to distant objects. However, existing studies ignore this problem.

III. METHOD

A. Pipeline of the CMG3D Algorithm

The pipeline of CMG3D is presented in Fig. 1. CMG3D includes three modules: the MC, OPF and PLRG. For the MC, the point clouds are converted into a point cloud voxel space and then processed by a voxel encoder, while the images are converted into the initial pseudo voxel space. Then, de-redundancy operations are performed on the initial pseudo voxel space to obtain the pseudo voxel space. Next, the pseudo voxel space is processed by a voxel encoder. Finally, we combine the two processed voxel datasets to obtain the final voxel space. When it comes to OPF, we first extract local features from the 3D backbone. Then, the OPF detection head processes these local features to provide us proposals (which include bounding boxes and their confidence values) and object features.

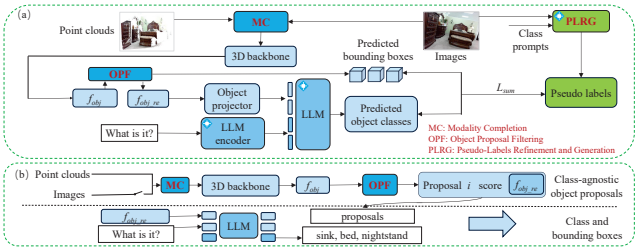

Fig. 1. CMG3D pipeline. CMG3D contains three modules: the MC, OPF and PLRG. (a) The CMG3D training network. (b) The CMG3D evaluation network.

Because of the noise in the 3D data, the 3D detector tends to confuse the foreground with the background, so the proposals are filtered via a threshold. When the confidence score is less than the set threshold, the proposal is considered background and removed. When the confidence score is greater than the set threshold, the proposal is considered foreground and is retained. The object features are subsequently used to obtain the object category via combination with the LLM. For the PLRG, the 2D detection result is first obtained by the open-vocabulary 2D detector, which is refined with the multimodal LLM SigLIP to eliminate the influence of incorrect categories. This 2D detection result, combined with multiscale image features, is processed via region of interest (ROI) pooling to obtain local object features, and then input to the monocular 3D detection head to obtain the 3D detection result. These results are then used as pseudo labels to train the class-agnostic 3D detector.

The CMG3D includes a training network and an evaluation network. As shown in Fig. 1 (a), the training network includes the MC, OPF and PLRG, while as shown in Fig. 1 (b), the evaluation network includes the MC and OPF. We adopt the FCAF3D [15] as the 3D detector when training and evaluating the network. FCAF3D uses a sparse convolutional network, including backbone, neck and head, and FCAF3D is an anchor-free 3D detector.

B. Multimodal Compensation

An illustration of the MC is presented in Fig. 2. The point clouds X_p are converted into a point cloud voxel space $V_p \in$

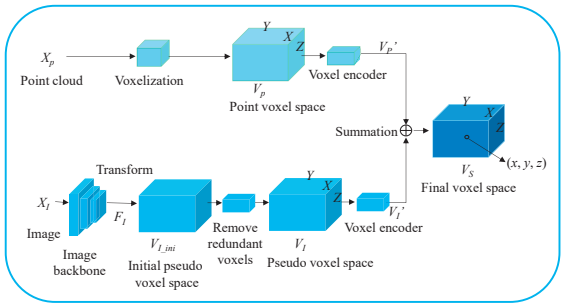

Fig. 2. Illustrations of the MC.

$\mathbb{R}^{X\times Y\times Z\times C}$ on the basis of the method presented in [15]. The point cloud voxel space is then processed by a voxel encoder, which can promote the interaction between the local features. The images are converted into initial pseudo voxel space, which is used as the base of the pseudo voxel space. We retain the voxels related to distant objects and remove the voxels related to near objects to obtain the pseudo voxel space, which is then processed by a voxel encoder. Finally, the two processed voxel spaces are added together to obtain the final voxel space.

For the images, we use ResNet50 [16] and the FPN [17] to extract features $f_I \in \mathbb{R}^{H\times W\times C}$ from the input images X_I. For the feature map, H is the height, and W is the width. f_I are then transformed to obtain the initial pseudo voxel space $V_{I\_ini} \in \mathbb{R}^{X\times Y\times Z\times C}$, which includes two steps: the calculation of the depth distribution $d_i \in \mathbb{R}^{D\times H\times W}$, and the multiplication of the image features and depth distribution. The depth distribution is calculated via the method presented in [18] as follows:

$$d_i = Softmax(Con(f_I(u,v)) \tag{1}$$

where, D is the perceptual limit of the indoor scene and where (u,v) represents the coordinates of the image plane.

(x,y,z) corresponds to the 3D coordinate of the sampling point, which is systematically computed at the centroid of each voxel bin. The image plane coordinates (u,v,d) are then determined through geometric projection of the voxel space coordinates (x,y,z) via the calibration matrix P, which establishes the coordinate transformation between the 3D space and the 2D image plane. The parameter d represents the depth value measured along the depth axis d_i which servers as the reference value for depth calculations. The multiplication of image features and depth distribution is calculated as follows:

$$V_{I\_ini}(x,y,z) = d_i(u,v,d) \times f_I(u,v) \tag{2}$$

where, $d_i(u,v,d)$ is the probability of voxel occupation for the multichannel image feature of $f_I \in \mathbb{R}^{H\times W\times C}$ at the 3D voxel coordinate (x,y,z) of the voxel space.

We perform de-redundancy operations for the initial voxel space to obtain the pseudo voxel space V_I. To interact with the local features of voxels V_P and V_I from these different modalities, a 3D convolutional block is used to process the voxel features from different modalities. The 3D convolutional block consists of 3D convolution, batch normalization [19] and ReLU [20] activation. Thus, the processed voxel spaces $V_P{}'$ and $V_I{}'$ are obtained, and the interaction is calculated as follows:

$$\begin{cases} V_P{}' = ReLU(BN(Conv_{3D}(V_p))) \\ V_I{}' = ReLU(BN(Conv_{3D}(V_I))) \end{cases} \tag{3}$$

The voxel spaces of the two different modalities are summed to obtain the final voxel space V_s as follows:

$$V_s = V_P{}' + V_I{}' \tag{4}$$

C. Object Proposal Filtering

The role of OPF is to generate preliminary detection results by first extracting the OPF extracts the results $\{b_i, f^i_{obj\_re}\}^{N_{obj}}_{i=1}$ through the local feature f_{obj}, and the transformation is calculated as follows:

$$\{b^i_{3D}, f^i_{obj\_re}\}^{N_{obj}}_{i=1} = OPF(f_{obj}) \tag{5}$$

where, b_i and $f^i_{obj\_re}$ are the bounding box and feature of the i-th object, respectively, and where N_{obj} denotes the quantity of objects.

The inherent sensor noise present in 3D data can induce ambiguity in the discrimination between foreground and background elements, leading to the generation of inaccurate object proposals by the detection algorithm. To filter the object proposals, the OPF is proposed in this work. In OPF, the object proposals $\{\hat{b}_i, \hat{o}_i\}^{N_q}_{i=1}$ with uncertain categories are obtained by processing the local features f_{obj} with several detection heads, where, $\hat{b}_i$ represents the predicted bounding box coordinates for the i-th detected object instance and where $\hat{o}_i$ quantifies the confidence score associated with the i-th detection proposals. Then, the results corresponding to confidence values below the threshold φ_{obj} are removed, the final category-agnostic detection result $\{b^i_{3D}, o_i\}^{N_{obj}}_{i=1}$ is obtained, and N_{obj} is the number of objects.

The object category c_i is determined through the LLM inference process, which operates as follows:

$$c_i = LLaMA\left(t_{obj}, projector\left(f^i_{obj\_re}\right)\right) \tag{6}$$

where t_{obj} represents the carefully engineered textual input provided to the LLaMA, which serves as the guiding prompt for generating the object-related responses. We use "What is it?" as the t_{obj}. The projected layer $projector(\cdot)$ implemented as a parameterized linear transformation, serves to map the extracted object features $\{f^i_{obj\_re}\}^{N_{obj}}_{i=1}$ into a latent space that is compatible with the semantic embedding space of the LLM. The training of the OPF and projected layers is guided by 3D pseudo labels that are generated via the PLRG, which serves as the supervisory signal for training these components.

In conventional closed-set object detection, confidence scores are naturally derived from supervised learning via ground-truth annotations. Conversely, open-vocabulary object detection paradigms often require the construction of supervisory signals through human-defined criteria or pre-specified confidence thresholds, as direct label-based supervision becomes infeasible due to the expanded vocabulary

scope. First, a bipartite matching [21] between the detection results $\{b^i_{3D}\}^{N_{obj}}_{i=1}$ and pseudo labels $\{\tilde{b}^i_{3D}\}^{\tilde{N}}_{i=1}$ is constructed on the basis of the IoU [22]. Then all class-agnostic detection results are matched with the pseudo labels as follows:

$$y_i = \begin{cases} 1 & \exists j, b_i \ and \ \tilde{b}_j \ are \ matched, \\ 0 & otherwise, \end{cases} \tag{7}$$

where, $y_i = 1$ signifies that the i-th proposal is classified to be positive, representing a foreground object. Conversely, when $y_i = 0$, the i-th proposal is negative, corresponding to a background object.

D. Pseudo Labels Refinement and Generation

An illustration of the MC is presented in Fig. 3. To obtain 3D pseudo labels, we refine the predictions from the open-vocabulary 2D detector and then transform the refined 2D predictions into 3D pseudo labels. We input image X_I and text X_T, then the open-vocabulary 2D detector can obtain 2D detection results $\{(b^i_{2D}, c^i_{2D})\}^N_{i=1}$. N is the number of 2D labels, and b^i_{2D} and c^i_{2D} represent the geometric parameters and categorical label of the i th detected bounding box in the 2D image space, respectively. The 2D detection results directly affect the quality of the pseudo labels. Furthermore, to obtain high-quality pseudo labels, the raw 2D detection outputs undergo a refinement process. The refinement is realized on the basis of the multimodal LLM SigLIP [23]. First, according to $\{(b^i_{2D}, c^i_{2D})\}^N_{i=1}$, the corresponding patches obtained from the image are $\{p_i\}^N_{i=1}$. To refine the detection results, we establish two sets of prompt templates $t^+(\cdot)$ and $t^-(\cdot)$ as follows:

$$\begin{cases} t^+(category) : It \ \ is \ a \ \{category\} \\ t^-(category) : It \ \ is \ not \ a \ \{category\} \end{cases} \tag{8}$$

The two templates, image patches $\{p_i\}^N_{i=1}$ and $\{c^i_{2D}\}^N_{i=1}$ are fed into SigLIP for the calculation of the confidence score, which is expressed as follows:

$$[\varphi^+_i, \varphi^-_i] = Softmax(SIGLIP(t^+(c_i), p_i), SIGLIP(t^-(c_i), p_i)) \tag{9}$$

where, φ^+_i signifies the confidence score indicating the probability of proposal p_i being classified as category c_i, while φ^-_i represents the complementary confidence score that indicates the likelihood that p_i dose not belong to the same category c_i. We keep the predictions that φ^+_i are above the threshold φ_{SIGLIP}, and we also use φ^-_i to assist the refinement for a few difficult predictions. Finally, we obtain the refined results $\{(\tilde{b}^i_{2D}, \tilde{c}^i_{2D})\}^{\tilde{N}}_{i=1}$, where $\tilde{N}$ is the number of 2D labels after refinement.

Moreover, the image X_I is processed with a visual backbone based on ViT [24] to obtain the features $f \in \mathbb{R}^{\frac{H}{14} \times \frac{W}{14} \times C}$. Subsequently, the enhanced feature pyramid network [25] is subsequently employed to process the features, thereby extracting multiscale features. The scaling factors for obtaining a multiscale hierarchical feature pyramid network are 0.5, 1 and 2. Then, the multiscale features and the refined single detection result $(\tilde{b}^i_{2D}, \tilde{c}^i_{2D})$ are used to obtain a 2D region proposal, which is processed by a Cube R-CNN [26] head to obtain a 3D bounding box $\{(\tilde{b}^i_{3D}, \tilde{c}^i_{3D})\}^{\tilde{N}}_{i=1}$. Finally, we

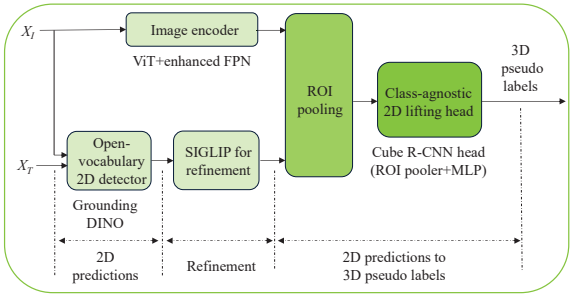

Fig. 3. Illustration of the PLRG.

adopt $\{(\tilde{b}^i_{3D}, \tilde{c}^i_{3D})\}^{\tilde{N}}_{i=1}$ as the 3D pseudo labels. The Cube R-CNN is a monocular 3D detector, and its head consists of an ROI pooler and an MLP.

E. Loss Function

The overall training objective is formulated as a multi-task loss function comprising three distinct components: the bounding box regression loss L_{bbox}, confidence prediction loss L_{conf}, and object classification loss L_{cls}. In this work, we apply the regression loss function $Loss_{reg}$ based on the IoU [15, 27] to calculate the bounding box regression loss L_{bbox}. The confidence loss L_{conf} is formulated to quantify the discrepancy between the predicted confidence scores and ground-truth objectness labels. It is computed via a binary cross-entropy (BCE) function across all candidate bounding boxes. Since the object category is obtained from the LLM, L_{cls} is derived through the maximization of the probability for the labeled text tokens. The classification loss L_{cls} is then calculated on the basis of text loss. In summary, the total loss function is calculated as follows:

$$L_{sum} = \xi_1 L_{bbox} + \xi_2 L_{conf} + \xi_3 L_{cls} \tag{10}$$

where, ξ_1, ξ_2 and ξ_3 are the weight coefficients in the total loss function.

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

In this letter, we implement comprehensive evaluations of CMG3D across two benchmark indoor datasets: SUN RGB-D [28] and ScanNet [29]. SUN RGB-D includes approximately 800 objects and 10,335 samples, of which 5,285 scenes are used for training, 5,050 scenes are used for testing. ScanNet comprises 1,513 3D indoor scenes, which are partitioned into 1,201 training scenes and 312 testing scenes. This dataset contains over 200 distinct object categories, providing a rich environment for evaluating semantic segmentation and object recognition algorithms in complex indoor settings. In this letter, mAP with an IoU threshold of 0.25 is used as the evaluation metric for detection performance. Futhermore, we follow the methods described in CoDA [14] and OV-3DET [30] for the experiments.

We follow the methods defined in CoDA to execute the experiments. Evaluations are performed on three distinct categories. The novel categories contain unseen object classes from the training set, and the base categories contain

TABLE I

PERFORMANCE OF CMG3D AND THE EXISTING METHODS WITH THE SUN RGB-D AND SCANNET. MODALITY DENOTES THE MODALITY OF THE INPUT DATA FOR THE EVALUATION. P REPRESENTS THE POINT CLOUDS, AND I REPRESENTS THE IMAGES.

Methods	Modality	SUN RGB-D			ScanNet		
		mAP_{novel}	mAP_{base}	mAP_{all}	mAP_{novel}	mAP_{base}	mAP_{all}
Det-PointCLIP [33]	P	0.09	5.04	1.17	0.13	2.38	0.50
Det-PointCLIPv2 [34]	P	0.12	4.82	1.14	0.13	1.75	0.40
Det-CLIP^2 [35]	P	0.88	22.74	5.63	0.14	1.76	0.40
3D-CLIP [13]	P+I	3.61	30.56	9.47	3.74	14.14	5.47
CoDA [14]	P	6.71	38.72	13.66	6.54	21.57	9.04
CoDAv2 [36]	P	9.17	42.04	16.31	9.12	23.35	11.49
CMG3D (ours)	P	10.54	49.33	19.34	13.62	31.54	16.27
	P+I	**13.89**	**50.41**	**21.74**	**16.32**	**32.94**	**18.93**

TABLE II

COMPARISON RESULTS OF CMG3D AND THE EXISTING METHODS WITH SINGLE CATEGORY ON SUN RGB-D.

Methods	mAP_{25}^{10cls}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk
OV-3DETIC [37]	13.03	43.97	6.17	0.89	45.75	2.26	8.22	0.02	8.32	0.07	14.60
FM-OV3D [7]	21.47	55.00	38.80	19.20	41.91	23.82	3.52	0.36	5.95	17.40	8.77
OV-3DET [30]	31.06	72.64	66.13	34.80	44.74	42.10	11.52	0.29	12.57	14.64	11.21
CMG3D (P)	36.24	72.72	72.43	42.25	50.98	48.32	17.64	4.32	17.26	19.17	17.34
CMG3D (P+I)	**38.63**	**74.23**	**73.78**	**45.13**	**53.26**	**51.54**	**19.11**	**8.34**	**20.78**	**20.81**	**19.32**
Methods	mAP_{25}^{20cls}	table	stand	cabinet	counter	bin	bookshelf	pillow	microwave	sink	stool
OV-3DET [30]	20.46	23.31	2.75	3.40	0.75	23.52	9.83	10.27	1.98	18.57	4.10
CMG3D (P)	26.65	30.13	10.89	12.81	8.24	30.46	15.65	17.34	10.67	19.15	15.20
CMG3D (P+I)	**28.98**	**32.44**	**12.74**	**15.39**	**12.54**	**32.12**	**17.54**	**18.12**	**13.43**	**21.36**	**17.55**

TABLE III

COMPARISON RESULTS OF CMG3D AND THE EXISTING METHODS WITH SINGLE CATEGORY ON SCANNET.

Methods	mAP_{25}^{10cls}	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DETIC [37]	12.65	48.99	2.63	7.27	18.64	2.77	14.34	2.35	4.54	3.93	21.08
FM-OV3D [7]	21.53	62.32	41.97	22.24	31.80	1.89	10.73	1.38	0.11	12.26	30.62
OV-3DET [30]	24.36	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA [14]	28.76	68.09	44.04	**28.72**	44.57	3.41	20.23	**5.32**	0.03	27.95	45.26
CoDAv2 [36]	30.06	**77.24**	43.96	15.05	53.27	11.37	19.36	1.42	0.11	34.42	44.38
CMG3D (P)	32.42	73.31	47.45	22.34	55.72	14.54	21.13	3.54	4.23	36.38	45.54
CMG3D (P+I)	**33.64**	73.69	**48.23**	23.33	**56.64**	**15.72**	**23.44**	4.94	**5.69**	**38.54**	**46.21**
Methods	mAP_{25}^{20cls}	bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET [30]	18.02	56.28	10.99	19.72	0.77	0.31	9.59	**10.53**	3.78	2.11	2.71
CoDA [14]	19.32	50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
CoDAv2 [36]	22.72	55.60	24.41	20.67	20.72	0.28	13.54	0.92	4.16	4.37	9.20
CMG3D (P)	26.43	58.42	30.51	25.33	24.98	3.41	**19.63**	6.87	9.23	9.54	**16.43**
CMG3D (P+I)	**27.57**	**59.23**	**32.42**	**26.45**	**25.77**	**4.56**	18.12	9.67	**10.26**	**12.59**	15.98

categories from the training set. All categories contain both novel categories and base categories. Furthermore, for the three categories, we select mAP_{novel}, mAP_{base} and mAP_{all} as metrics. In the SUN RGB-D dataset, the top-10 most frequent object categories are designated as base classes, whereas the other 46 categories are categorized as novel classes. Similarly, for the ScanNet dataset, the top-10 prevalent object types constitute the base classes, with the other 50 classes identified as novel entities.

We also follow the methods described in OV-3DET to perform the experiments, and select the top-10 and top-20 categories as novel classes on the two indoor datasets: SUN RGB-D and ScanNet. To later facilitate the description of the experimental results, the top-10 metrics and top-20 metrics are expressed as mAP_{25}^{10cls} and mAP_{25}^{20cls}.

B. Implemented Details

We implement CMG3D by employing the MMDetection3D [31] framework, and we use AdamW as the optimizer. The voxel size is set as 0.01 m. Furthermore, the training process is divided into two phases. In the first phase, we train the 3D backbone network and geometric prediction header over 400 epochs, with each GPU processing a batch size of 4. In the second phase, the object confidence prediction header and local projector are trained over 50 epochs, with each GPU processing a batch size of 2. The base learning rate is configured at $1e^{-4}$, with carefully calibrated weight coefficients applied to the loss function components to balance their contributions during optimization. The weight coefficients of the loss functions ξ_{conf}, ξ_1, ξ_2 and ξ_3 are set to 0.3, 3, 9 and 1, respectively. The thresholds φ_{obj}, φ_{SIGLIP}, φ^- and φ^+ are set to 0.1, 0.5, 0.25 and 0.6, respectively. In this letter, we choose FCAF3D to construct the 3D backbone network, the prediction heads of the bounding boxes and the confidence score. We adopt the Grounding DINO as an open-vocabulary 2D detector. For the LLM, we use LLaMA 3.1 [32] to construct the LLM, and use SigLIP to construct the PLRG. The experimental studies are conducted on NVIDIA A800 GPUs, leveraging their

advanced computational architecture.

C. Comparison with Existing Methods

We evaluate CMG3D on the SUN RGB-D and ScanNet datasets following the methods described in CoDA, with 56 categories and 60 categories for the SUN RGB-D and ScanNet datasets, respectively. Furthermore, we use mAP_{25} as the metric for evaluation, and the results of the comparison with the existing algorithms are shown in Table I. For CMG3D, we use two different kinds of inputs, one using only the point cloud as inputs, and the other using the point cloud and image as inputs. For ease of presentation, we simplify the CMG3D with the single-modality input as CMG3D(P), and CMG3D with the multimodal input as CMG3D(P+I). For CMG3D(P), the MC is applied during training and evaluation, the point clouds and images are applied for training, and only the point clouds are applied for evaluation. We then use the results of CMG3D(P) as the reference for CMG3D(P+I)

For CMG3D(P), the detection results exceed those of CoDAv2 and the other existing methods. For CMG3D(P), PLRG is used to obtain high-quality pseudo labels, and OPF is used to filter the low-quality proposals. For CMG3D(P+I), the detection results achieve a further improvement over CMG3D(P). For CMG3D(P+I), the detection performance of far objects is improved by complementing the point clouds related to these far objects with the features from the images, which further improves the OVI3DOD performance.

To further evaluate CMG3D, we use the experimental setting described in OV-3DET, corresponding to the top 10 categories and the top 20 categories, which are unseen. This evaluation employs mAP_{25} as the metric. The comparison results of individual categories with the SUN RGB-D and ScanNet datasets for CMG3D and the existing algorithms are provided as shown in Tables II and III. We Follow the Method Described in the OV-3DET [30] to Implement the Experiment. The comparison results indicate that CMG3D(P) achieves improvements in most of the categories over the existing methods, while CMG3D(P+I) achieves further improvement over CMG3D(P).

D. Ablation Studies

To assess how the various network components influence the effectiveness of CMG3D, we conduct component-wise ablation studies using the ScanNet dataset. We use the method described in CoDA to conduct an ablation study by designing 8 distinct method configurations to investigate component-wise contributions, and we establish different methods for comparison, where the MC, OPF and PLRG are not used in Method 0. Method 0 uses only the point cloud as input, with no filtering for the proposals, and the pseudo labels are not refined. 2D-3D spatial conversion of open-vocabulary 2D detection is implemented with the projection matrix and open-vocabulary 2D detection results. The individual components from the MC, OPF and PLRG are added to Methods 1, 2 and 3, respectively. While two components in the MC, OPF and PLRG are added in Methods 4, 5 and 6, respectively. In method 7, all of the components are included.

The experimental results indicate that the performances of Methods 1, 2 and 3 are better than that of Method 0, because the MC, OPF and PLRG are included into the Methods 1, 2 and 3, respectively. Among the Methods 1, 2 and 3, Method 1 achieves the maximum performance gains in mAP_{novel}, and Method 2 achieves the maximum performance gains in mAP_{base}. These results indicate that MC specializes in boosting the detection performance of unseen categories, and the OPF specializes in boosting the detection performance for the base categories. Methods 4, 5 and 6 yield further performance improvements over Methods 1, 2 and 3 due to the interactions among the different modules. Finally, in Method 7, all of the modules are included to CMG3D, which yielded the best results.

TABLE IV

ABLATION STUDY OF THE IMPACT OF DIFFERENT COMPONENTS OF THE CMG3D DETECTION PERFORMANCE ON SCANNET.

Methods	Modality	MC	OPF	PLRG	ScanNet mAP_{novel}	ScanNet mAP_{base}	ScanNet mAP_{all}
0	P				8.13	22.14	10.44
1	P+I	✓			14.27	27.52	16.42
2	P		✓		10.71	28.33	13.44
3	P			✓	12.54	24.11	14.56
4	P+I	✓	✓		14.97	32.15	17.14
5	P+I	✓		✓	15.97	30.35	17.89
6	P		✓	✓	12.77	29.43	15.12
7	P+I	✓	✓	✓	**16.32**	**32.94**	**18.93**

TABLE V

PERFORMANCE OF CMG3D AND THE EXISTING METHODS FOR FAR OBJECTS WITH SUN RGB-D AND SCANNET.

Methods	Modality	SUN RGB-D mAP_{novel}	SUN RGB-D mAP_{base}	SUN RGB-D mAP_{all}	ScanNet mAP_{novel}	ScanNet mAP_{base}	ScanNet mAP_{all}
CoDA [14]	P	2.18	30.60	7.32	2.72	15.35	4.46
CoDAv2 [36]	P	5.45	35.35	11.21	5.86	17.57	7.92
CMG3D(ours)	P	7.47	43.32	14.65	10.43	26.38	13.56
	P+I	**11.16**	**46.36**	**17.95**	**14.11**	**29.67**	**16.21**

TABLE VI

PERFORMANCE OF CMG3D AND THE EXISTING METHODS FOR FAR OBJECTS WITH SUN RGB-D AND SCANNET.

Methods	Modality	SUN RGB-D mAP_{25}^{10cls}	SUN RGB-D mAP_{25}^{20cls}	ScanNet mAP_{25}^{10cls}	ScanNet mAP_{25}^{20cls}
OV-3DET [30]	P	25.18	13.19	18.59	12.25
CoDA [14]	P	27.15	16.24	23.24	14.43
CoDAv2 [36]	P	29.47	18.73	25.16	18.84
CMG3D(ours)	P	32.13	22.47	28.35	22.38
	P+I	**35.63**	**25.55**	**30.16**	**24.44**

E. Detection Performance of Far objects

To evaluate the detection performance of CMG3D and the existing methods for distant objects, we implement ablation experiments with SUN RGB-D and ScanNet. We use the methods described in CoDA to conduct the experiments and two comparison methods are established: CoDA and CoDAv2. We also use the methods in OV-3DET to conduct the experiments, establishing three comparison methods in total: OV-3DET, CoDA and CoDAv2. We consider objects

with a distance greater than 2 m from the camera to be far objects. As shown in Tables V and VI, compared with the comparison methods, our proposed method remarkably enhances the detection performance for far objects. Compared with the results produced by CoDA and CoDAv2, the detection performance for distant objects obtained by CMG3D(P) is better. While compared with CMG3D(P), the detection performance of CMG3D(P+I) for far objects is further improved. These results are obtained because the MC enhances the learning of features for far objects in CMG3D(P) and CMG3D(P+I) during training, and image features promote further improvement by CMG3D(P+I).

V. CONCLUSION

In this work, we propose the CMG3D which consists of the MC, OPF and PLRG. For the MC, the features of far objects are compensated by the pseudo voxels obtained from the 2D images to eliminate the modality gap between the 2D and 3D data, which in turn improves the detection performance for far objects. We also apply OPF and PLRG to further improve the detection performance. To evaluate CMG3D, comparison experiments with existing algorithms are implemented on two indoor datasets, and under two different experimental settings, CMG3D achieves SOTA results. For fault mode analysis, we implement model inference experiments with our analysis and found that severe occlusion scenarios weaken the detection performance of far targets. In our follow-up work, we attempt to solve this problem through 3D reconstruction methods.

REFERENCES

[1] J. Qiao, B. Liu, J. Yang et al., "MonoSample: Synthetic 3D Data Augmentation Method in Monocular 3D Object Detection," *IEEE Robot. Automat. Lett.*, 2024.

[2] D. Rukhovich, A. Vorontsova, A. Konushin, "TR3D: Towards real-time indoor 3d object detection," in *Proc. IEEE Int. Conf. on Image Process.*, 2023, pp. 281-285.

[3] J. Cai, Y. He, W. Yuan, et al., "Open-Vocabulary Category-Level Object Pose and Size Estimation," *IEEE Robot. Automat. Lett.*, 2024.

[4] J. Wu, X. Li, S. Xu, et al., "Towards open vocabulary learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[5] P. Simonetto, A. Polato, F. Pasti, et al., "OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation," 2024, *arXiv preprint arXiv:2408.13936*.

[6] C. Zhu, W. Zhang, T. Wang, et al., "Object2scene: Putting objects in context for open-vocabulary 3D detection," 2023, *arXiv preprint arXiv:2309.09456*.

[7] D. Zhang, et al. "FM-OV3D: Foundation Model-Based Cross-Modal Knowledge Blending for Open-Vocabulary 3D Detection." In *Proc. AAAI Conf. on Artif. Intell.*, 2024.

[8] A. Zareian, K. D. Rosa, D. H. Hu, et al. "Open-vocabulary object detection using captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, PP. 14393-14402.

[9] Y. Zhong, J. Yang, P. Zhang, et al., "Regionclip: Region-based language-image pretraining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16793-16803.

[10] L. H. Li, P. Zhang, H. Zhang, et al., "Grounded language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10965-10975.

[11] S. Liu, Z. Zeng, T. Ren, et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 38-55.

[12] T. Cheng, L. Song, Y. Ge, et al., "Yolo-world: Real-time open-vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024: 16901-16911.

[13] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in *Proc. IEEE Int. Conf. Mach. Learn.*, 2021, pp. 8748-8763.

[14] Y. Cao, Y. Z, H. Xu, et al. "Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2024.

[15] D. Rukhovich, A. Vorontsova, A. Konushin, "Fcaf3d: Fully convolutional anchor-free 3d object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 477-493.

[16] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

[17] T. Y. Lin, P. Dollár, R. Girshick, et al., "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117-2125.

[18] J. Philion, S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 194-210.

[19] S. Santurkar, D. Tsipras, A. Ilyas, et al., "How does batch normalization help optimization?" in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, 31.

[20] Y. Li, Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, 30.

[21] N. Carion, F. Massa, G. Synnaeve, et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020: 213-229.

[22] D. Zhou, J. Fang, X. Song, et al., "Iou loss for 2d/3d object detection," in *Proc. Int. Conf. on 3D Vis.*, 2019: 85-94.

[23] X. Zhai, B. Mustafa, A. Kolesnikov, et al., "Sigmoid loss for language image pre-training," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 11975-11986.

[24] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv preprint arXiv:2010.11929*.

[25] Y. Zhao, F. Zhu, Y. Mi, et al., "Simple-FPN: An Image Anomaly Detection and Localization Network based on SimpleNet and Feature Pyramid," in *Proc. IEEE Int. Conf. Digital Twins and Parallel Intell.*, 2024, pp. 417-422.

[26] G. Brazil, A. Kumar, J. Straub, et al., "Omni3d: A large benchmark and model for 3d object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13154-13164.

[27] I. Misra, R. Girdhar, A. Joulin, "An end-to-end transformer model for 3d object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2906-2917.

[28] S. Song, S. P. Lichtenberg, J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567-576.

[29] A. Dai, A. X. Chang, M. Savva, et al., "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828-5839.

[30] Y. Lu, C. Xu, X. Wei, et al., "Open-vocabulary point-cloud object detection without 3d annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1190-1199.

[31] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," 2020. [Online]. Available: *https: //github.com/openmmlab/mmdetection3d*.

[32] A. Dubey, A. Jauhri, A. Pandey, et al., "The llama 3 herd of models," 2024, *arXiv preprint arXiv:2407.21783*.

[33] R. Zhang, Z. Guo, W. Zhang, et al., "Pointclip: Point cloud understanding by clip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8552-8562.

[34] X. Zhu, R. Zhang, B. He, et al. "Pointclip v2: Adapting clip for powerful 3d open-world learning," 2022, *arXiv preprint arXiv:2211.11682*.

[35] Y. Zeng, C. Jiang, J. Mao, et al., "CLIP2: Contrastive language-image-point pretraining from real-world point cloud data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15244-15253.

[36] Y. Cao, Y. Zeng, H. Xu, et al., "Collaborative Novel Object Discovery and Box-Guided Cross-Modal Alignment for Open-Vocabulary 3D Object Detection," 2024, *arXiv preprint arXiv:2406.00830*.

[37] Y. Lu, C. Xu, X. Wei, et al., "Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning," 2022, *arXiv preprint arXiv:2207.01987*.