Vision-Language Pre-training with Object Contrastive Learning for 3D Scene Understanding

Taolin Zhang[1*], Sunan He[2*], Tao Dai[3†], Zhi Wang[1], Bin Chen[4], Shu-Tao Xia[1, 5]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Department of Computer Science and Engineering , Hong Kong University of Science and Technology
[3]College of Computer Science and Software Engineering, Shenzhen University
[4]Harbin Institute of Technology, Shenzhen
[5]Research Center of Artifcial Intelligence, Peng Cheng Laboratory
zhangtlin3@gmail.com,sunan.he@connect.ust.hk,daitao.edu@gmail.com
wangzhi@sz.tsinghua.edu.cn,chenbin2021@hit.edu.cn,xiast@sz.tsinghua.edu.cn

Abstract

In recent years, vision language pre-training frameworks have made significant progress in natural language processing and computer vision, achieving remarkable performance improvement on various downstream tasks. However, when extended to point cloud data, existing works mainly focus on building task-specific models, and fail to extract universal 3D vision-language embedding that generalize well. We carefully investigate three common tasks in semantic 3D scene understanding, and derive key insights into the development of a pre-training model. Motivated by these observations, we propose a vision-language pre-training framework 3DVLP (3D vision-language pre-training with object contrastive learning), which transfers flexibly on 3D vision-language downstream tasks. 3DVLP takes visual grounding as the proxy task and introduces Object-level IoU-guided Detection (OID) loss to obtain high-quality proposals in the scene. Moreover, we design Object-level Cross-Contrastive alignment (OCC) task and Object-level Self-Contrastive learning (OSC) task to align the objects with descriptions and distinguish different objects in the scene, respectively. Extensive experiments verify the excellent performance of 3DVLP on three 3D vision-language tasks, reflecting its superiority in semantic 3D scene understanding. Code is available at https://github.com/iridescenttt/3DVLP.

Introduction

Semantic 3D scene understanding has recently attracted increasing research interest due to its wide applications such as automatic driving, human-machine interaction, *etc*. Much progress has been made in semantic 3D scene understanding, with task-specific models continuously pushing the state-of-the-art in various downstream tasks including visual grounding (Chen, Chang, and Nießner 2020; Zhao et al. 2021; Cai et al. 2022), dense captioning (Chen et al. 2021b), and question answering (Azuma et al. 2022).

While effective on their benchmarks, the task-specific representations obtained by existing approaches prevent them from generalizing well to other tasks. A common practice for extracting joint multimodal representation is to adopt the pre-training plus fine-tuning paradigm (Tan and Bansal 2019; Zhai et al. 2022; Alayrac et al. 2022; Zha et al. 2023). Existing works on semantic 3D scene understanding are still limited, which motivates us to introduce this paradigm in an appropriate way. However, 3D vision-language pre-training differs from pre-training in 2D vision-language tasks since point cloud data is introduced (Guo et al. 2020). The objectives designed in previous works cannot be directly applied to 3D vision-language pre-training due to the gap of downstream tasks. Therefore, it is essential to identify the shared

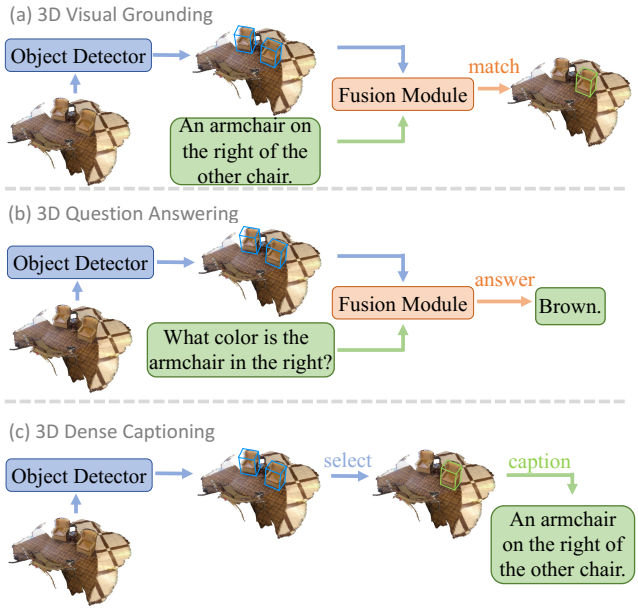


Figure 1: Relationship between 3D vision-language tasks. Firstly, all the tasks rely heavily on the object detector to locate object in the scene. Secondly, 3D vision-language tasks require an effective fusion module to understand the connection between point cloud and language.

*These authors contributed equally.
†Corresponding author: Tao Dai.

nature across different tasks in semantic 3D scene understanding to determine the appropriate pre-training model.

Figure 1 provides an intuitive depiction of the relationships among three 3D vision-language tasks and two key observations emerages from the comparision. Firstly, all of these tasks rely on object detection when applying two-stage pipeline models, which is a common practice in semantic 3D scene understanding (Chen, Chang, and Nießner 2020; Chen et al. 2021b). Secondly, an effective fusion module is required to enable information interaction across modals for a deeper understanding of the relationships between objects in the scene, such as the matching stage in visual grounding (Zhao et al. 2021; Cai et al. 2022) and the classification stage in question answering (Azuma et al. 2022).

These observations in semantic 3D scene understanding pose several challenges in designing an effective training paradigm for the pre-training model. Firstly, high-quality bounding boxes are required for object detection. These boxes represent the model's ability to segment the scene at the object level, as demonstrated by works that use a detection-then-matching pipeline (Cai et al. 2022; Chen, Chang, and Nießner 2020). Secondly, object detection requires the model to distinguish between different objects in the scene, especially when there are many objects similar to the target(Chen, Chang, and Nießner 2020). This means the model needs to be able to identify what makes objects distinct in the scene, which is challenging and has not yet been fully addressed. Thirdly, the fusion module suffers from the issue that the data come from different modalities are unaligned, as similar to 2D vision language learning (Li et al. 2021; Chen et al. 2020). Point cloud features and word token embeddings exist in different spaces, making it challenging for the fusion module to model their interactions.

To this end, we propose 3DVLP: vision-language pre-training with object contrastive learning in semantic 3D scene understanding. (1) To obtain better object bounding boxes, we introduce **Object-level IoU-guided Detection** (OID) loss. Specifically, we leverage visual grounding as the proxy task, as it shares the same objective of localizing high-quality bounding boxes. Additionally, we incorporate Distance IoU (DIoU) loss (Zheng et al. 2020) and label smoothing at the object level to achieve faster convergence and better performance. (2) We further introduce **Object-level Self-Contrastive learning** (OSC) task to distinguish the target object from others. The self-contrastive learning is performed at the object level, where boxes with an IoU higher than a specific threshold are considered positive samples and others are regarded as negative ones. (3) To enable fully information intereaction between point cloud and language, we design **Object-level Cross-Contrastive alignment** (OCC) task to align the unimodal representation across these two modalities. We use a similar IoU filter as in OSC to generate positive and negative samples, which are then fed as inputs to calculate the cross-contrastive loss.

To apply the contrastive loss at the object level, we leverage the large amount of the proposals generated by the object detector and filter positive ones with the IoU filter. The positive proposals can be regarded as diverse data augmentations of the ground truth and share similarity to some extent. The positive and negative sample pairs contain sufficient information for training the contrastive loss and help the model better distinguish different objects in the scene.

The contributions of this study are summarized as follows: (1) A 3D vision-language pre-training framework called 3DVLP has been proposed, achieving the unification of the tasks in semantic 3D scene understanding. (2) We introduce Object-level IoU-guided Detection loss to obtain high-quality bounding boxes. We also present two proxy tasks at the object level, including the Object-level Cross-Contrastive alignment task and Object-level Self-Contrastive learning task, which facilitate cross-modal alignment and help the model distinguish objects more accurately, respectively. (3) We conduct extensive experiments and empirically demonstrate the effectiveness of 3DVLP.

Related Work

3D Visual-Langauge Tasks

Recently, semantic 3D scene understanding has raised great interest and has been widely explored in recent approaches across various tasks, including 3D visual grounding (Chen, Chang, and Nießner 2020; Zhao et al. 2021; Cai et al. 2022; Luo et al. 2022), 3D dense captioning (Chen et al. 2021b), and 3D question answering (Azuma et al. 2022).

3D visual grounding aims to locate a region of interest in a scene based on a referring description. Chen et al. (Chen, Chang, and Nießner 2020) introduces ScanRefer dataset, while Achlioptas et al. (Achlioptas et al. 2020) collects two datasets containing Nr3D and Sr3D. Most existing methods rely on a detection-then-match pipeline to tackle the grounding task. 3DVG-Transformer (Zhao et al. 2021) introduces coordinate-guided contextual aggregation module to enhance proposal generation. HAM(Chen et al. 2022b) shifts attention to contextual information and develops both local and global attention module, while BUTD-DETR(Jain et al. 2022) presents a DETR-like (Zhu et al. 2020) referential grounding model that incorporates guidance from language, points, and objects. 3D-SPS(Luo et al. 2022) proposes the first one-stage end-to-end framework and mines the cross-modal relationship based on points.

Dense captioning in 3D scene requires model to derive high-quality object bounding box from point cloud data and generates corresponding descriptions. Scan2Cap (Chen et al. 2021b) extends the dense captioning task to 3D scenes based on ScanRefer and establishes a messege-passing network. SpaCap3D(Wang et al. 2022) investigates the relative spatiality of objects and builds a spatiality-guided transformer. Importantly, it designs a object-centric decoder by using a vision token as information carrier of the target object.

3D question answering requires model to generate a correct answer provided with point cloud and a question. ScanQA(Azuma et al. 2022) collects 41k question-answer pairs and brings the question-answering task into 3D scenes. Besides, it propose a baseline model by casting the task as a classification problem. FE-3DGQA(Zhao et al. 2022) proposes anthoer datasets and predicts the answer through a token encoding and fusion module based on attention.

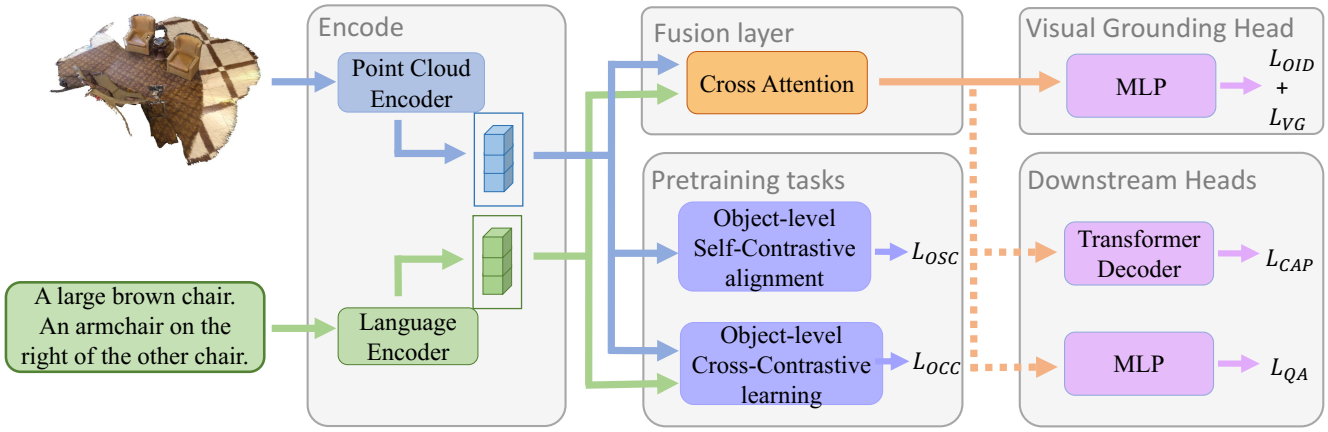


Figure 2: Pipeline of 3DVLP in semantic 3D scene understanding. 3DVLP takes visual grounding as the proxy task and utilizes Object-level IoU-guided Detection (OID) loss to boost the performance of the object detector. We also introduce Object-level Cross-Contrastive alignment task and Object-level Self-Contrastive learning task in the pre-training stage, which facilitate cross-modal alignment and enable the model to distinguish objects more accurately, respectively.

3D Vision-Language Pre-training

Recently, there have been some studies focusing on vision-language pre-training of point clouds. PointCLIP (Zhang et al. 2022), PointCLIP V2 (Zhu et al. 2022) and CLIP2point (Huang et al. 2022b) utilize CLIP to align point cloud with text. CrossPoint (Afham et al. 2022) renders the point cloud into image and apply contrastive loss for intra-modal and cross-modal alignments. They mainly focus on tasks over a single object, while our research deals with multiple objects in semantic scene understanding. The models mentioned above are incapable of tackling downstream tasks in the scene such as visual grounding and dense captioning.

A similar pre-training framework in semantic scene understanding is 3D-language pre-training (3DLP) (Jin et al. 2023), which utilize semantic-level and contextual alignment for cross-modal fusion. Moreover, it applies masked modeling in both proposals and language to get better understanding across modalities. In contrast, the introduction of the OID loss in 3DVLP during the pre-training phase markedly improves the performance of the object detector for scenes understanding. Consequently, 3DVLP surpasses 3DLP in (Jin et al. 2023) by a large margin in unique scenarios, especially in terms of Acc@0.5.

Method

As demonstrated in Figure 2, both the point cloud and linguistic data are encoded and fed into a cross-attention module for fusion. The training can be mainly divided into the pre-training stage and the fine-tuning stage. In the pre-training stage, 3DVLP utilizes visual grounding as the proxy task and employs Object-level IoU-guided Detection loss for high-quality object detection. Additionally, 3DVLP is pre-trained on other designed proxy tasks, including Object-level Cross-Contrastive alignment and Object-level Self-Contrastive learning. In the finetuning stage, we transfer the backbone to downstream tasks with task-specific heads.

Object-level IoU-guided Detection Loss

We consider visual grounding as the proxy task since it shares the same objective of obtaining high-quality proposals. Additionally, we propose Object-level IoU-guided Detection loss to enhance the performance of the object detector, as demonstrated in Fig. 4a. Specifically, we introduce the Distance IoU (DIoU) loss (Zheng et al. 2020) for bounding box regression. Given the predicted proposal $\mathbf{b}_p$ and ground truth $\mathbf{b}_{gt}$, we calulate the IoU between them and have the following regression loss:

$$\mathcal{L}_{DIoU}(\mathbf{b}_p, \mathbf{b}_{gt}) = 1 - IoU + \frac{\rho^2(\mathbf{b}_p, \mathbf{b}_{gt})}{c^2}, \quad (1)$$

where c is the diagonal length of the smallest enclosing box covering the two boxes and $\rho^2(\cdot,\cdot)$ is the Euclidean distance. However, previous approaches(Zhao et al. 2021; Cai et al. 2022) treats the matching stage as a classification problem and use the proposal with the highest IoU as a supervised label to train the fusion module. In this case, the DIoU loss can only be applied to a single proposal, which weakens its efforts in optimization. Additionally, due to the large number of proposals generated by the detector, there can be multiple boxes pointing to the target object, and these boxes may share similar semantic information, making it difficult to achieve accurate matching with a one-hot label.

We take inspiration from label smoothing (Müller, Kornblith, and Hinton 2019) and address such matching problems by introducing an IoU filter. As shown in Fig. 3, given a pre-defined IoU threshold δ and the weight factor ε, positive proposals are filtered according to their IoU with the ground truth, and weights are assigned to them based on their total

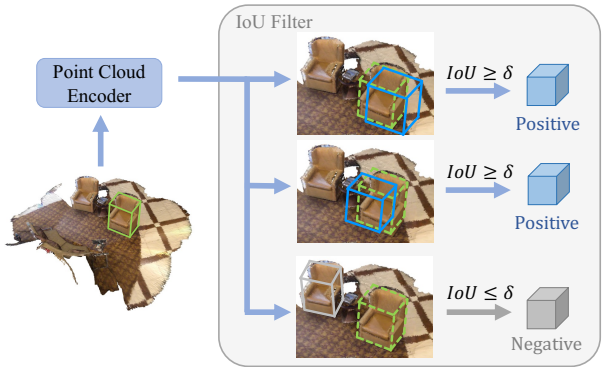


Figure 3: Illustration of the IoU filter in 3DVLP. To apply label smoothing and contrastive loss at the object level, proposals with IoU higher than a threshold δ are considered positive samples while others are regarded as the negative ones.

count, denoted by K. The weight of proposal p in the soft label is shown in Eq. (2).

$$y_p = \begin{cases} 1-\varepsilon & \text{if } IoU_p = IoU_{max} \\ \frac{\varepsilon}{K} & \text{if } IoU_p \geq \delta \text{ and } IoU_p \neq IoU_{max} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We further combine DIoU loss and label smoothing to obtain our OID loss, as demonstrated in Eq. (3).

$$\mathcal{L}_{OID} = \sum_p y_p \cdot \mathcal{L}_{DIoU}(\mathbf{b}_p, \mathbf{b}_{gt}). \quad (3)$$

Usually, label smoothing in common classification tasks assigns weight to all negative classes and aims to prevent overfitting of the dataset. Completely different from traditional goals, label smoothing in 3DVLP is motivated by the need to optimize multiple proposals that point towards the ground truth simultaneously. Thus, we only assign weights to the positive proposals obtained by the IoU filter during label smoothing. This distinct approach of using label smoothing enables faster convergence and better performance.

Object-level Cross-contrastive Alignment

As a common practice (Zhao et al. 2021; Cai et al. 2022), a cross-modal attention module is applied for feature fusion between language and point cloud embedding. However, it is observed that the data distribution across modalities is not well-aligned, resulting in insufficient interaction between the embedding of proposals and the language feature. To address this issue, contrastive learning can provide insights for embedding alignment across different distributions. However, naive implementation over proposals is not effective, since multiple proposals pointing at the target object might contain semantically similar information, thereby conflicting with the optimization objective of contrastive loss.

Based on these observations, we reconsider contrastive learning at the object level and introduce the Object-level Cross-Contrastive alignment (OCC) task to enhance the performance of the cross fusion module, as shown in Fig. 4b. The OCC task is proposed to align the distribution of cross-modal data. Specifically, in the training stage, we introduce the target detection boxes of real objects and select all the predicted boxes with IoU greater than a pre-defined threshold as positive samples since they semantically point to the target object and should have similar features. The remaining predicted boxes are considered negative samples, representing the proposals of other objects or background. We then align the features of positive samples with the language embedding and push the features of negative samples away with the contrastive loss to achieve better cross-modal understanding. Formally, we have the following contrastive loss, which serves as the loss function for our OCC task.

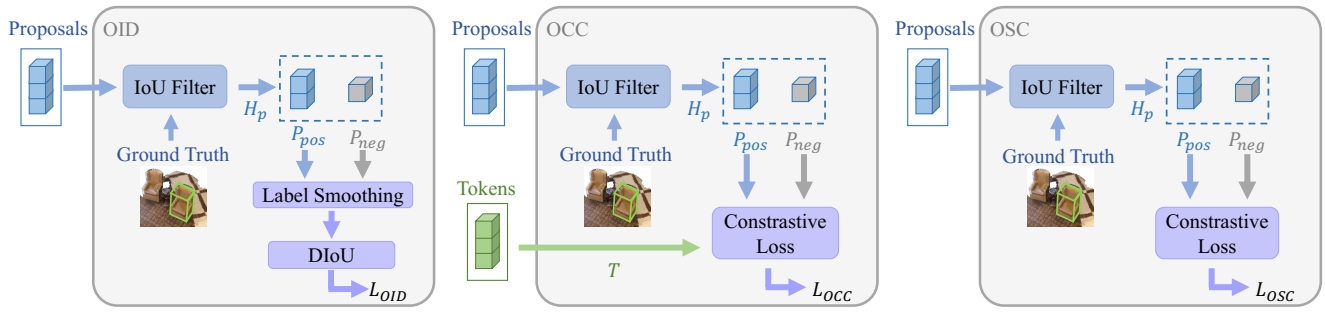
$$\mathcal{L}_{\text{OCC}} = -\frac{1}{2}\mathbb{E}_{(\mathbf{b_{gt}},T)\sim D}\Big[\log \frac{\sum_{p\in P_{pos}} exp(s(H_p,T))}{\sum_{\hat{p}\in P_{pos}\cup P_{neg}} \exp(s(H_{\hat{p}},T))} + \log \frac{\sum_{p\in P_{pos}} \exp(s(T,H_p))}{\sum_{\hat{p}\in P_{pos}\cup P_{neg}} \exp(s(T,H_{\hat{p}}))}\Big]. \quad (4)$$

where H_p represents the embedding of proposal p, and T denotes the language embedding. Given $\mathbb{I}$ as the indicator function, $IoU(\cdot,\cdot)$ as the IoU score between two boxes, and δ as the IoU threshold, we have $P_{pos} = \{p|IoU(\mathbf{b}_p, \mathbf{b}_{gt}) \geq \delta\}$ as the set of proposals containing positive samples while $P_{neg} = \{p|IoU(\mathbf{b}_p, \mathbf{b}_{gt}) < \delta\}$ containing the negative ones. $s(\cdot,\cdot)$ represents the similarity score function for measuring the similarity between two types of features, such as by performing a dot product operation. Note that the threshold δ determines how close positive samples should be to align with the language embedding. Specifically, when $\delta = IoU_{max}$, Eq. (4) only considers the proposal with the highest IoU to be the positive sample and reverts to the original formula of traditional pairwise contrastive loss.

With incorporating the IoU filter, we utilize large amount of proposals generated by an object detector to calculate contrastive loss. The positive proposals selected through the IoU filter can be regarded as diverse data augmentations of the ground truth. Therefore, the sufficient information contained in these positive proposals aids the model in better extracting the intrinsic characteristics of objects for alignment with textual embedding. Such meaningful information fulfills the substantial data sample requirements of contrastive learning, significantly boosting the generalization and robustness of the cross-modal fusion module in 3DVLP.

Object-level Self-contrastive Learning

In semantic 3D scene understanding, the presence of similar objects in the scene can significantly affect the matching performance of the model. Therefore, a well-designed pre-training model should be capable of accurately distinguishing between objects in the scene and understanding what makes them similar or different. To address this issue, we utilize self-contrastive loss that incentivizes the model to capture features that differentiate objects. Similarly, we require an object-level self-contrastive loss instead of the pairwise loss to effectively differentiate between objects and improve the model's semantic understanding of the scene.



(a) Object-level IoU-guided Detection (b) Object-level Cross-contrastive Alignment (c) Object-level Self-Contrastive Learning

Figure 4: Illustration of Object-level IoU-guided Detection (OID) loss, Object-level Cross-contrastive alignment (OCC) and Object-level Self-Contrastive learning (OSC) pre-training tasks. All the modules utilize a IoU filter to select positive proposals.

Therefore, we introduce the Object-level Self-Contrastive learning (OSC) task for object detection, as shown in Fig. 4c. The OSC task is proposed for unimodal point cloud data and aims to optimize the embedding generated by the point cloud encoder. Based on the idea in OCC task, we utilize the IoU threshold to select positive samples and negative ones for self contrastive learning. By optimizing the self-contrastive loss, 3DVLP encourages the features of the proposals targeting the ground truth object to be as dissimilar as possible from those of other proposals, thereby enabling the fusion module to distinguish different objects easily. Following Eq. (4), we replace the language embedding with the embedding of proposals to obtain the corresponding contrastive loss for OSC module, as shown in Eq. (5).

$$\mathcal{L}_{\text{OSC}} = -\mathbb{E}_{\mathbf{b}_{gt}\sim D}\Big[\log \frac{\sum_{p,\hat{p}\in P_{pos}} \exp(s(H_p, H_{\hat{p}}))}{\sum_{p,\hat{p}\in P_{pos}\cup P_{neg}} \exp(s(H_p, H_{\hat{p}}))}\Big]. \quad (5)$$

The sufficient information within multiple positive proposals, complemented by their contrast against the negative proposals, helps the model in more effectively discerning the similarities and differences between objects in the scene. Consequently, the OSC task at the object level enhances the performance of the object detector for downstream tasks.

Experiment

Datasets and Implementation Details

Visual Grounding Dataset: We select the benchmark dataset ScanRefer (Chen, Chang, and Nießner 2020) for visual grounding task. It consists of 800 3D scenes from the ScanNet dataset (Dai et al. 2017), each annotated with object bounding boxes and corresponding text descriptions. To evaluate our results, we employ two evaluation metrics: IoU@0.25 and IoU@0.5, which measure the percentage of times the proposals have an IoU greater than the threshold.

Dense Captioning Dataset: We conduct experiments on Scan2Cap dataset (Chen et al. 2021b) for the dense captioning task. We jointly measure the quaility of the generated model with captioning matrics including CiDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), BlEU-4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005) and ROUGE (Lin 2004), cited as C, B-4, M and R, respectively. We combine the metrics above with an IOU threshold and adopt the m@kIoU metric: $m@kIoU = \frac{1}{N}\sum_{i=1}^{N} m_i \cdot \mathbb{I}(IoU \geq k)$, where m represents the captioning metric, k is the threshold of IoU and $\mathbb{I}$ stands for the indicator function.

Question Answering Dataset: We perform question answering tasks over the ScanQA dataset (Azuma et al. 2022), which consists of 41363 questions and 32337 unique answers from 800 scenes derived from the ScanNet scenes. Following the evaluation in ScanQA, EM@1 and EM@10 are used as the evaluation metric. EM@K is the percentage of predictions where the top K predicted answers exactly match any of the ground-truth answers.

Implementations Details

We first train 3DVLP over the proposed proxy tasks including visual grounding, OCC and OSC in the pre-training stage for 200 epochs. We then evaluate our methods on the dense captioning and question answering tasks by finetuning with tasks-specific loss. Importantly, we use VoteNet (Qi et al. 2019) as our point cloud encoder and a frozen BERT(Devlin et al. 2018) as the language encoder to avoid over-fitting on short-length sentences. For grounding task, we model it as a classification problem and use a 3-layers MLP as the head. For captioning task, we use a Transformer decoder with 6 layers and 128 as the hidden size. For QA task, a lightweight MLP is adopted to predict the score for each answer and the answer with the highest score is selected as the final answer. More details of the downstream heads can be found in the appendix. We set the batch size as 8 and the initial learning rate is set to be 0.002 for the detector and 5e-4 for other modules in the 3DVLP. Codes are implemented by Pytorch and run on a Nvidia 3090 GPU.

Comparison with State-of-the-art Methods

3D Visual Grounding Task We present the results of 3D visual grounding in Table 1. The "3D" models only utilizes raw attributes in point cloud input features, while "2D+3D" models use 2D multi-view features as additional inputs. Note that the results of BUTD-DETR (Jain et al. 2022) is re-evaluated by removing the GT object labels in the text

Method	Data	Unique		Multiple		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
InstanceRefer (Yuan et al. 2021)	3D	77.45	66.83	31.27	24.77	40.23	32.93
3DVG-Transformer (Zhao et al. 2021)	3D	77.16	58.47	38.38	28.70	45.90	34.47
3DJCG (Cai et al. 2022)	3D	78.75	61.30	40.13	30.08	47.62	36.14
3D-SPS (Luo et al. 2022)	3D	81.63	64.77	39.48	29.61	47.65	36.43
BUTD-DETR (Jain et al. 2022)	3D	82.77	63.81	**44.01**	**33.51**	49.69	38.01
3DLP (Jin et al. 2023)	3D	79.35	62.60	42.54	32.18	49.68	38.08
3DVG-Transformer (Zhao et al. 2021)	2D + 3D	81.93	60.64	39.30	28.42	47.57	34.67
Multi-View Trans (Huang et al. 2022a)	2D + 3D	77.67	66.45	31.92	25.26	40.80	33.26
3D-SPS (Luo et al. 2022)	2D + 3D	84.12	66.72	40.32	29.82	48.82	36.98
3DJCG (Cai et al. 2022)	2D + 3D	83.47	64.34	41.39	30.82	49.56	37.33
D3Net (Chen et al. 2021a)	2D + 3D	-	**70.35**	-	30.05	-	37.87
3DLP (Jin et al. 2023)	2D + 3D	84.23	64.61	43.51	33.41	51.41	39.46
3DVLP	2D + 3D	**85.18**	70.04	43.65	33.40	**51.70**	**40.51**

Table 1: Comparison of different methods in 3D visual grounding task. We measure the percentage of the correctly predicted bounding boxes whose IoU with the ground-truth boxes are larger than 0.25 and 0.5, respectively.

Method	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5
MORE (Jiao et al. 2022)	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3D (Wang et al. 2022)	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG (Cai et al. 2022)	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80
D3Net (Chen et al. 2021a)	-	-	-	-	46.07	30.29	24.35	51.67
3DLP (Jin et al. 2023)	**70.73**	**41.03**	28.14	59.72	**54.94**	32.31	24.83	51.51
3DVLP	66.63	40.85	**36.12**	**61.03**	54.41	**34.10**	**34.34**	**54.28**

Table 2: Comparison of different methods in 3D dense captioning task. We report the result with the percentage of the predicted bounding boxes whose IoU with the ground truth are greater than 0.25 and 0.5.

Method	EM@1	EM@10	Acc@0.25	Acc@0.5
ScanQA	21.05	51.23	24.96	15.42
FE-3DGQA	22.26	54.51	26.62	18.83
3DLP	21.65	50.46	-	-
3DVLP	**24.03**	**57.91**	**33.38**	**26.12**

Table 3: Comparison of different methods in 3D question answering task. The results are presented with the percentage of predictions where the top K predicted answers exactly match any of the ground-truth answers. We also report Acc@0.25 and Acc@0.5 similar to visual grounding.

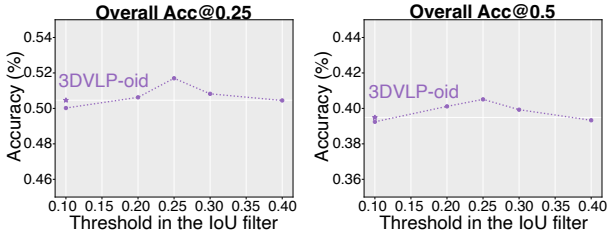


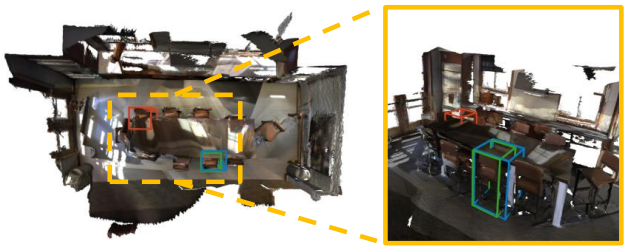
Figure 5: Comparison of the performance when using different threshold in the IoU filter. We also compare a variant of 3DVLP with only OID loss, referred as 3DVLP-oid.

queries for a fair comparison, provided by UniT3D (Chen et al. 2022a). The results indicate that 3DVLP performs remarkably well and outperforms the baselines by a large margin. In terms of unique scenes, 3DVLP achieves the highest accuracy in Acc@0.5 and ranks second in Acc@0.25, indicating the significant impact of our OID loss in developing the model's ability to identify high-quality bounding boxes. Furthermore, when comparing multiple and unique metrics, previous works suffers from issues related to the presence of similar objects in the scene, leading to poor matching results. However, the introduction of OSC and OCC tasks in 3DVLP enables it to achieve competitive performance in multiple metrics, showcasing its ability to accurately locate objects in complex scenes. In the overall metric, 3DVLP's performance surpasses the baseline by 0.29% in Acc@0.5 and 1.15% in Acc@0.25.

3D Dense Captioning Task As presented in Table 2, it is evident that 3DVLP shows excellent transfer performance in dense captioning task. Importantly, the point cloud encoder in 3DVLP extracts universal features that generalize well in dense captioning, enabling 3DVLP to outperform other baselines. Specifically, 3DVLP achieves an improvement of 7.98% and 1.31% in terms of M@0.25, and R@0.5, respectively. Moreover, the results show that 3DVLP outperforms the second baseline by 8.46% in M@0.25 and 9.51% in M@0.5. Among various evaluation metrics, METEOR focuses on capturing the semantic similarity and fluency between the output and the ground truth, thereby indicating the generalization ability of 3DVLP.

Module			Visual Grounding		Dense Captioning			
OID	OCC	OSC	Acc@0.25	Acc@0.5	C@0.5	B-4@0.5	M@0.5	R@0.5
			50.59	37.96	53.12	31.90	33.93	52.27
✓			50.46	39.49	52.91	33.91	34.28	54.08
	✓		51.15	38.44	53.24	32.79	33.98	52.99
		✓	50.91	38.28	51.41	32.93	34.00	52.94
✓	✓	✓	**51.70**	**40.51**	**54.41**	**34.10**	**34.34**	**54.28**

Table 4: Quantitative results of the overall accuracy in visual grounding and the metric in dense captioning.



(a) There is a tall chair pulled up to the table in the room. It is the second from the right.

(b) This is a brown ottoman in front of a brown sofa. The ottoman has a black backpack, and a black duffel bag, and a box of tissues.

Figure 6: Qualitative results in Visual Grounding. We mark the ground truth in blue, 3D-SPS in red and 3DVLP in green.

3D Question Answering Task Based on the results in Table 3, 3DVLP consistently outperforms other methods in the question answering task. For example, 3DVLP achieves approximately 1.7%-2.4% improvement in EM@1 and EM@10 compared to the baseline. Moreover, question answering benefits from the pre-training model when compared to ScanQA, as 3DVLP utilizes the same classification head. Furthermore, 3DVLP provides a boost by 6.76% and 7.23% in Acc@0.25 and Acc@0.5, respectively.

Ablation Study

Does the OID loss and the designed proxy tasks benefit downstream tasks? We investigate the contribution of each module in 3DVLP and the results in Table 4 demonstrate that both visual grounding and dense captioning tasks benefit from each proposed module. In visual grounding, the OID loss significantly improves the quality of the proposals, thereby enhancing Acc@0.5 to a large degree. Furthermore, neither the introduction of OSC nor OCC provides a remarkable boost in Acc@0.25, indicating the superiority of optimization at the object level in complex scenes. In dense captioning, the improvement of the model is consistent with that in visual grounding by combining the modules together. **Is the improvement in OSC and OCC sensitive to the threshold used the IoU filter?** To have a better understanding of the threshold δ used in the IoU filter, we estimate the results of the overall Acc in visual grounding with the varying δ. Moreover, we include 3DVLP with only OID loss as a base variant, referred as 3DVLP_oid. As shown in Fig. 5, the performance improves when increasing the threshold from 0.1 to 0.25. This is because proposals targeting other objects can be incorrectly considered as positive samples and thus mislead the training optimization when using a low threshold. However, we further increase the threshold and observe that the improvement is not consistent. The performance drops with a large threshold since model will regard proposals that are not good enough as negative samples, resulting in semantic divergence. This is similar to what happens with the traditional pairwise contrastive loss. Therefore, based on our results, we believe that selecting a threshold of 0.25 in the IoU filter is a reasonable tradeoff.

Qualitative Results

To further explore how 3DVLP improves the performance in visual grounding, we provide the comparison results with 3D-SPS as shown in Figure 6. These examples demonstrate that 3DVLP has a better understanding of the relationship between scene and language as a result of incorporating OSC and OCC, leading to better performance.

Conclusion

In this paper, we investigates the shared nature across different tasks in semantic 3D scene understanding and proposes 3DVLP, a contrastive 3D vision-language pre-training framework. 3DVLP introduces the object-level IoU-guided detection loss to obtain high-quaility proposals, aligns the point cloud representation and language representation by training over object-level cross-contrastive alignment task and develops its ability to distinguish different objects in the scene through object-level self-contrastive learning task. Comprehensive experiments reveal the generalization ability and superiority of 3DVLP over all downstream tasks in semantic 3D scene understanding, leading to a new state-of-the-art performance. Future work needs to focus on dealing with the fusion of point cloud and language, desirably about the full interaction of multi-level information.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China, under Grant No. 2023YFF0905502, National Natural Science Foundation of China, under Grant (62302309, 621712488), Shenzhen Science and Technology Program (Grant No. RCYX20200714114523079,JCYJ20220818101014030, JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1), and Tencent "Rhinoceros Birds" - Scientific Research Foundation for Young Teachers of Shenzhen University.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 422–440. Springer.

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9902–9912.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19129–19139.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Cai, D.; Zhao, L.; Zhang, J.; Sheng, L.; and Xu, D. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16464–16473.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, 202–221. Springer.

Chen, D. Z.; Hu, R.; Chen, X.; Nießner, M.; and Chang, A. X. 2022a. UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding. *arXiv preprint arXiv:2212.00836*.

Chen, D. Z.; Wu, Q.; Nießner, M.; and Chang, A. X. 2021a. D3Net: a speaker-listener architecture for semi-supervised dense captioning and visual grounding in RGB-D scans. *arXiv preprint arXiv:2112.01551*.

Chen, J.; Luo, W.; Wei, X.; Ma, L.; and Zhang, W. 2022b. HAM: Hierarchical Attention Model with High Performance for 3D Visual Grounding. *arXiv preprint arXiv:2210.12513*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, 104–120. Springer.

Chen, Z.; Gholami, A.; Nießner, M.; and Chang, A. X. 2021b. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3193–3203.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12): 4338–4364.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022a. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15524–15533.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2022b. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv preprint arXiv:2210.01055*.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 417–433. Springer.

Jiao, Y.; Chen, S.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, 528–545. Springer.

Jin, Z.; Hayat, M.; Yang, Y.; Guo, Y.; and Lei, Y. 2023. Context-aware Alignment and Mutual Masking for 3D-Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10984–10994.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16454–16463.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, H.; Zhang, C.; Yu, J.; and Cai, W. 2022. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1791–1800.

Zha, Y.; Wang, J.; Dai, T.; Chen, B.; Wang, Z.; and Xia, S.-T. 2023. Instance-aware Dynamic Prompt Tuning for Pre-trained Point Cloud Models. *arXiv preprint arXiv:2304.07221*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18123–18133.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8552–8562.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2928–2937.

Zhao, L.; Cai, D.; Zhang, J.; Sheng, L.; Xu, D.; Zheng, R.; Zhao, Y.; Wang, L.; and Fan, X. 2022. Towards Explainable 3D Grounded Visual Question Answering: A New Benchmark and Strong Baseline. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhu, X.; Zhang, R.; He, B.; Zeng, Z.; Zhang, S.; and Gao, P. 2022. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*.