UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding

Dave Zhenyu Chen[1] Ronghang Hu[2] Xinlei Chen[2] Matthias Nießner[1] Angel X. Chang[3]

[1]Technical University of Munich [2]Meta AI [3]Simon Fraser University

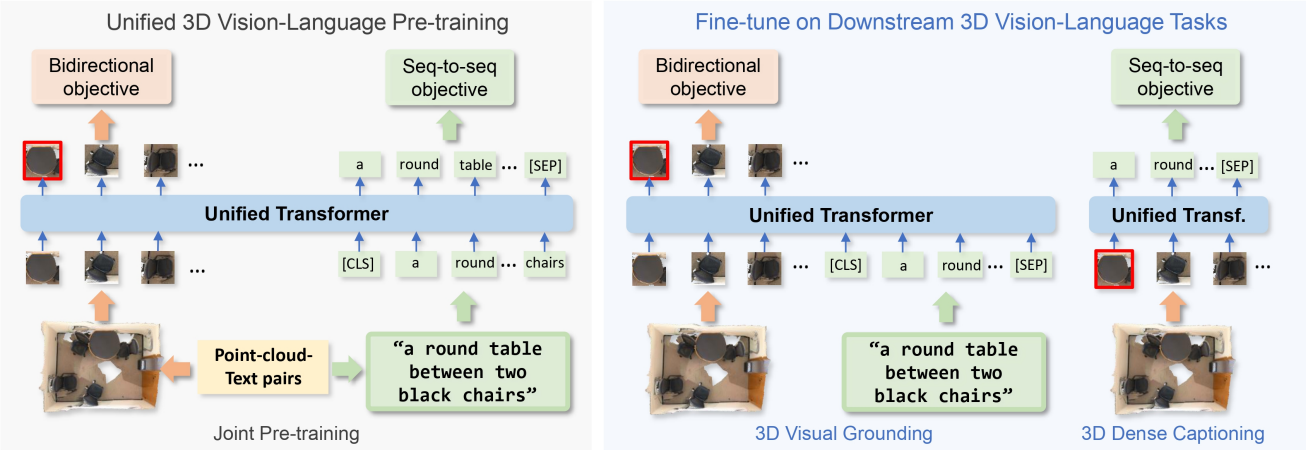


Figure 1: We present UniT3D, a unified transformer for 3D dense captioning and visual grounding. UniT3D is pre-trained with both bidirectional and seq-to-seq objectives on point-cloud-text pairs. Afterwards, it is further fine-tuned for 3D visual grounding and dense captioning. We show that our proposed architecture and pre-training scheme largely improve the performance on both downstream tasks.

Abstract

Performing 3D dense captioning and visual grounding requires a common and shared understanding of the underlying multimodal relationships. However, despite some previous attempts on connecting these two related tasks with highly task-specific neural modules, it remains understudied how to explicitly depict their shared nature to learn them simultaneously. In this work, we propose UniT3D, a simple yet effective fully unified transformer-based architecture for jointly solving 3D visual grounding and dense captioning. UniT3D enables learning a strong multimodal representation across the two tasks through a supervised joint pre-training scheme with bidirectional and seq-to-seq objectives. With a generic architecture design, UniT3D allows expanding the pre-training scope to more various training sources such as the synthesized data from 2D prior knowledge to benefit 3D vision-language tasks. Extensive experiments and analysis demonstrate that UniT3D obtains significant gains for 3D dense captioning and visual grounding.

1. Introduction

The 3D vision-language field has been drawing increasing research interest in jointly understanding 3D scenes [11, 34, 26, 35, 17, 14, 36, 18, 9] and natural language [41, 13, 28, 46, 2], such as 3D visual grounding [5] and 3D dense captioning [8]. The task of 3D visual grounding takes as input a point-cloud-text pair and outputs a bounding box of the referred object. As its sibling task, 3D dense captioning expects a point cloud as input and densely generates object bounding boxes and descriptions in the scene. Both tasks enable applications such as assistive robots and natural language control in AR/VR systems.

Although these two 3D vision-language tasks are naturally complementary to each other, previous attempts to connect them [6, 3] only exploit partly shared object spatial information, leaving the joint nature between object relationships and textual semantics underdeveloped. More concretely, to localize the target object in the scene, 3D visual

grounding methods require a joint understanding of object attributes and spatial relationships. This requirement persists for 3D dense captioning when densely describing the appearance and spatial aspects of the objects. Therefore, it is naturally desirable to develop a shared representation between two tasks with a unified task-agnostic framework, where the two input modalities are jointly encoded and enhanced. Moreover, such generic multimodal representation is not only beneficial for sharing knowledge between visual grounding and dense captioning, but could also enable 3D vision-language research beyond specific domains, as in the case of VisualBERT [27] or CLIP [38] on 2D images.

To this end, we propose UniT3D, a joint transformer-based solution to facilitate vision-language representation learning for 3D visual grounding and dense captioning, as illustrated in Fig. 1. Unlike prior works with two distinct and task-specific neural modules [3, 6], our method tackles the tasks of 3D dense captioning and visual grounding via a task-agnostic unified transformer with light-weight output heads. To enable joint vision-language representation learning, we design a supervised training scheme that combines the bidirectional objective with query-aware object matching supervision and the seq-to-seq objective with object-aware sequence generation supervision. Through fine-tuning on specific tasks with corresponding output heads, the joint vision-language representation learned by our task-agnostic multimodal transformer is well capable of supporting both the localization of the referred objects in the scenes and the dense generation of object descriptions.

One challenge of the joint 3D vision-language representation learning is that existing 3D vision-language datasets [5, 1] are relatively limited in size and variety compared to their 2D counterparts such as MSCOCO [7] and Conceptual Captions [39]. To address this challenge, we build a large-scale 3D vision-language dataset with text annotations generated from an image captioner learned on abundant 2D image and text datasets. Concretely, we apply the image captioner from Mokady et al. [32] to ScanNet images to obtain synthetic image-text pairs and convert them to point-cloud-text pairs by cropping the reconstructed point clouds in ScanNet within the image frustum using camera parameters as our synthetic 3D vision-language data. We show that along with jointly pre-training on such synthesized data with the proposed bidirectional and seq-to-seq objectives, our UniT3D model obtains significant performance gains on the downstream 3D vision-language tasks. To summarize, our contributions are threefold:

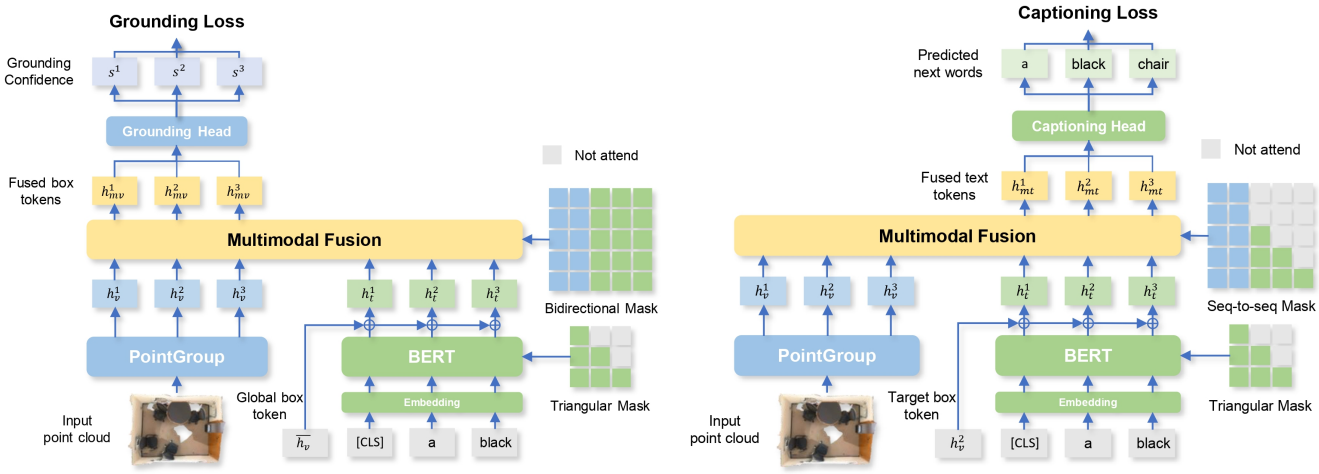
- We introduce a multimodal transformer architecture to solve 3D visual grounding and dense captioning in a fully unified fashion.
- We propose a supervised joint pre-training scheme with bidirectional and seq-to-seq objectives to facilitate multimodal feature learning.
- We construct a large-scale synthetic point-cloud-text dataset, showing that the distilled 2D prior knowledge is beneficial to 3D vision-language tasks.

2. Related work

3D visual grounding and dense captioning. Recently, there has been a thriving research interest in 3D vision-language [5, 1]. Seminal works [5, 1] concurrently propose ScanRefer and ReferIt3D dataset consisting of free-form descriptions of 3D real-world object from ScanNet [12] scenes. Chen et al. [5] propose the 3D visual grounding task to localize a target object in a 3D environment using text queries. The majority of the previous work on 3D visual grounding [5, 1, 20, 47, 50, 3, 6] concentrate on distinguishing the multimodal relationships among objects and text queries. As its reversed task, the task of 3D dense captioning predicts the bounding boxes and the associated descriptions for all objects in the input 3D scene [8]. Similar to 3D visual grounding, most prior works in this track [8, 48, 24] learn the multimodal relationships among the objects in the 3D scene and decode them as text outputs.

We observe that these two tasks are complementary in nature, with minor discrepancies in their outputs (boxes vs. descriptions). There are several initial attempts in bridging them together. Chen et al. [6] apply a speaker-listener architecture to self-critically improve the overall performance of visual grounding and dense captioning. However, there is no shared representation between the visual and text modalities. Cai et al. [3] link both tasks with a shared detection backbone attached to two task-specific neural modules. Despite the shared object relationships within the visual modality, the multimodal representations between objects and text inputs are not explicitly modeled. In contrast, our method directly builds multimodal representations with a joint multimodal fusion transformer, enabling both visual grounding and dense captioning prediction from shared multimodal knowledge.

Vision-language pre-training. In light of the shared nature of vision-language tasks, developing a joint multimodal representation between the vision and language modalities has become a popular research domain in the 2D vision-language community [27, 30, 51, 44, 19, 40, 43]. Inspired by the masked language modeling in Devlin et al. [13], seminal works such as VisualBERT [27] and ViLBERT [30] introduce the masked vision-language modeling scheme for learning the multimodal representation. However, such pre-training schemes only concentrate on multimodal encoding, but neglect multimodal decoding, which is essential for generation-based tasks (e.g. image captioning). Based on masked vision-language modeling, Zhou et al. [51] design the bidirectional and seq-to-seq pre-training objectives to



(a) Training flow for 3D visual grounding

(b) Training flow for 3D dense captioning

Figure 2: Task-specific training schemes for UniT3D. The input point cloud is fed into a PointGroup [23] detection backbone to generate box tokens $\{h_v^i\}$. Parallelly, the text query is processed with a pre-trained BERT [13] to produce text tokens $\{h_t^i\}$. For visual grounding, as shown in 2a, the multimodal fusion module takes in the concatenated box and text tokens to produce the fused box tokens $\{h_{mv}^i\}$. Then, a lightweight grounding head predicts the grounding confidence scores for each proposal. Similarly, for dense captioning, the current target box token from PointGroup (such as h_v^2) is added to all BERT outputs as the captioning cue. The captioning head then takes in the fused text tokens $\{h_{mt}^i\}$ from the multimodal fusion module, and predicts the next tokens in the input sequence, as shown in 2b. Note that a multimodal triangular masking scheme is applied here to enforce the left-to-right sequence modeling.

empower both multimodal encoding and decoding. Hu and Singh [19] further expand the spectrum of the applicable downstream tasks using the learned multimodal representation. However, due to the limited amount of existing vision-language data in 3D, such pre-training strategies entangled with large transformer architectures cannot be directly migrated to the 3D vision-language domain. Hence, We propose a synthetic vision-language data generation scheme featuring the distilled 2D multimodal knowledge to facilitate the multimodal pre-training in the 3D domain.

3D vision with 2D priors. Leveraging 2D distilled knowledge such as CLIP [38] for 3D vision is a trending research topic. PointCLIP [49] transfers 2D knowledge to 3D by conducting alignments between the CLIP-encoded point cloud image and category texts. This scheme enables zero-shot classification on point cloud without training on any 3D datasets such as ShapeNet [4]. CLIP2Point [21] aligns depth image features to CLIP-encoded point cloud image features to facilitate image-depth pre-training for 3D shape understanding. SemAbs [15] extracts relevancy maps from CLIP for open-vocabulary 3D scene understanding and CLIP-NeRF [42] leverages CLIP embedding space for 3D scene manipulation. More advanced methods utilize pixel-to-point correspondence [29, 16] or kernel inflation technique [45] to tackle 3D shape/scene understanding tasks leveraging 2D distilled knowledge. In this work, we propose to use 2D vision-language priors to address the limited scale and variety of existing 3D data, enabling multimodal representation learning for 3D vision-language.

3. Method

Our architecture stem consists of three main opponents as shown in Fig. 2: a PointGroup detection backbone, a pre-trained BERT [13] encoder, and a transformer-based multimodal fusion module. Given a point-cloud-text pair as input, PointGroup takes in the point cloud $\mathbf{P} \in \mathcal{R}^{N\times(3+K)}$ to produce a sequence of M box tokens $\{h_v^1, ..., h_v^i, ..., h_v^M\}$. We follow Chen et al. [5] to construct the auxiliary point features with multi-view features and point normals ($k = 134$). In the meantime, the BERT encoder encodes the input text of length L into a sequence of text tokens $\{h_t^1, ..., h_t^j, ..., h_t^L\}$. We use [CLS] and [SEP] as the start and end tokens of the input text in the BERT encoder, and apply a triangular mask to enforce the left-to-right sequence modeling. We also add a visual context token (from a global box for grounding or a target box for captioning) to every text token h_t^j. Finally, the multimodal fusion module enriches the concatenated box-text input sequence and produces a sequence of fused tokens as output.

3.1. A unified model with task-specific objectives

3D visual grounding with bidirectional objective. As shown in Fig. 2a, the input point cloud is processed by PointGroup to produce the instance masks and the box to-

kens $\{h_v^1, ..., h_v^M\}$. In the meantime, an object text query is fed into BERT to produce the text tokens $\{h_t^1, ..., h_t^L\}$. Here, we average all box tokens as the global box token $\hat{h_v}$ and add it to all text tokens as the global visual cue. To be consistent with the captioning task, we constantly apply a triangular mask to enforce the left-to-right modeling. Then, the box tokens are concatenated with the text tokens and fed into the multimodal fusion module to acquire the fused multimodal features. We apply a bidirectional mask in the multimodal fusion module to enable the sequence encoding in both directions. Afterwards, we feed the fused box tokens $\{h_{mv}^1, ..., h_{mv}^M\}$ to the lightweight grounding head to predict the grounding confidence scores for each object proposal. The box with the highest grounding confidence will be taken as the grounding output. Here, we use a conventional cross-entropy loss L_{G} for supervision, where the object proposal with the highest IoU with the GT queried box is treated as training GT. Inspired by Chen et al. [5], we also attach a lightweight MLP taking the encoded [CLS] token as input to predict the object semantic class from the text query. We supervise the language object classification with another cross-entropy loss L_{cls}. Since the PointGroup is fine-tuned end-to-end, we apply the original PointGroup loss identical to Jiang et al. [23], where $L_{\text{PG}} = L_{\text{sem}} + L_{\text{o\_dir}} + L_{\text{o\_reg}} + L_{\text{c\_score}}$. Overall, we combine the overall visual grounding loss as: $L_{\text{FG}} = L_{\text{PG}} + L_{\text{G}} + L_{\text{cls}}$.

3D dense captioning with seq-to-seq objective. We follow the next word prediction strategy to train our model, where a lightweight MLP is attached to the task-agnostic multimodal fusion module as the captioning head. As shown in Fig. 2b, the input text is padded with [CLS] and [SEP] token at the beginning and the end of the sentence, respectively. Following the teacher-forcing scheme, for a word sequence $\{w^1, ..., w^L\}$ of length L, we take the 1^{st} to the $(L-1)^{\text{th}}$ words as input and choose the 2^{nd} to the L^{th} words as the modeling target. Note that a triangular-style mask is applied to the multimodal fusion module. This way, each text token can only attend to the context box tokens and the other text token before itself in the sequence. We apply the seq-to-seq objective loss L_{C} on the predicted sequence, where a word-level cross-entropy loss is applied on each predicted next word against the target word. Similar to Sec. 3.1, we also apply the same PointGroup loss L_{PG} for fine-tuning PointGroup. In the end, we combine the overall dense captioning loss as: $L_{\text{FC}} = L_{\text{PG}} + L_{\text{C}}$.

Joint training with both objectives. On top of the task-agnostic multimodal fusion module, our architecture can easily enable joint training with both bidirectional and seq-to-seq objectives. In this case, both output heads are attached to the multimodal fusion module. We first input the point cloud and the padded text into PointGroup and BERT to encode the box and text tokens, respectively. To accumulate gradients for both objectives before back-propagation, we apply two forward passes through the multimodal fusion module. In the first pass for the bidirectional objective, the global box token and bidirectional mask are applied. Similarly, the target box token and seq-to-seq mask are applied for the seq-to-seq objective in the second pass. We combine all losses for both objectives as: $L_{\text{FJ}} = L_{\text{PG}} + L_{\text{G}} + L_{\text{cls}} + L_{\text{C}}$.

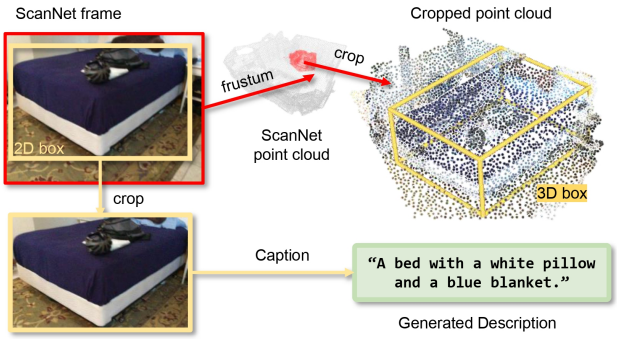


Figure 3: We propose a novel vision-language data synthesis method to enable more generic vision-language representation learning for 3D visual grounding and dense captioning, In particular, we crop out the dominant objects from the ScanNet frames and feed the image crops to CLIPCap [32], an off-the-shelf image captioner equipping CLIP [38] and GPT-2 [37]. In the end, we obtain the point-cloud-text pairs by combining the cropped point clouds in the camera frustums and the generated object captions.

3.2. 3D pre-training with synthetic data

Synthesize 3D vision-language data from images. Since the existing 3D vision-language data [5] is relatively limited in scale and variety, augmenting the training data with vision-language knowledge acquired from existing 2D datasets and models can potentially enable more generic vision-language representation learning for aforementioned 3D vision-language tasks. As illustrated in Fig. 3, given a ScanNet frame, we crop out the dominant objects and feed the image crops to CLIPCap [32], an image captioner empowered by pre-trained CLIP [38] and GPT-2 [37], to generate its description. Then, we crop out the points in the camera frustum from ScanNet point clouds as the 3D context. We pair the generated description and the cropped point cloud as a synthetic point-cloud pair for this view. To ensure the quality of the created point-cloud-text pairs, we remove pairs whose CLIP similarities between the original image crops and descriptions are lower than 0.3. We crop multiple objects from a single ScanNet frame, and repeatedly perform the description generation and point cloud cropping on every 20 frames in ScanNet. This process produces a large synthetic 3D vision-language dataset for the pre-training purpose.

Pre-training on synthetic data with both objectives. We feed the point cloud into a PointGroup [23] detection module to produce a sequence of encoded object tokens. To secure the quality of each object token, the PointGroup module is first trained on the ScanNet instance segmentation task until convergence. During pre-training, we freeze the PointGroup module and use the GT instance masks for encoding the object tokens. In the meantime, the object text query is fed into the BERT [13] to form a sequence of encoded text tokens. These two sets of tokens are concatenated as a sequence for vision-language modeling. Then, a multimodal transformer module takes the concatenated sequence to produce a sequence of fused multimodal tokens. To uniformly supervise the joint representation learning, we apply bidirectional objective ($L_{\text{G}} + L_{\text{cls}}$) and seq-to-seq objective (L_{C}) introduced in Sec. 3.1. We combine the overall pre-training loss function as: $L_{\text{PT}} = L_{\text{G}} + L_{\text{cls}} + L_{\text{C}}$.

As the synthetic 3D vision-language data contain many noisy samples, after convergence on the synthetic data, we continue the joint training with both objectives on the ScanRefer data with clearer annotations as a second pre-training stage. Then, we separately fine-tune the model on the 3D grounding task and 3D dense captioning task with bidirectional objective and seq-to-seq objective, respectively. We show that this pre-training scheme further boosts the overall accuracy on the downstream tasks.

3.3. Inference

We use task-specific heads for decoding visual grounding and dense captioning outputs from the task-agnostic multimodal fusion module. When inferencing the object captions, we autoregressively generate words from the [CLS] token until the [SEP] token or reaching the maximum length. Following Chen et al. [6], we take the minimum and maximum coordinates in the predicted instance masks to construct the object bounding boxes. For the bounding boxes that are assigned to the same GT box, we keep only the box with the highest IoU with the GT box. Note that the bounding boxes used for validating the detection and dense captioning performance are deliberately kept identical.

4. Experiments

4.1. Implementation details

Following Chen et al. [6], we use PointGroup implemented with the Minkowski Engine [10]. The PointGroup backbone is trained on the ScanNet [12] train set for 500 epochs using the Adam optimizer [25] with a learning rate of 2e-3 and a batch size of 4. In the pre-training stage on synthetic data, we acquire the object features with a trained and frozen PointGroup. For computational simplicity, we feed the PointGroup with GT instance masks. We consider up to 16 instances in the cropped point clouds. The pre-training takes 40k iterations to converge, with a learning rate of 1e-5 and a batch size of 128. During fine-tuning, we set the learning rate of PointGroup and the rest of the UniT3D to 1e-4 and 1e-5, respectively. To boost the fine-tuning speed, we pair each point cloud with 16 descriptions, as in Chen et al. [6]. Our full architecture contains 120M trainable parameters. All our experiments are conducted on an RTX A6000 GPU with PyTorch [33].

4.2. Dataset

We use the ScanRefer [5] dataset consisting of around 51k descriptions for over 11k objects in 800 ScanNet [12] scans for the visual grounding and dense captioning tasks. The descriptions include information about the appearance of the objects, as well as the object-to-object spatial relationships. We follow the official splits of ScanRefer for training and validation, and report visual grounding and dense captioning results on the validation split.

Synthetic data. We repeatedly generate the object descriptions in every 20 frames from ScanNet [12]. In each frame, we arrange the objects in descending order by the instance mask areas. Up to 3 most dominant objects are cropped for description generation. For consistency, we only generate synthetic data from scans in the ScanNet training split. In the end, we generate 265 693 synthetic samples from 92 766 frames in 1 199 ScanNet [12] scans. We visualize some synthetic samples in Fig. 4.

4.3. Evaluation metrics

Localization. Following Chen et al. [5], we measure the thresholded accuracy Acc@kIoU where the positive predictions have a higher intersection over union (IoU) with the ground truths than the thresholds. We set the threshold value k for IoU to 0.25 and 0.5 in our experiments.

Dense captioning. To jointly measure the quality of the generated descriptions and the detected bounding boxes, we evaluate them by evaluating standard image captioning metrics such as CIDEr and BLEU under different Intersection-over-Union (IoU) scores between predicted bounding boxes and the matched ground truth bounding boxes.

For N^{pred} predicted bounding boxes and N^{GT} ground truth bounding boxes, we define the captioning precision $M^{\text{P}}@k\text{IoU}$ as $M^{\text{P}}@k\text{IoU} = \frac{1}{N^{\text{pred}}} \sum_{i=1}^{N^{\text{pred}}} m_i u_i$, where $u_i \in \{0, 1\}$ is set to 1 if the IoU score for the i^{th} box is greater than k, otherwise 0. We use m to represent the captioning metrics such as CIDEr, BLEU, METEOR, and ROUGE.

Similarly, we define the captioning recall $M^{\text{R}}@k\text{IoU}$ as $M^{\text{R}}@k\text{IoU} = \frac{1}{N^{\text{GT}}} \sum_{i=1}^{N^{\text{GT}}} m_i u_i$. Note that previous dense captioning metrics proposed in Chen et al. [8] are analogous to $M^{\text{R}}@k\text{IoU}$. It solely measures the caption quality against the matched grounding truth, without taking the false positive predictions into account.

Figure 4: Our generated synthetic samples. Objects in clear views can be well captured by CLIPCap [32] as shown in the first three boxes in green. However, CLIPCap also fails to describe blurry or incomplete objects as shown in the last red box. Figure best viewed in color.

	Val Acc@0.25IoU			Val Acc@0.5IoU		
	Unique	Multiple	Overall	Unique	Multiple	Overall
ScanRefer [5]	76.33	32.73	41.19	53.51	21.11	27.40
TGNN [20]	68.61	29.84	37.37	56.80	23.18	29.70
InstanceRefer [47]	75.72	29.41	38.40	66.83	24.77	32.93
3DVG-Trans [50]	81.93	39.30	47.57	60.64	28.42	34.67
3DJCG [3]	78.75	40.13	47.62	61.30	30.08	36.14
D3Net [6]	-	-	-	72.04	27.11	35.58
3D-SPS [31]	84.12	40.32	48.82	66.72	29.82	36.98
BUTD-DETR [22]	82.77	**44.01**	**49.69**	63.81	**33.51**	38.01
Ours (from scratch)	**85.84**	32.21	42.31	**74.78**	27.60	36.49
Ours (w/ pre-training)	82.75	36.36	45.27	73.14	31.05	**39.14**

Table 1: Quantitative results on 3D visual grounding. We follow the evaluation setting in Chen et al. [5]. All accuracies are thresholded by the IoU 0.25 and 0.5. In comparison with the previous SOTA methods, our method has a notably higher overall grounding accuracy thresholded by IoU 0.5. Note that the grounding results from BUTD-DETR [22] are re-evaluated by removing the GT object labels in the text queries from the original implementation.

Finally, we adopt the captioning F1-score combining both captioning precision and captioning recall as the final metric for dense captioning:

$$M@k\text{IoU} = \frac{2 \times M^{\text{P}}@k\text{IoU} \times M^{\text{R}}@k\text{IoU}}{M^{\text{P}}@k\text{IoU} + M^{\text{R}}@k\text{IoU}} \quad (1)$$

Object detection. We also report the mean average precision (mAP) of the detected objects from our PointGroup backbone on ScanRefer val split. These detected boxes are the same as the ones used for dense captioning evaluation.

4.4. Comparison with the state-of-the-art methods

We compare our proposed UniT3D method with several state-of-the-art methods for 3D visual grounding and dense captioning tasks on ScanRefer dataset [5] in Tab. 1 and Tab. 3. For 3D visual grounding, we divide the validation set into "Unique", "Multiple", and "Overall", following the evaluation protocol in Chen et al. [5]. For 3D dense captioning, we report the aforementioned dense captioning F1-score and the object detection mAP.

3D visual grounding. Tab. 1 compares our method with the prior state-of-the-art 3D visual grounding methods on ScanRefer. For our method trained from scratch (Ours (from scratch)), we observe that it already achieves on-par performance with the previous SOTA method [22]. In this case, our method even achieves the best grounding accuracy in the "Unique" subset, where there is only one unique object belonging to a specific semantic class in the scene. After pre-training the network on the synthetic 3D vision-language data and fine-tuning on the visual grounding tasks (Ours (w/ pre-training)), we observe a clear performance boost in the visual grounding accuracies. Despite a drop in the "Unique" subset, our method performs clearly better in the more challenging "Multiple" subset, resulting in an improvement in the overall grounding accuracy. The improvement in "Multiple" subset indicates that pre-training on a large amount of synthetic data provides more multimodal knowledge for disambiguating the objects in the scene with language cues. Note that BUTD-DETR [22] is re-evaluated by removing the GT object labels in the text queries for a fair comparison (more details in supplementary material). Besides the modulated object detector, BUTD-DETR also takes pre-computed object bounding boxes as extra visual cues. In contrast, our method follows a much simpler design philosophy and relies purely on the end-to-end fine-tuned object detector. We note that the baseline joint 3DJCG [3] and D3Net [6] involves either sophisticatedly designed heavy neural heads or complicated self-critical training strategy. Our method outperforms both of them with a significantly simpler architecture design.

3D dense captioning. We present a comparison of our method against the previous state-of-the-art 3D dense captioning methods on ScanRefer in Tab. 3. To keep the comparison consistent and fair, all methods presented here are trained with the cross-entropy objective only, including D3Net [6]. Even without any pre-training, our architecture (Ours (from scratch)) already achieves similar dense captioning results in comparison with the previous SOTA D3Net [6]. We observe a significant performance boost if we initialize the network with the pre-trained weights on

# objects	Visual Grounding Accuracy			Dense Captioning F1-Scores			
	Unique@0.5IoU	Multiple@0.5IoU	Overall@0.5IoU	CIDEr@0.5IoU	BLEU-4@0.5IoU	ROUGE-L@0.5IoU	METEOR@0.5IoU
up to 1	26.72	7.24	11.02	3.90	0.53	18.70	8.17
up to 3	**30.84**	**9.34**	**13.51**	**5.91**	**0.69**	**18.91**	**8.32**
up to 5	29.54	7.62	11.87	2.81	0.29	18.22	7.89

Table 2: Comparison of 3D visual grounding and dense captioning results evaluated on ScanRefer [5] val set with UniT3D pre-trained on different amount of synthetic data. When using synthetic data generated from up to 3 object crops in the ScanNet [12] frames, the pre-trained UniT3D demonstrates the best initial performance in comparison with other object cropping configurations.

	Captioning F1-score				Detection
	C@0.5IoU	B-4@0.5IoU	R@0.5IoU	M@0.5IoU	mAP@0.5
Scan2Cap [8]	15.71	9.01	14.92	7.18	32.09
X-Trans2Cap [48]	17.64	9.68	15.25	7.21	35.31
MORE [24]	16.46	8.86	14.71	7.12	31.93
3DJCG [3]	21.17	14.18	19.49	10.19	39.75
D3Net [8]	26.13	16.18	27.48	13.06	50.93
D3Net [8] (CIDEr loss)	41.32	22.75	35.30	15.87	53.85
Ours (from scratch)	26.68	14.64	27.10	12.92	53.91
Ours (w/ pre-training)	**30.28**	**18.23**	**30.72**	**14.74**	**54.03**

Table 3: Quantitative results of 3D dense captioning on ScanRefer [5]. We measure the dense captioning F1-scores and the PointGroup detection mAP against the ground truth bounding boxes and descriptions. All reported metrics are thresholded by IoU 0.5. Our method outperforms the previous SOTA methods even without any pre-training, and our proposed pre-training scheme on the synthetic data gives a further improvement the dense captioning performance. Note that we compare to D3Net [6] trained only with the cross-entropy objective for a fair comparison.

the synthetic data (Ours (w/ pre-training)). We observe that our method clearly outperforms previous joint architectures D3Net and 3DJCG [3] without any complex design in the architecture or training objectives. Note that D3Net [6] applies the same PointGroup detection backbone as in our method. Compared to D3Net [6], the results show that our method is more capable of modeling the multimodal and spatial relationships from similar visual cues.

4.5. Ablation and analysis

Does more synthetic data help? Since synthetic data can be acquired at almost zero cost in terms of human annotation, there is no theoretical upper bound for the amount of synthetic data we can generate. We conduct an ablation experiment where we pre-train the proposed network on the different amounts of synthetic data. We crop up to 1, 3, or 5 dominant objects from the ScanNet [12] frames to generate synthetic descriptions. The proposed network is pre-trained on these three synthetic datasets and evaluated on the ScanRefer [5] validation set afterwards. As shown in Tab. 2, we empirically find that the network pre-trained with the dataset where up to 3 objects are cropped from the frame achieves the best performance for both visual grounding and dense captioning. This indicates that more synthetic data is not necessarily helpful since more noise resulting from it could overwhelm the network in the pre-training stage.

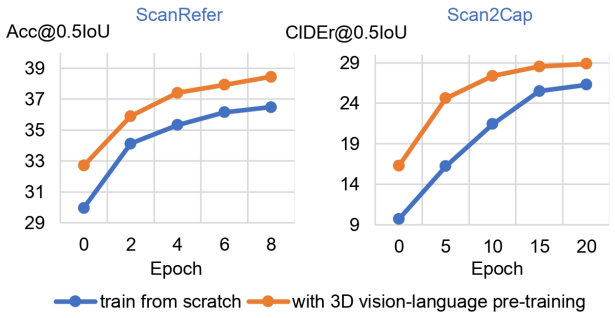


Figure 5: Compared to training from scratch, the pre-training scheme on synthetic data with joint training objectives shows notable benefits on both 3D dense captioning and visual grounding.

Do pre-training and joint training objectives help? We report the ablation results on different pre-training schemes for visual grounding and dense captioning in Tab. 4 and Tab. 5, respectively. In general, we validate 5 scenarios: (a) training the network from scratch on each task in ScanRefer with target objective; (b) training the network from scratch on both tasks in ScanRefer with joint objectives; (c) pre-training the network on synthetic data without further any fine-tuning; (d) fine-tuning the network from (c) directly on the downstream objective; and (e) our full setting – fine-tuning the network from (c) joint training objectives on ScanRefer.

For visual grounding in Tab. 4, our model in (a) already demonstrates strong performance on par with the previous SOTA even without pre-training. And the visual grounding accuracy is further improved by having the joint training objectives on ScanRefer in (b), showing that our unified architecture effectively connects the two tasks through joint learning. After being pre-trained only on the synthetic data, our model in (c) demonstrates non-trivial results when directly evaluated on the ScanRefer [5] validation split. Then, the model fine-tuned with only the bidirectional objective in (d) achieves the highest visual grounding accuracy in the "Unique" split, indicating a great capability of distinguishing the objects via semantic information. Fine-tuned jointly with both training objectives, our final model in (e)

Training setup	Training Dataset(s)		Training Objective(s)		Visual Grounding Accuracy		
	Synthetic	ScanRefer	Bidirectional	Seq-to-Seq	Unique@0.5IoU	Multiple@0.5IoU	Overall@0.5IoU
(a) direct from scratch		✓	✓		74.78	27.60	36.49
(b) joint from scratch		✓	✓	✓	73.68	28.84	37.45
(c) initial pre-trained	✓		✓	✓	30.84	9.34	13.51
(d) direct fine-tuned		✓	✓		**75.51**	29.63	38.45
(e) joint fine-tuned		✓	✓	✓	73.14	**31.05**	**39.14**

Table 4: 3D visual grounding results on ScanRefer [5] with different training schemes. (a) Without any pre-training, our model already has a strong performance on par with the previous SOTA. (b) The visual grounding accuracy is further improved by having the joint training objectives, indicating the effectiveness of our architecture in unifying two downstream tasks. (c) Pre-trained on the synthetic data alone gives non-trivial visual grounding performance on the ScanRefer [5] validation split. (d) Fine-tuned with bidirectional objective from weights in (c), it achieves the highest visual grounding accuracy in the "Unique" subset. (e) Fine-tuned jointly with both objectives, our final setting in achieves the highest visual grounding accuracy in the "Multiple" subset, leading to the best overall visual grounding results.

Training setup	Training Dataset(s)		Training Objective(s)		Dense Captioning F1-Scores			
	Synthetic	ScanRefer	Bidirectional	Seq-to-Seq	CIDEr@0.5IoU	BLEU-4@0.5IoU	ROUGE-L@0.5IoU	METEOR@0.5IoU
(a) direct from scratch		✓	✓		26.68	14.64	27.10	12.92
(b) joint from scratch		✓	✓	✓	27.28	17.22	29.12	13.74
(c) initial pre-trained	✓		✓	✓	5.91	0.69	18.91	8.32
(d) direct fine-tuned		✓		✓	28.13	17.69	30.33	14.30
(e) joint fine-tuned		✓	✓	✓	**30.28**	**18.23**	**30.72**	**14.74**

Table 5: 3D dense captioning results on ScanRefer [5] with different training schemes. (a) Our model without pre-training already demonstrates competitive performance against the previous SOTA. (b) The dense captioning accuracy is further improved by having the joint training objectives, indicating the effectiveness of our architecture in unifying two downstream tasks. (c) Due to the domain gap, the descriptions generated by the model pre-trained only on the synthetic data obviously deviate from the GT samples in ScanRefer [5] validation split. (d) Plausible dense captioning performance is achieved after being directly fine-tuned on ScanRefer [5] with the seq-to-seq objective. (e) A further performance boost is observed if the network is fine-tuned jointly with both training objectives in our final setting.

achieves the highest visual grounding accuracy in the "Multiple" subset, leading to the best overall visual grounding results. Compared with directly training from scratch in (a), this pre-training and joint optimization scheme clearly demonstrates its advantage.

For dense captioning in Tab. 5, our method in (a) already demonstrates competitive performance against the previous SOTA without pre-training, and can be further improved by having the joint training objectives on ScanRefer in (b), indicating the effectiveness of our architecture in unifying two downstream tasks. Due to the domain gap, the descriptions generated by the model after pre-training only on the synthetic data in (c) obviously deviate from the GT samples in ScanRefer [5] validation split. Plausible dense captioning performance is achieved after the network is directly fine-tuned on ScanRefer [5] with the seq-to-seq objective in (d). A further performance boost is observed if the pre-trained network is fine-tuned jointly with both the bidirectional and the seq-to-seq objectives in (e). Here, we also demonstrate the strong advantage of having the pre-training and joint optimization scheme in Fig. 5. Compared to training from scratch, our pre-training scheme on synthetic data with joint training objectives effectively improves the overall accuracy on 3D dense captioning and visual grounding.

5. Conclusion

We present UniT3D, a unified transformer architecture to connect 3D dense captioning and visual grounding. UniT3D enables learning a strong joint multimodal representation across two tasks through a supervised joint pre-training scheme with bidirectional and seq-to-seq objectives. The generic representation of UniT3D expands pre-training scope to more various training sources such as the synthesized data via 2D priors, showing that the distillate 2D knowledge is beneficial to 3D vision-language tasks. Extensive experiments and analysis demonstrate the strength of our UniT3D model for 3D dense captioning and visual grounding. We hope our work can inspire more future work in exploring the 3D vision-language field.

Limitations. Although UniT3D takes a promising step towards unifying the two discussed tasks, UniT3D could be still expanded to other tasks. Further, in addition to generating synthetic data from a 2D captioner, more sources of distillate 2D data could be explored in the future.

Acknowledgements

This work is funded by the ERC Starting Grant Scan2CAD (804724), a Hans Fischer Fellowships (Focus Group Visual Computing), as well as the the German Research Foundation (DFG) under the Grant *Making Machine Learning on Static and Dynamic 3D Data Practical*. This work is also supported in part by the Canada CIFAR AI Chair program and an NSERC Discovery Grant.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *European Conference on Computer Vision*, pages 422–440. Springer, 2020. 2

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[3] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 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*, pages 16464–16473, 2022. 1, 2, 6, 7

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[6] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. *arXiv preprint arXiv:2112.01551*, 2021. 1, 2, 5, 6, 7

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[8] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 1, 2, 5, 7

[9] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3D object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8963–8972, 2021. 1

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 5

[11] Angela Dai and Matthias Nießner. 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 1

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5, 7

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 3, 5

[14] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-proposal aggregation for 3D semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 1

[15] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *6th Annual Conference on Robot Learning*, 2022. 3

[16] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3D understanding. *arXiv preprint arXiv:2111.15363*, 2021. 3

[17] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 1

[18] Ji Hou, Angela Dai, and Matthias Nießner. RevealNet: Seeing behind objects in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 1

[19] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 2, 3

[20] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3D instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021. 2, 6

[21] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv preprint arXiv:2210.01055*, 2022. 3

[22] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022. 6

[23] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-set point grouping for 3D instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, pages 4867–4876, 2020. 3, 4, 5

[24] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. *arXiv preprint arXiv:2203.05203*, 2022. 2, 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3D instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019. 1

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[29] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3D-to-2D distillation for indoor scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4464–4474, 2021. 3

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[31] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 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*, pages 16454–16463, 2022. 6

[32] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4, 6

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[34] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3D object detection from RGB-D data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 1

[35] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep Hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1

[36] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. ImVoteNet: Boosting 3D object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 1

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[40] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[42] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3

[43] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2

[44] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[45] Chenfeng Xu, Shijia Yang, Bohan Zhai, Bichen Wu, Xiangyu Yue, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3D point-cloud understanding with pretrained 2D convnets. *arXiv preprint arXiv:2106.04180*, 2021. 3

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 1

[47] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 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*, pages 1791–1800, 2021. 2, 6

[48] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-

modal knowledge transfer using transformer for 3D dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 2, 7

[49] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 3

[50] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 2, 6

[51] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2