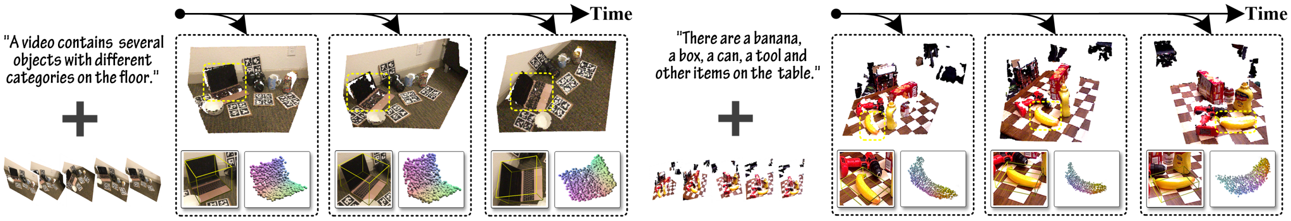
L4D-Track: Language-to-4D Modeling Towards 6-DoF Tracking and Shape Reconstruction in 3D Point Cloud Stream

Jingtao Sun[1,2] Yaonan Wang[1*] Mingtao Feng[3] Yulan Guo[4] Ajmal Mian[5] Mike Zheng Shou[2*]
[1]NERC-RVC, Hunan University [2]Show Lab, National University of Singapore
[3]Xidian University [4]Sun Yat-Sen University [5]The University of Western Australia


"A video contains several objects with different categories on the floor."
+
Time
"There are a banana, a box, a can, a tool and other items on the table."
+
Time

(a) Unseen objects with known categories (*e.g., laptop*). (b) Unseen objects with unknown categories (*e.g., banana*).

Figure 1. Given a 3D point cloud stream and the language-3D captions, our method achieves real-time, causal 6-DoF pose tracking while reconstructing the 3D shape in the current observation. We demonstrate that: **(a)** our method not only enables zero-shot inference for unseen objects with known categories, **(b)** but also perfectly showcases the zero-shot capabilities for unseen objects with unknown classes.

Abstract

3D visual language multi-modal modeling plays an important role in actual human-computer interaction. However, the inaccessibility of large-scale 3D-language pairs restricts their applicability in real-world scenarios. In this paper, we aim to handle a real-time multi-task for 6-DoF pose tracking of unknown objects, leveraging 3D-language pre-training scheme from a series of 3D point cloud video streams, while simultaneously performing 3D shape reconstruction in current observation. To this end, we present a generic Language-to-4D modeling paradigm termed L4D-Track, that tackles zero-shot 6-DoF Tracking and shape reconstruction by learning pairwise implicit 3D representation and multi-level multi-modal alignment. Our method constitutes two core parts. 1) Pairwise Implicit 3D Space Representation, that establishes spatial-temporal to language coherence descriptions across continuous 3D point cloud video. 2) Language-to-4D Association and Contrastive Alignment, enables multi-modality semantic connections between 3D point cloud video and language. Our method trained exclusively on public NOCS-REAL275 dataset, achieves promising results on both two publicly benchmarks. This not only shows powerful generalization performance, but also proves its remarkable capability in zero-shot inference. The project is released at L4D-Track.

1. Introduction

The integration of the 3D physical world with coherent natural language constitutes a pivotal advancement in field of 3D computer vision and robotics, playing a crucial role in the domains of embodied artificial intelligence [14]. Recent research endeavors have increasingly focused on the task of 3D vision-language (3D-VL) learning. These tasks contain a wide array of objectives, *e.g.,* 3D visual grounding [30], dense captioning [56] and question answering [1], often employing task-specific model designs.

However, the absence of large-scale human-annotated 3D-text pairs has led to the prevalent adoption of an alternative approach [53, 61], that involves pre-training a generic 3D model leveraging large scale image-text paired data, alongside 2D-3D back-projection. Despite the existing approaches have made the promising progress in terms of handling 3D visual understanding, this line of methods suffers from several major challenges. Firstly, the paradigm of mapping 3D data into 2D modalities utilizing the vision-language (VL) foundation models, leads to lose the vital information in 3D space and renders heavier computational and memory costs. Secondly, given the irregular and unstructured nature of 3D point cloud data, a complete 3D-language foundation that learns a unified 3D representation emerges as a more viable option compared to the indirect VL-based approach. Futhermore, existing methods are only capable of understanding concrete

*Corresponding Author

instances with known classes in single-frame 3D data, thus exhibiting limited performance in real-world scenes with unseen objects. Moreover, few studies have delved into the modeling from language-to-3D video (point cloud stream) in the existing literatures [8, 20]. On the other hand, 6-DoF pose tracking and 3D reconstruction from a 3D video are also the fundamental open problems, and the intra-class shape variation restricts their ability for category-level pose predictions. While these category-level methods [43, 49, 50] enable generalization to new objects within the same category, they face difficulties when confronted with out-of-distribution unseen instances with unknown categories.

Thus, we wondered *if it was possible to establish an unified paradigm bridging 3D video and language for zero-shot simultaneous 6-DoF pose tracking and 3D shape reconstructing in a real-time manner.* In this work, our core idea is to discover and construct a model linking 3D video and language descriptions to distill semantic-rich language captions and compensate for lack of language-3D data pairs in the 3D domain. In the end, we introduce a language-to-4D modeling paradigm that seeks to establish fine-grained interactions between 3D point cloud steam and their corresponding language captions. This enables learning of universal multi-modal inter-frame feature pairs for zero-shot tasks involving 6-DoF pose tracking and 3D shape reconsrtuction. As depicted in Fig. 1, given the point cloud stream, we first propose a Pairwise Implicit 3D Space Representation module (Sec. 3.2) to construct an implicit distribution for the change of 6-DoF pose and 3D shape field between consecutive frames. We futher propose a GPT-assisted Language-to-4D Association module (Sec. 3.3) that bridges the modal association between 3D point cloud video and language semantics. To align 3D geometric feature pairs with language features from spatial semantics to temporal perspectives, we propose to build multi-level contrastive alignment (Sec. 3.4) to make our approach more generalizable. Through experiments on two public datasets, we demonstrate the effectiveness of our proposed method. In conclusion, our main contributions are:

- We introduce a language-4D modeling paradigm that learns multi-modal characteristics to achieve object 6-DoF pose tracking and the corresponding 3D shape reconstruction in a real-tme manner.
- We propose a pairwise pose/shape 3D space implicit representation strategy, ensuring that the learned multi-modal feature pairs are both language-aligned and spatio-temporally coherent for every frame.
- We propose a GPT-assisted 3D point cloud video-language association and alignment scheme to tackle the limitations of lack of language-to-3D paired data, while guaranteeing the zero-shot generalization to unseen instances with known or unknown classes.

2. Related Work

Object 6-DoF Pose Estimation and Tracking. The existing studies related to 6-DoF pose estimation have primarily focused on either the instance-level [3, 12] or the category-level [7, 17, 21, 22, 54, 60, 62]. These methods operate under the assumption of the availability or unavailability of precise CAD models. Several recent works conducted the Normalized Object Coordinate Space (NOCS) [45] for all instances within the same category. Aside from single-frame pose estimation, researchers have exploited pose tracking techniques that leverage temporal information to estimate inter-frame change of poses across 2D video, *i.e.,* 6-PACK [43], CAPTRA [50], BundleTrack [47], ICK-Track [39], CatTrack [55], and BundleSDF [49]. Our objective is to address the zero-shot tasks of categorical pose tracking and shape reconstruction from 3D point cloud stream (3D video). This entails the ability to generalize from annotated (seen) object classes to other (unseen) categories, with guidance from languages.

3D Vision-Language Pre-training. Recently, Vision-Language Pre-training (VLP) has received significant attention in the 2D domain [15, 25, 36, 38, 57]. Thanks to the vast repositories of publicly available datasets, which comprise billions of web-crawled images with semantic-rich annotations, VLP methods have made it possible to establish meaningful image-text embeddings. Notable examples include VinVL [59] and VILLA [10]. However, the domain of 3D vision-language pre-training (3D-VLP) focused on directly learning a unified embedding space has not yet been explored in the current literatures [8, 20]. Initial attempts by Pointclip [61], CLIP2 [58] and ULIP [53] aimed to project 3D data into 2D modalities (*e.g., RGB(D) or depth*) using the 2D-VLP foundations. In our work, we also focus on addressing 3D understanding challenges by learning the pairwise modeling of 3D video, encompassing spatio-temporal consistency and language descriptions.

Neural Implicit Representation. Neural implicit representation functions have found extensive application in various domains, addressing challenging problem statements such as View Synthesis [40], 3D Rendering [33]. NeRF [31] introduced the concept of a neural radiance field, representing a static scene as a 5D function that provides radiance information in terms of direction, point, and density. NeRF and it variants [9,11,26,28,51] have achieved remarkable generalization ability across a wide range of tasks, particularly in dynamic scene synthesis studies. With such great advances in existing NeRF-related technologies, researchers have recently delved into utilizing NeRFs for 3D locialization [13, 19], 3D semantic segmentation [63] and applications in robotics [27]. Our work leverages the power of neural implicit fields to model the continuous spatio-temporal consistency within dynamic 3D video.

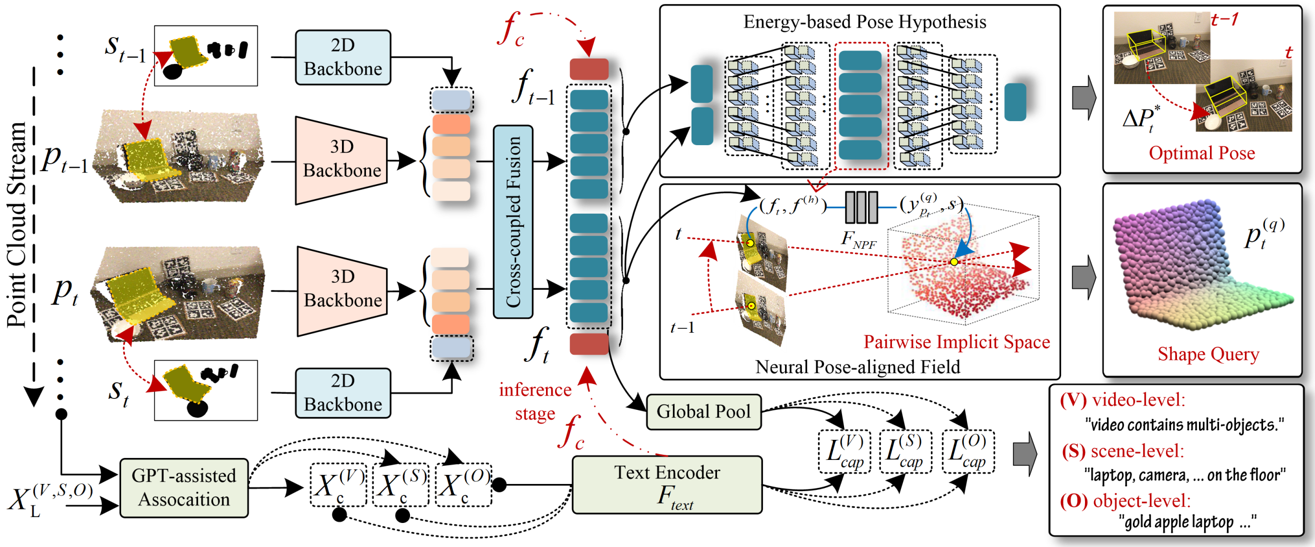


Figure 2. Illustration of the pipeline of proposed methodology. Given input point cloud stream along with the corresponding segmented mask, we first encode them with both 2D/3D backbone separately and a cross-coupled fusion module to obtain inter-frame embeddings f_{t-1}, f_t. These paired embeddings are then used to model the energy-based hypothesis about changes in pose and learn a neural pose-aligned field that generates shape query while aligning its pose for an arbitrary object. Meanwhile, these embeddings will be aligned with the extra input multi-level language captions using proposed GPT-assisted assocaition and alignment modules to achieve zero-shot inference. It's noteworthy that the caption embeddings f_c are added into f_{t-1}, f_t to enhance its performance during the inference stage.

3. Methodology

3.1. Preliminary

Problem Formulation. In this paper, we address a challenging zero-shot task: the simultaneous tracking of categorical 6-DoF poses and the corresponding 3D shape reconstruction for previously unseen instances in a point cloud video. This is accomplished by leveraging natural language instructions and a Vision-Language (VL) foundation model. Formally, given a point cloud video $P = \{p_t\}$ along with its segmentation masks, and a natural language instuction list X_L, our objective is to estimate and track the 6-DoF pose and 3D shape of the observable objects descirbed from the language instructions X_L. The 6-DoF pose, denoted as $\mathcal{P} = \{T, R\} \in \mathbf{SE}(3)$, encompasses the translation vector $T \in \mathbb{R}^3$ and the rotation matrix $R \in \mathbf{SO}(3)$. We aim to track the change of pose $\Delta\mathcal{P}_t$ and subsequently reconsrtuct the corresponding 3D shape $p_t^{(q)}$ of previously unseen object in the online manner. Given the estimated pose from previous frame $\mathcal{P}_{t-1}$, our method needs to estimate $\mathcal{P}_t$:

$$\mathcal{P}_t = \Delta\mathcal{P}_t \cdot \mathcal{P}_{t-1} = \Delta\mathcal{P}_t \cdot \Delta\mathcal{P}_{t-1} \cdots \mathcal{P}_0. \qquad (1)$$

An overview of our method is depicted in Fig. 2. Given the inter-frame pairs of observable inputs $(p_{t-1},\ p_t)$ from a 3D video, along with the segmented mask $(s_{t-1},\ s_t)$ of arbitrary object of interest, we first extract spatial features by a 3D backbone (PointNet [35]). Meanwhile, the masks are converted into image features by a 2D backbone, which are back-projected and added to spatial features. To build the pairwise connection between continous frames, the pairs of these features $\bar{f}_{t-1}, \bar{f}_t$ are then fed into a cross-attention based cross-coupled fusion module to construct the pair of inter-frame embeddings $f_{t-1}, f_t \in \mathbb{R}^K$:

$$\begin{cases} f_{t-1} = \sum\limits_{i\in\Omega_t} \alpha((\Gamma(\bar{f}_{t-1})+\Delta)\cdot\Gamma(\bar{f}_t^i)/\sqrt{K})\cdot\Gamma(\bar{f}_t^i) \\ f_t = \sum\limits_{i\in\Omega_{t-1}} \alpha((\Gamma(\bar{f}_t)-\Delta)\cdot\Gamma(\bar{f}_{t-1}^i)/\sqrt{K})\cdot\Gamma(\bar{f}_{t-1}^i) \\ \Delta = linear(\bar{f}_t - \bar{f}_{t-1}) \end{cases}, \qquad (2)$$

where Γ are the linear projection layer, α denotes softmax operation and Ω is feature space. Δ represents the inter-frame point feature difference, which is used to offset data drifting between consecutive frames.

3.2. Pairwise Implicit 3D Space Representation

Energy-based Pose Hypothesis. Due to there may be ambiguities when inferring the 6-DoF pose's change given two 3D views, we introduce an energy-based formulation that can model these uncertainty. Given a pair of inter-frame embeddings depicting an arbitrary object, we wish to recover a pose hypothesis distribution over the relative change of pose between two consecutive timesteps $t-1$ and t. Inspired by the idea of implicitly representing the distribution using a neural network in recent work [32], we propose modelling this conditional distribution of 6-DoF pose hypotheses $P(\Delta\mathcal{P}_t^{(h)}|f_{t-1}, f_t)$ as the unnormalized joint log-probability (*w.r.t., energy*):

$$P(\Delta\mathcal{P}_t^{(h)}, f_{t-1}, f_t) = \sigma \cdot exp(F_\eta(\Delta\mathcal{P}_t^{(h)}, f_{t-1}, f_t)). \qquad (3)$$

To this end, we aim to predict this energy via training a network F_η parameterised by η. Where σ is the constant of integration. From the product rule, these distribution can also be represented as:

$$\begin{aligned} P(\Delta\mathcal{P}_t^{(h)}|f_{t-1},f_t) &= \frac{P(\Delta\mathcal{P}_t^{(h)},f_{t-1},f_t)}{P(f_{t-1},f_t)} \\ &= \frac{exp\cdot F_\eta(\Delta\mathcal{P}_t^{(h)},f_{t-1},f_t)}{\sum\limits_i exp\cdot F_\eta((\Delta\mathcal{P}_t^{(h)})_i,f_{t-1},f_t)}, \end{aligned} \tag{4}$$

according to the Eq. (4), we compute these conditional distribution $\Delta P_t^{(h)}$ of 6-DoF pose change by sampling multiple homogeneous pose transformation matrices over $\mathbf{SO(3)}$ and $\mathbb{R}^3$, and the number of pose sampling matrices should be large for accurate approximation. Ultimately, we recover the optimal change of pose from frame $t-1$ to t by optimizing F_η over the 6D space of implicit hypothesis:

$$\begin{aligned} \Delta\mathcal{P}_t^* &= \underset{\Delta\mathcal{P}_t^{(h)}\in\mathbf{SE(3)}}{\arg\max} \quad P(\Delta\mathcal{P}_t^{(h)}|f_{t-1},f_t) \\ &= \underset{\Delta\mathcal{P}_t^{(h)}\in\mathbf{SE(3)}}{\arg\max} \quad F_\eta(\Delta\mathcal{P}_t^{(h)},f_{t-1},f_t), \end{aligned} \tag{5}$$

we train above network F_η by minimizing the negative log-likelihood:

$$L_{pose} = -\log P(\tilde{\mathcal{P}}_t\cdot\tilde{\mathcal{P}}_{t-1}^{-1}|f_{t-1},f_t), \tag{6}$$

where $\tilde{\mathcal{P}}_t$ and $\tilde{\mathcal{P}}_{t-1}$ are the ground-truth pose of the targeted object in frame $t-1$ and t. $\tilde{\mathcal{P}}_t\cdot\tilde{\mathcal{P}}_{t-1}^{-1}$ denotes the pose change from frame $t-1$ to t.

Neural Pose-aligned Fields (NPFs). The key to our approach is learning a pairwise implicit field that learns inter-frame consistent 3D query shape with pose alignment properties, that is related to the current observation. This correspondences are established using NPF defined as a mapping from the pairwise feature f_t and the pose hypothesis feature $f^{(h)}$ to the corresponding point $y_{p_t}^{(q)}$ in 3D query shape $p_t^{(q)}$ in mentioned implict 3D space, and its signed distance s:

$$F_{NPF}(f_t,f^{(h)};\theta) = (y_{p_t}^{(q)},s), \tag{7}$$

$$F_{NPF}:\mathbb{R}^K\times\mathbb{R}^K\mapsto\mathbb{R}^3\times\mathbb{R}, \tag{8}$$

where F_{NPF} denotes a full-connected neural network parameterised by θ. As shown in Fig. 2, the pose hypothesis feature $f^{(h)}$ is extracted from the encoder of the network F_η, as shown in Fig. 2. $y_{p_t}^{(q)}$ denotes the point of generated shape query. Owing to the 3D query shape is sampled in the same coordinate frame with the corresponding shape prior, the signed distance s is defined as the distance between $y_{p_t}^{(q)}$ and the surface of shape prior p_r, which is generated using the Shape Auto-Encoder proposed in [17] and the number of its point is set to N_p. With these field, we can then calculate shape regress loss as Eq. (10) and L_1 loss for s similar to [13]:

$$L_s = \frac{1}{N}\sum_i|clamp(\varphi(\tilde{y}_i),\delta) - clamp(s,\delta)|, \tag{9}$$

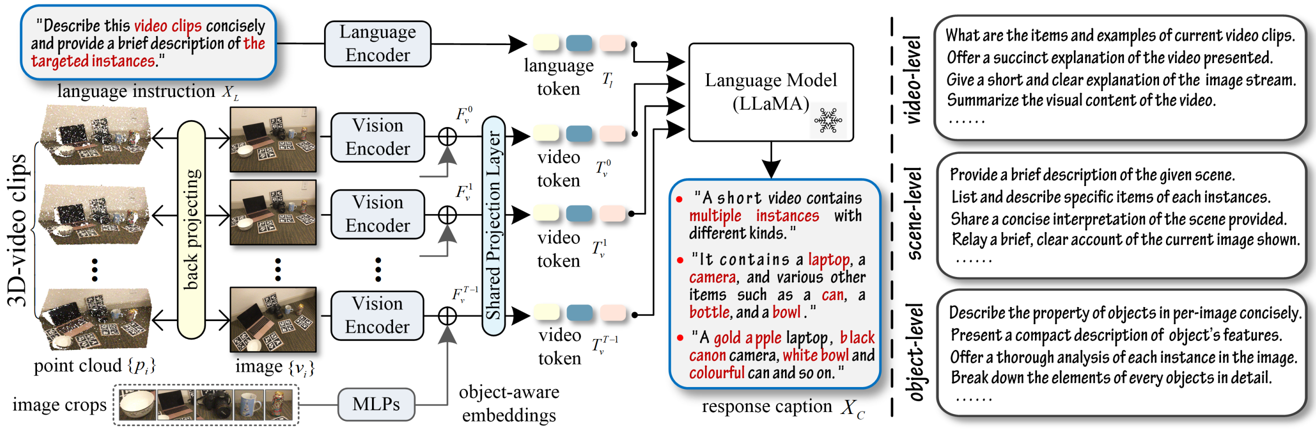
$$\begin{aligned} L_{shape} = &\frac{1}{N_p}\sum_{x_i\in p_r}\min_{x_j\in p_t^{(q)}}||x_i-x_j||_2^2 \\ &+\frac{1}{N}\sum_{x_i\in p_t^{(q)}}\min_{x_j\in p_r}||x_j-x_i||_2^2, \end{aligned} \tag{10}$$

where $\tilde{y}_i$ is the corresponding ground-truth 3D point and δ is a clamping parameter to maintain a metric SDF, as in [33]. We jointly train the energy-based model and NPFs model by the following objective:

$$\eta^*,\theta^* = \underset{\eta,\theta}{\arg\min}(L_{pose}+L_{shape}+L_s). \tag{11}$$

3.3. GPT-assisted Language-to-4D Association

The primary goal of this module is to enable meaningful bidirectional interaction between language and 3D point cloud stream and expect to extablish a universal 3D video-language association to address the inaccessibility of large-scale 3D-language data pairs. Following the statement in [29], a general-purpose assistant that follows the multi-modal vision-and-language instructions, which can effectively improve the zero- and few-shot generalization abilities. To this end, we introduce a GPT-assisted association manner to link the language supervision to all frames in the given 3D video clips (see in Fig. 3 (a)). This module play a crucial role in enhancing the zero-shot inference for pose tracking and shape reconstruction. Here, we first back-project each 3D video clip $P = \{p_0,p_1,\ldots,p_{\mathrm{T}-1}\}$ to the corresponding image space V using the depth information and produce a series of consecutive image frames. Given the image clips $V = \{v_0,v_1,\ldots,v_{\mathrm{T}-1}\}$ with T frames, we adopt BLIP-2 [24], a ViT backbone coupled with Q-Former, as the pre-trained visual encoder to provide the visual feature F_v^i for per-image frame, and F_v^i is also concatnated with the corresponding object-aware embeddings, which generated from all image crops in each frame. And then, a shared linear layer is considered as the trainable projection to convert these visual features into video embedding tokens $T_v^i\in R^{T\times H^i\times W^i\times 300}$, individually, which have the same dimensionality of the language embedding space from the language encoder. Moreover, we encode the input language instruction X_L with GloVe [34] to obtain the language embedding token $T_l\in R^{L\times 300}$, where L is the number



"Describe this video clips concisely and provide a brief description of the targeted instances."

- "A short video contains multiple instances with different kinds."
- "It contains a laptop, a camera, and various other items such as a can, a bottle, and a bowl."
- "A gold apple laptop, black canon camera, white bowl and colourful can and so on."

What are the items and examples of current video clips.
Offer a succinct explanation of the video presented.
Give a short and clear explanation of the image stream.
Summarize the visual content of the video.
......

Provide a brief description of the given scene.
List and describe specific items of each instances.
Share a concise interpretation of the scene provided.
Relay a brief, clear account of the current image shown.
......

Describe the property of objects in per-image concisely.
Present a compact description of object's features.
Offer a thorough analysis of each instance in the image.
Break down the elements of every objects in detail.
......

(a) The pipline of proposed association module. (b) Multi-level instructions.

Figure 3. Illustration of GPT-assisted association module (Sec. 3.3). (a) It mainly consists of a vision encoder, a language encoder, the shared linear projection layer, and an large language model. (b) We present multi-level language instructions with video-, scene- and object-level association manners to assign 3D video with mixed caption supervision.

of words in a language instruction. Finally, we choose LLaMA [42] as our basic language model parameterized by Φ, that predicts the corresponding response caption X_C:

$$X_c = F_\Phi(T_v^i, T_l),\ with\ T_v^i = W \cdot v_i, \tag{12}$$

Though we can obtain a acceptable language-to-4D association tool with the functions in Eq. (12), it is actually just a generic way, incapable of enabling the former 6-DoF tracking backbone to access abundant language features with the rich and detailed object-centric descriptions. In this regard, we hope to build a multi-level instruction-caption pairs, including the level of video, scene and object. As depicted in Fig. 3 (b), these multi-level instruction-caption pairs $(X_L^{(V,S,O)} \sim X_C^{(V,S,O)})$ can offer adequate captions from video content, spatial scene relationships to the details of instances (*e,g., shape, texture and colour*). Besides, in order to generate available instruction-following ground-truth caption pairs in the used datasets, we propose to leverage ChatGPT/GPT-4 to semi-automatically collect these instruction-caption pairs for the following pre-training. Finally, we only keep the LLM weight frozen in LLaMA and continue to update the weights of linear layer, language and visual encoder, and finetune our association model with these instruction-following data.

3.4. Contrastive Multi-Level Language Alignment

With the multi-level captions $X_C^{(V,S,O)}$, we are now ready to guide our core network to align with diverse language-instructions. Since the inter-frame feature pairs f_{t-1}, f_t and language captions X_C are generated separately in their own spaces and directly aligning them will lead to ambiguities. Meanwhile, unlike the conventional language-video 2D image-based modeling and aligning [16], a 3D video stream requires more complex spatial propagation. To this end, we introduce a 3D video-language constrastive alignment strategy that aligns inter-frame feature pairs and corresponding languange instructions from spatial-temporal perspectives. We first obtain the caption embeddings f_c with a pre-trained text-encoder F_{text} (ViT-L/14 in [58]) and leverage global average pooling to merge pairwise embeddings f_{t-1}, f_t as follows,

$$f_c = F_{text}(X_c),\ (\Delta f_{t-1}, \Delta f_t) = Pool(f_{t-1}, f_t). \tag{13}$$

We pool f_{t-1} and f_t separately. For the purpose of refining the zero-shot generalization of our method, we add the f_c into inter-frame embeddings f_{t-1}, f_t at the test stage, and then adopt the contrastive loss [36] on both f_c and Δf to achieve the single-level cross-modality alignment:

$$L_{cap} = -\frac{1}{2N_{bs}} \sum_{j=t-1}^{t} \sum_{i=1}^{N_{bs}} \left[\log \frac{\exp(f_c \cdot \Delta f_j^i / \tau)}{\sum_{n=1}^{N_{bs}} \exp(f_c \cdot \Delta f_j^n / \tau)} \right], \tag{14}$$

where N_{bs} is the number of the training batch size and τ is a learnable temperature. With Eq. (13) and Eq. (14), we can compute the multi-level caption aligning loss from video-, scene- to object-level, and the final constrastive loss can be combined as follows,

$$\hat{L}_{cap} = \lambda_{\mathrm{v}} \cdot L_{cap}^{(V)} + \lambda_{\mathrm{s}} \cdot L_{cap}^{(S)} + \lambda_{\mathrm{o}} \cdot L_{cap}^{(O)}. \tag{15}$$

4. Experiments

We present several experiments using model trained only on NOCS-REAL275 dataset to support the claims that our method is able to: (i) cover the novel categories not annotated in training set; (ii) achieve zero-shot inference for unseen instances with unknown (known) classes.

4.1. Experimental Setup

Datasets. To evaluate our L4D-Track, we consider both two public benchmarks of real-world datasets: NOCS-

Table 1. Quantitative comparison of category-level 6-DoF pose estimation on the pubilc NOCS-REAL275 dataset. Note that the best results and second best results are highlighted in **bold** and $\underline{\text{underlined}}$, respectively. The results are averaged over all six categories. The comparison results of current state-of-the-art baselines are all summarized from their original papers, and empty entries either could not be evaluated or were not reported in the original paper.

#	Method	Publication	Input	Outputs	Evaluation Metrics						
					$IoU25\uparrow$	$IoU50\uparrow$	$IoU75\uparrow$	$5°2cm\uparrow$	$5°5cm\uparrow$	$10°2cm\uparrow$	$10°5cm\uparrow$
1	NOCS [45]	CVPR2019	RGB	Pose	82.2	78.0	30.1	7.2	9.5	13.8	25.2
2	CASS [2]	CVPR2020	RGB-D	Shape+Pose	$\underline{84.2}$	77.7	-	-	23.5	-	58.0
3	SDP [41]	ECCV2020	RGB-D	Shape+Pose	83.4	77.3	53.2	19.3	21.4	43.2	54.1
4	SGPA [4]	ICCV2021	RGB-D	Shape+Pose	-	80.1	61.9	35.9	39.6	61.3	70.7
5	FS-Net [5]	CVPR2021	RGB-D	Pose	84.0	81.1	63.5	19.9	33.9	-	69.1
6	GPV-Pose [7]	CVPR2022	D	Pose	$\underline{84.2}$	$\underline{83.0}$	64.4	32.0	42.9	-	73.3
7	RBP-Pose [60]	ECCV2022	RGB-D	Pose	-	-	67.8	38.2	48.1	63.1	79.2
8	TTA-COPE [22]	CVPR2023	RGB-D	Pose	-	69.1	39.7	30.2	35.9	61.7	73.2
9	HS-Pose [62]	CVPR2023	3D Points	Pose	$\underline{84.2}$	82.1	$\underline{74.7}$	$\underline{46.5}$	$\underline{55.2}$	$\underline{68.6}$	$\underline{82.7}$
	Ours w/o seg.	-	3D Points	Shape+Pose	**84.5**	**83.1**	69.2	$\underline{42.3}$	**55.4**	$\underline{64.8}$	**83.1**
	Ours	-	3D Points	Shape+Pose	**86.6**	**83.4**	**76.0**	**47.7**	**56.2**	**68.7**	**85.5**

Table 2. Quantitative comparison of category-level 6-DoF pose tracking on the pubilc NOCS-REAL275 dataset.

Method	Input	Init.	$5°5cm\uparrow$	$IoU25\uparrow$	$R_{err}\downarrow$	$T_{err}\downarrow$
ICP	Depth	GT.	16.9	47.0	48.1	10.5
Oracle ICP [50]	Depth	GT.	0.65	14.7	40.3	7.7
6-PACK [43]	RGB-D	GT.	28.9	55.4	19.3	**3.3**
6-PACK w/o temporal	RGB-D	Pert.	22.1	53.6	19.7	$\underline{3.6}$
CAPTRA [50]	Depth	Pert.	62.2	64.1	$\underline{5.9}$	7.9
CAPTRA +RGB seg.	RGB-D	Pert.	**63.6**	$\underline{69.2}$	6.4	4.2
Mask Fusion [37]	RGB-D	GT.	26.5	64.9	28.5	8.3
Ours	3D points	GT.	$\underline{56.2}$	**86.6**	**5.6**	**3.3**

REAL275 [45] and YCB-Video dataset [52]. The NOCS-REAL275 dataset includes six categories, *i.e., bottle, bowl, camera, laptop, can and mug* and and comprises 7 training videos and 6 testing videos captured in real-world settings. This data contains 8K images that are collected in 18 defferent scenes. The YCB-Video dataset comprises both real-world and synthetic images (21 objects) and the real images include 92 videos captured in various scenes using an RGB-D camera.

Metrics. Following the evaluation metrics in [17], [50], we use five aspects of metrics respectively: 1) 3D-IoU, that measures the average percision for various IoU-overlap thresholds and we use 25%, 50% and 75% for this metric. 2) $a°b\ cm$, measures the average precision of objects for which the error is less than $a°$ for rotation and $b\ cm$ for translation and we adopt $5°2\ cm$, $5°5\ cm$, $10°2\ cm$ and $10°5\ cm$ for this metric. 3) R_{err} and T_{err}, average error of rotation and translation. 4) ADD (S), average distance (synmetric) for instance-level pose estimation. 5) the CD (Chamfer Distance) is used for shape reconstruction.

Implementation Details. Our method is only trained on the NOCS-REAL275 dataset and validated on each test sets. The experiments are conducted on Ubuntu 20.04 system with four NVIDIA RTX A6000 GPUs. We use a batch-size of 32 and trained the network for 30 epochs with early-stopping based on the performance of the model on the held out validation set and set the learning rate to be $1e-7$. We implement the number of hypothesis matrices as 5×10^4. Shape priors are obtained by sampling $N_p = 2048$ points.

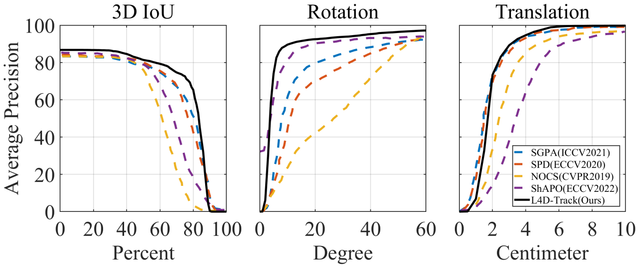


Figure 4. Comparison of mAP on NOCS-REAL275 dataset. Our method and typical baselines for 3D IoU, rotation and translation.

4.2. Comparison with State-of-the-Art Methods

Category-level 6-DoF Pose Tracking. We first conduct category-level 6-DoF pose estimation on the testing set of the NOCS-REAL275 dataset summarized in Tab. 1. To demonstrate the performance evaluation of our L4D-Track, we compared it with the nine main state-of-the-art estimation-based baselines. It's worth noting that most available methods such as GPV-Pose [7] only focus on single pose estimation using RGB-D/RGB whereas we focus on both dynamically tracking and reconstructing objects from 3D video. Notably, we report the results of our method using only point cloud and using both points and segmented mask, our method achieves the better performance compared with previous state-of-the-art baselines. Our approach achieves the most significant gains on metrics of $IoU75$ (+45.9%), $5°2\ cm$ (+40.5%) and $10°5\ cm$ (+60.3%) compared to NOCS, respectively, which indicates the effectiveness of our language-to-4D modeling. Tab. 2 and Fig. 4 summarize the additional comparisons with other related category-level 6-DoF pose tracking methods on the NOCS-REAL275 dataset. To make the comparison fair, we follow the assumption in [43], that the ground-truth values of the pose and size are known in initial frame. It can be concluded that L4D-Track also outperforms most state-of-the-art track-based baselines. Visualization results are shown in the left of Fig. 5.

Table 3. Quantitative comparison of instance-level 6-DoF pose tracking on the pubilc YCB-Video dataset. Note that the best results and second best results are highlighted in **bold** and underlined, respectively.

Object	PoseCNN [52]		DenseFusion [44]		se(3)-TrackNet [48]		PoseRBF [6]		CatTrack [55]		Ours	
	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S
002 master chef can	50.9	84.0	-	**96.4**	93.8	95.9	49.2	62.6	82.5	86.3	**96.2**	**96.4**
003 cracker box	51.7	76.9	-	95.5	**96.4**	**97.1**	74.4	85.2	86.2	91.7	86.3	91.8
004 sugar box	68.6	84.3	-	97.5	**97.6**	**98.1**	74.6	86.1	83.6	92.0	90.6	92.2
005 tomato soup can	66.0	80.9	-	94.6	94.8	97.1	75.0	84.5	84.3	88.6	**95.6**	**97.8**
006 mustard bottle	79.9	90.2	-	97.2	95.7	97.4	75.7	87.3	85.9	90.2	**96.1**	**98.7**
007 tuna fish can	70.4	87.9	-	96.6	86.5	91.1	70.8	86.6	84.7	91.5	**90.4**	**97.1**
008 pudding box	92.9	79.0	-	96.5	**97.9**	**98.4**	62.1	76.6	73.4	85.8	80.3	87.4
009 gelatin box	75.2	87.1	-	98.1	**97.7**	**98.5**	88.3	92.4	90.8	93.9	82.7	84.6
010 potted meat can	59.6	78.5	-	91.3	74.2	82.4	43.7	55.2	66.7	75.9	**87.2**	**92.6**
011 banana	72.3	85.9	-	**96.6**	**84.6**	95.2	40.3	59.7	76.8	82.4	62.5	70.3
019 pitcher base	52.5	76.8	-	97.1	**96.7**	**97.4**	74.9	87.5	84.1	92.8	84.3	89.2
021 bleach cleanser	50.5	71.9	-	95.8	**95.9**	**97.2**	52.7	67.8	73.4	80.5	70.6	75.8
024 bowl	6.5	69.7	-	88.2	39.1	95.6	24.9	87.6	33.6	89.8	**87.4**	**96.2**
025 mug	57.7	78.0	-	97.1	91.6	**96.9**	64.4	82.1	72.1	83.9	**91.8**	**96.9**
035 power drill	55.1	72.8	-	96.0	**96.4**	**97.4**	60.0	77.1	71.3	86.0	70.1	73.3
036 wood block	31.8	65.8	-	89.7	33.9	**95.9**	7.7	18.4	28.6	62.3	**50.2**	64.3
037 scissors	35.8	56.2	-	95.2	**95.7**	**97.5**	28.5	43.7	64.9	74.3	70.3	77.8
040 large marker	58.0	71.4	-	97.5	**89.0**	**94.2**	51.3	60.1	70.8	83.4	72.3	79.2
051 large clamp	25.0	49.9	-	72.9	71.6	**96.9**	55.6	73.7	66.8	78.1	**80.9**	89.1
052 extra large clamp	15.8	47.0	-	69.8	**64.6**	**95.8**	51.2	71.4	49.8	77.2	64.3	68.7
061 foam brick	40.4	87.8	-	92.5	40.7	**94.7**	77.7	88.9	86.0	93.4	**78.6**	89.4
Average	53.7	75.9	-	93.1	**87.8**	**95.5**	60.4	75.4	72.2	84.8	80.4	86.1

Table 4. Quantitative comparison of 3D shape reconstruction on the pubilc NOCS-REAL275 dataset: Evaluated with CD (10^{-2}).

Method	Bottle	Bowl	Camera	Can	Laptop	Mug	Average
6D-ViT [64]	0.24	0.11	0.61	0.16	0.14	0.11	0.21
CASS [2]	0.17	0.09	0.53	0.18	0.19	0.24	0.23
C3R-Net [46]	0.30	0.10	0.76	0.13	0.13	0.12	0.26
ShAPO [18]	0.10	**0.08**	0.40	**0.07**	0.08	**0.06**	0.13
SPD [41]	0.34	0.12	0.89	0.15	0.29	0.10	0.32
SGPA [4]	0.29	0.09	0.55	0.17	0.16	0.11	0.23
CenterSnap [17]	0.13	0.10	0.43	0.09	**0.07**	**0.06**	0.15
GCASP [23]	0.21	0.16	0.11	0.16	0.21	0.29	0.19
Ours w/o seg.	0.10	0.12	0.15	0.14	0.09	0.10	0.13
Ours	**0.09**	**0.08**	**0.10**	**0.07**	**0.07**	**0.06**	**0.08**

Table 5. Comparison of tracking speed (FPS) for typical methods on the both NOCS-REAL275 and YCB-Video datasets.

Dataset	NOCS [45]	6-PACK [43]	SGPA [4]	CAPTRA [50]	Ours
NOCS-REAL275	5.24	4.03	14.12	10.35	**20.45**
YCB-Video	6.39	5.01	16.52	12.44	**19.28**

Instance-level 6-DoF Pose Tracking. Our method has already shown the excellent potential ability in solving the category-level pose estimation task for unseen objects with unknown class shifts. However, it's limited to certain six categories and the transferable zero-shot learners across different categories and datasets also merit exploration. In this regard, we then conduct zero-shot instance-level 6-DoF pose estimation experiments that only train the model on NOCS-REAL275's base classes and test it on the YCB-Video's classes without fine-tuning. As depicted in Tab. 3, it can be observed that our L4D-Track performs well in generalizing to unknown categories and can accurately etimate the pose of the unseen objects with seen classes. Note that we use the model without segmention head and 2D backbone during the inference stage. Specifically, our method consistently outperforms CatTrack [55], the latest instance-level method, by 20.0% ∼ 30.0% for the metric of ADD and 5.0% ∼ 15.0% for ADD-S in response to unseen objects with the seen category before: "002 master chef can", "005 tomato soup can", "006 mustard bottle" and so on (marked in gray in Tab. 3), respectively. It further illustrates the zero-shot capability of our approach. Since our model was trained with 6 specific categories on NOCS-REAL275, its performance to other categories is weaker than the current best method se(3)-TrackNet [48].

3D Shape Reconstruction. We further test our approach's ability of reconstructing complete 3D shapes by comparing against RGB-D based and depth-based state-of-the-art baselines on the NOCS-REAL275 dataset, *i.e.,* 6D-ViT [64], SPD [41], SGPA [4], CenterSnap [17] and so on. The results are reported in Tab. 4, it is observed that the 3D shapes reconstructed by our complete model and it variant (without segmented head) obtain average CD metrics of 0.13 and 0.08, respectively, compared to the 0.13 of the current best baseline ShAPO [18], meanwhile our consistently lower CD compared to other baselines. It indicates that the pairwise implicit 3D field representation proposed in our network can greatly improve the quality of 3D model reconstruction. We also show a qualitative comparison in the Fig. 5. For more comparative results in the YCB-Video dataset, please refer to the appendix.

Runtime Analysis. We analyze the inference time of pose tracking performance, as summarized in Tab. 5. Our L4D-Track achieves an average speed of approximately 20 FPS on the NOCS-REAL275 dataset and 19 FPS on the YCB-Video dataset, respectively. For a fair evaluation, we compared the runtime with four available methods using their officially released codes and all method are tested on the same device, *i.e.,* an NVIDIA RTX A6000 GPU.

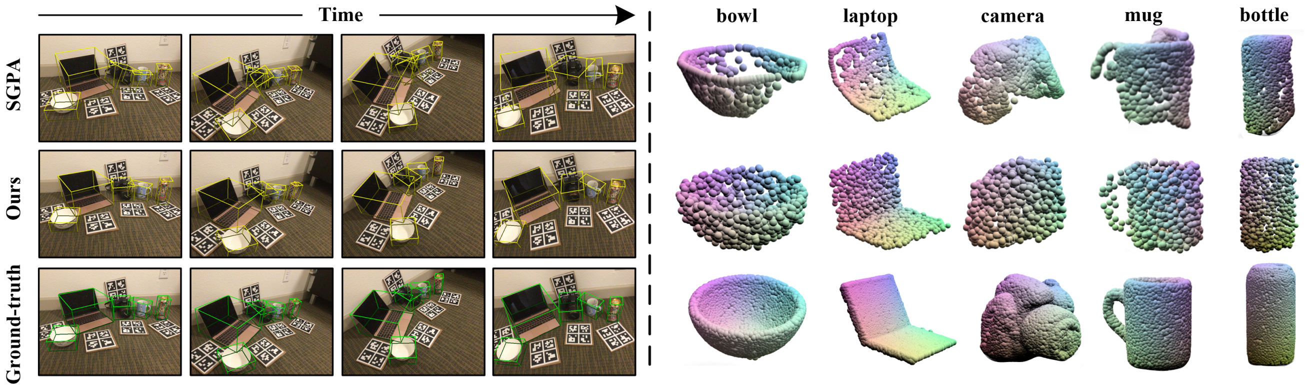


Figure 5. Visual comparison with competitive baseline on NOCA-REAL275 dataset. **Left**: the qualitative comparison of 6-DoF pose tracking. For a clearer comparison, we have expressed the results in RGB images. Yellow and green represent the results from SGPA [4], ours and ground-truth label, respectively. **Right**: the output of final reconstructed 3D shape, that are rendered from the same viewpoint.

Table 6. Ablation studies on different configurations of network structure and loss terms on both two public datasets. CCF refers to Cross-coupled Fusion. EPH refers to Energy-based Pose Hypothesis. Without this module, we recover pose with a simple MLP. NPFs refers to Neural Pose-aligned Fields. Without this module, we directly reconstruct the 3D shape. The loss L_{base} contains L_{pose}, L_{shape} and L_s.

Method	Network			Loss Terms				NOCS-REAL275						YCB-Video	
	CCF	EPH	NPFs	L_{base}	$L_{cap}^{(V)}$	$L_{cap}^{(S)}$	$L_{cap}^{(O)}$	$IoU25 \uparrow$	$5°5cm \uparrow$	$10°5cm \uparrow$	$R_{err} \downarrow$	$T_{err} \downarrow$	CD ↓	ADD ↑	ADD-S ↑
①	✗	✗	✗	✓	✓	✓	✓	66.3	32.1	48.5	24.3	17.9	0.45	58.6	62.3
②	✗	✓	✓	✓	✓	✓	✓	86.0	54.3	84.9	5.8	4.0	0.10	75.0	84.9
③	✓	✗	✓	✓	✓	✓	✓	68.4	39.2	55.1	12.3	10.8	0.08	64.5	70.3
④	✓	✓	✗	✓	✓	✓	✓	86.6	56.2	85.5	5.6	3.3	0.36	76.0	85.9
⑤	✓	✓	✓	✓	✓	✗	✗	86.2	55.8	84.2	4.1	4.3	0.11	72.6	84.0
⑥	✓	✓	✓	✓	✗	✓	✗	85.7	55.4	84.1	6.6	4.5	0.11	74.3	85.0
⑦	✓	✓	✓	✓	✗	✗	✓	86.0	54.3	83.2	5.9	4.9	0.12	75.0	84.6
Ours	✔	✔	✔	✔	✔	✔	✔	**86.6**	**56.2**	**85.5**	**5.6**	**3.3**	**0.08**	**80.4**	**86.1**

4.3. Ablation Study

Effect of the pairwise implicit representation. As depicted in Tab. 6, we evaluate the performance of our method's variants, *w.r.t.,* different configurations of network architectures and the choice of loss terms. We examine key components of our model through in-depth ablation studies, as shown from #① to #④ in Tab. 6. We investigate the effectiivenesss of proposed fusion module and two core network module, and the overall performance shows a significant leap forward from 0.08 to 0.36 referring to CD metric, idicating designed NPFs is needed to enable shape reconstruction performance. Our model without pose hypothesis under-performs the complete one also indicates that implict space leads to more robust pose.

Impact of the multi-level caption supervision. We also reduce each constrastive loss from video-level to object-level in order (see from #⑤ to #⑦), as can be served that all ablated versions exhibit their poorer performance on both two datasets when corresponding loss were removed. It not only indicates the effectiveness of the supervision from the multi-level captions but also indirectly shows that our designed language-to-4D association module improved the performance of our complete model (*Please check out the appendix for more ablation analyses*).

5. Conclusion

In this work, we present a zero-shot language-to-4D modeling framework to learn universal representation that can achieve spatio-temporal model across the entire raw point cloud video by aligning the embedded pairwise features and language features from multi-level contextual perspectives. Our method jointly trains a pairise implict 3D space representation and a pre-training language association with multi-level instructions to achieve simultaneously 6-DoF pose tracking and 3D shape reconstruction for unseen objects in a 3D video. Extensive experiments demonstrate the effectiveness of our L4D-Track on modeling between point cloud video and language, and the zero-shot inference for pose estimation and shape reconstruction. With the development of large language model, we will explore the deeper potiential of our approach in the future research.

Acknowledgement: This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. Prof. Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (No.FT210100268) funded by the Australian Government.

References

[1] Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117, 2022. 1

[2] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 6, 7

[3] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2781–2790, 2022. 2

[4] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 6, 7, 8

[5] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 6

[6] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021. 7

[7] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 2, 6

[8] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2023. 2

[9] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 2

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 2

[11] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[12] Yang Hai, Rui Song, Jiaojiao Li, and Yinlin Hu. Shape-constraint recurrent flow for 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4831–4840, 2023. 2

[13] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *European Conference on Computer Vision*, pages 585–603. Springer, 2022. 2, 4

[14] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[15] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2021. 2

[16] Tianrui Hui, Si Liu, Zihan Ding, Shaofei Huang, Guanbin Li, Wenguan Wang, Luoqi Liu, and Jizhong Han. Language-aware spatial-temporal collaboration for referring video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[17] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640. IEEE, 2022. 2, 4, 6, 7

[18] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 7

[19] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 2

[20] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 2

[21] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. Uda-cope: unsupervised domain adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14891–14900, 2022. 2

[22] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. Tta-cope: Test-time adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21285–21295, 2023. 2, 6

[23] Guanglin Li, Yifeng Li, Zhichao Ye, Qihang Zhang, Tao Kong, Zhaopeng Cui, and Guofeng Zhang. Generative

category-level shape and pose estimation with semantic primitives. In *Conference on Robot Learning*, pages 1390–1400. PMLR, 2023. 7

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[26] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in Neural Information Processing Systems*, 34:7838–7851, 2021. 2

[27] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*, pages 112–123. PMLR, 2022. 2

[28] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4

[30] 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. 1

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[32] Kieran Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*, 2021. 3

[33] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2, 4

[34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[36] 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, 5

[37] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, 2018. 6

[38] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17929–17938, 2022. 2

[39] Jingtao Sun, Yaonan Wang, Mingtao Feng, Danwei Wang, Jiawen Zhao, Cyrill Stachniss, and Xieyuanli Chen. Ick-track: A category-level 6-dof pose tracker using inter-frame consistent keypoints for aerial manipulation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1556–1563. IEEE, 2022. 2

[40] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[41] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 6, 7

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[43] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020. 2, 6, 7

[44] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 7

[45] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 2642–2651, 2019. 2, 6, 7

[46] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 7

[47] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021. 2

[48] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373. IEEE, 2020. 7

[49] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 2

[50] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13209–13218, 2021. 2, 6, 7

[51] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2

[52] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 6, 7

[53] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023. 1, 2

[54] Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6875, 2022. 2

[55] Sheng Yu, Di-Hua Zhai, Yuanqing Xia, Dong Li, and Shiqi Zhao. Cattrack: Single-stage category-level 6d object pose tracking via convolution and vision transformer. *IEEE Transactions on Multimedia*, 2023. 2, 7

[56] 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. 1

[57] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[58] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023. 2, 5

[59] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2

[60] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022. 2, 6

[61] 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. 1, 2

[62] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17163–17173, 2023. 2, 6

[63] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2

[64] Lu Zou, Zhangjin Huang, Naijie Gu, and Guoping Wang. 6d-vit: Category-level 6d object pose estimation via transformer-based instance representation learning. *IEEE Transactions on Image Processing*, 31:6907–6921, 2022. 7