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

Outline

In this supplementary material, we will provide more experimental results and the details that are not elaborated on the main manuscript due to the length limitation of page:

- Sec. A: We present the implementation details of our L4D-Track, that comprises extra metrics, core network design and limitation analysis.
- Sec. B: We show the additional experiment results on both NOCS-REAL275 and YCB-Video datasets, and the qualitative results of more ablation study.

A. Implementation Details

A.1. Metrics for Instance-Level Pose Tracking

As mentioned in Sec. 4.1 in our main manuscript, we use the extra evaluation metrics *i.e., ADD and ADD-S* for the comparison of instance-level 6-DoF pose estimation, which have been extensively utilized in prior works [6, 8].

1) **ADD**: ADD measures the distance between the ground truth 3D model and corresponding posed points using our predictions. The prediction is considered correct if this distance is within a certain threshold. The calculation process is defined as:

$$ADD = \frac{1}{m}\sum_{p\in M} ||(\tilde{R}p + \tilde{t}) - (Rp + t)||, \qquad (1)$$

where M is the set of points in the 3D model, m is the number of points p, $\tilde{R}$ and $\tilde{t}$ are the ground truth rotation and translation, R and t are the predicted values.

2) **ADD-S**: For the symmetrical objects, such as bowl and can, the average distance needs to be adapted for multiple appropriate poses due to symmetry axes. In this regard, the metric of ADD-S can be defined as follows:

$$ADD-S = \frac{1}{m}\sum_{p_1\in M} \min_{p_2\in M} ||(\tilde{R}p_2 + \tilde{t}) - (Rp_1 + t)||, \qquad (2)$$

where p_1 and p_2 are the points on the 3D model.

A.2. Model Architecture

As to 2D and 3D backbone, we use CNN-based encoder-decoder framework and the variant of PointNet++ [4], respectively. Their detailed architecture is as follows:
1) **2D-Backbone**:

$$\begin{aligned}&\mathrm{Res}net(block, layer = [2, 2, 2, 2]) \rightarrow\\ &PSPModule([512, \mathrm{bins} = (1, 2, 3, 6)]) \rightarrow Dropout(0.15)\\ &PSPUpsample([1024, 256, 64, 64, 32]) \rightarrow\\ &FP(mlp = [32, 512])\end{aligned}$$

2) **3D-Backbone**:

$$\begin{aligned}&SA(npoints = 2048, redius = 0.2, mlp([64, 64, 128])) \rightarrow\\ &SA(npoints = 2048, redius = 0.4, mlp([128, 128, 256])) \rightarrow\\ &SA(npoints = 2048, redius = 0.4, mlp([256, 256, 512])) \rightarrow\\ &BatchNorm([512]) \rightarrow Dropout(0.4)\end{aligned}$$

We use LeakyReLU for each layer in set abstraction (SA), feature propagation (FP) and batch normalization.

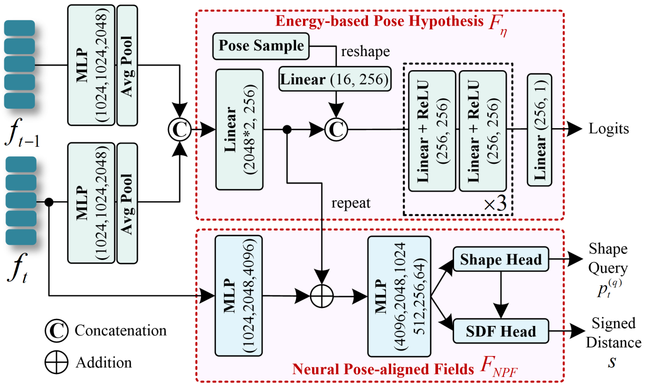

Figure 1. The detailed structure of pairwise implicit 3D space representation (*i.e.,* Sec. 3.2 in our manuscript).

As to the Pairwise Implicit 3D Shape Representation in our core design, mentioned in Sec. 3.2 of the main manuscript, we exploit a network F_η and a full-connnected neural network F_{NPF} to achieve pose hypothesis and shape query reconstruction, respectively. The concrete pipline can be found in Fig. 1.

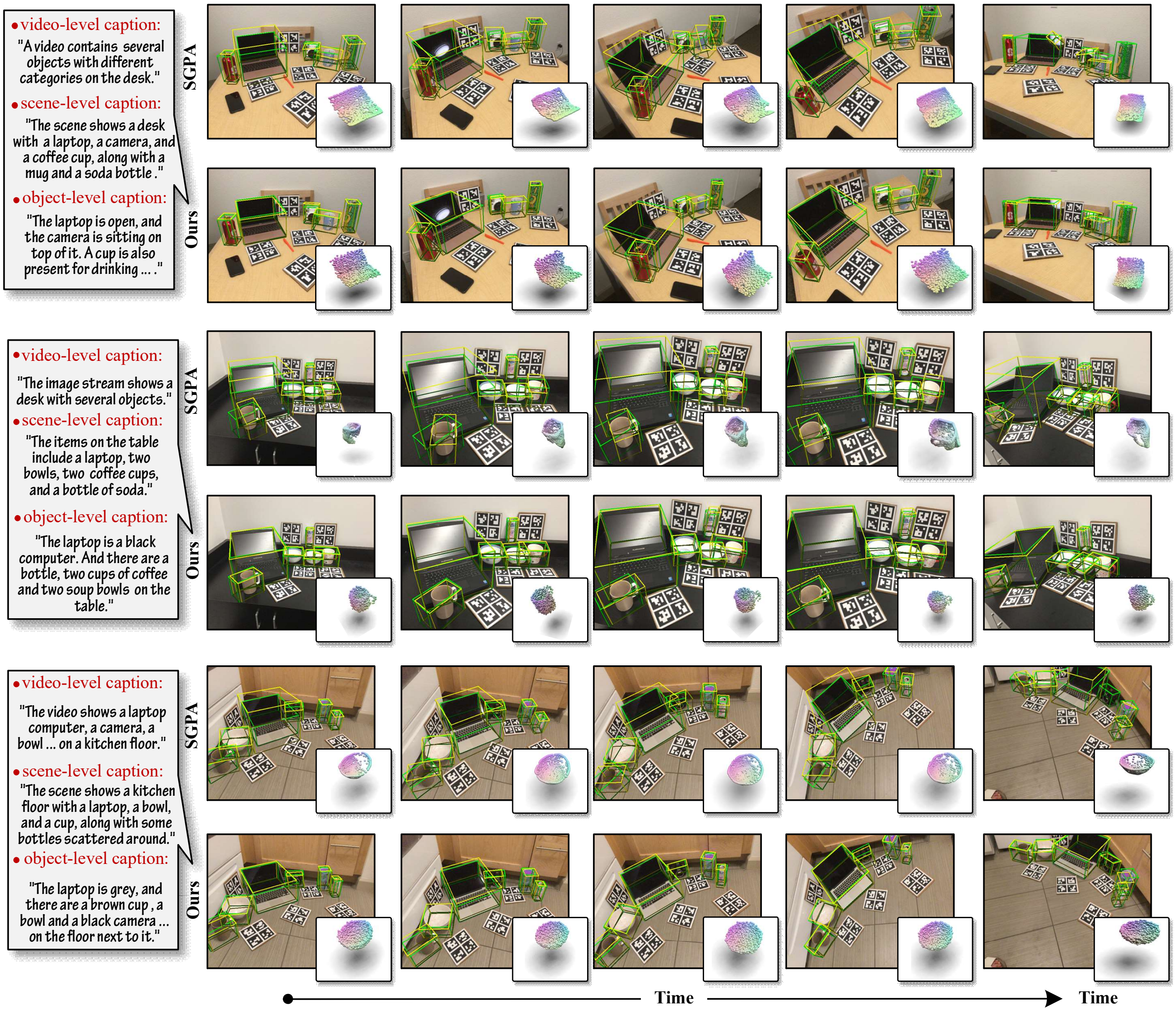
• video-level caption:
"A video contains several objects with different categories on the desk."
• scene-level caption:
"The scene shows a desk with a laptop, a camera, and a coffee cup, along with a mug and a soda bottle."
• object-level caption:
"The laptop is open, and the camera is sitting on top of it. A cup is also present for drinking ... ."
SGPA
Ours
• video-level caption:
"The image stream shows a desk with several objects."
• scene-level caption:
"The items on the table include a laptop, two bowls, two coffee cups, and a bottle of soda."
• object-level caption:
"The laptop is a black computer. And there are a bottle, two cups of coffee and two soup bowls on the table."
SGPA
Ours
• video-level caption:
"The video shows a laptop computer, a camera, a bowl ... on a kitchen floor."
• scene-level caption:
"The scene shows a kitchen floor with a laptop, a bowl, and a cup, along with some bottles scattered around."
• object-level caption:
"The laptop is grey, and there are a brown cup, a bowl and a black camera ... on the floor next to it."
SGPA
Ours
Time
Time

Figure 2. More visualization results on the NOCS-REAL275 dataset. We compare our method with the representative baseline (SGPA [1]). **Left:** The corresponding language caption generated from our method. **Right:** 6-DoF pose tracking and reconstructed shape visualization, that are presented in RGB image for clearer comparison. Yellow and green represent the results from SGPA, ours and ground-truth label.

A.3. Limitation and Open Problems

While our L4D-Track framework effectively handles zero-shot 6-DoF tracking and 3D shape reconstructing by incorporating the language semantics and pairwise implict representation, it still faces limitations in certain aspects. The key limitation is related to the available ground-truth labels for point cloud-language data pairs. Although our proposed association method provides the accurate description of the infromation in 3D point cloud, the performance of our method is still limited. We believe that pre-training our network on a large dataset with rich 3D-language semantic information will be a promising alternative, which will be explored in our future works. Additionally, the model tends to generate the non-perfect shapes based on the current observation. This motivates us to explore shape completion module in the following research.

B. Additional Results

B.1. Extra Qualitative Results

Tab. 1 summarizes the detailed per-category results of our approach for object pose tracking on the NOCS-REAL275 dataset. We also show more detailed quantitative comparisons of 3D shape reconstruction on YCB-Video as

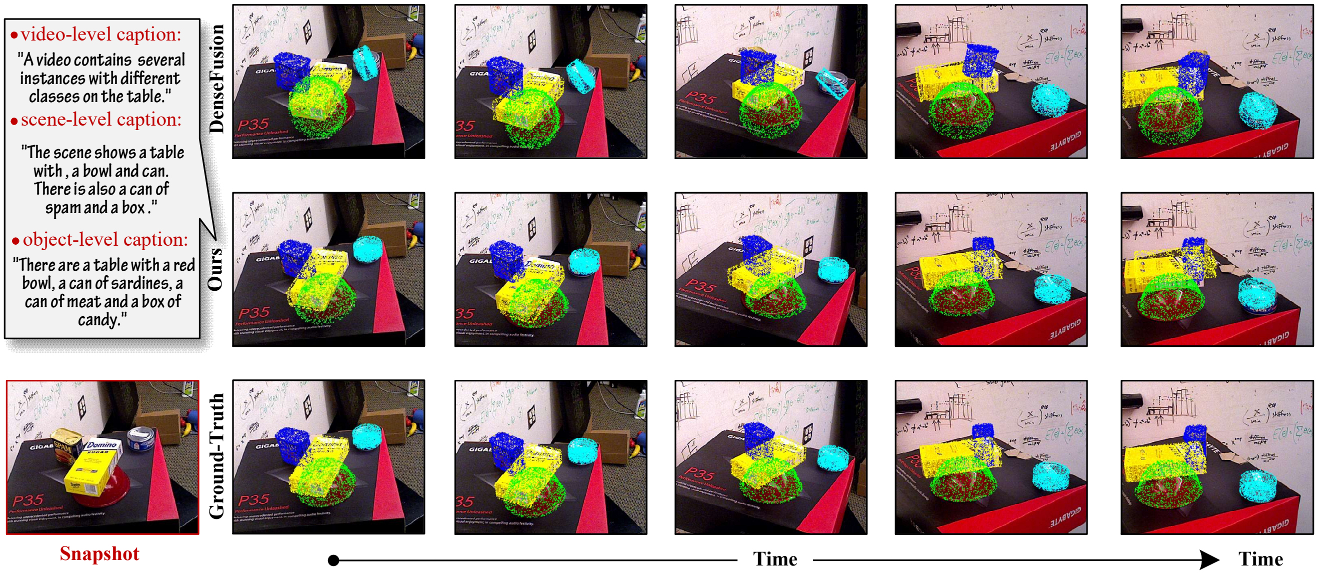
• video-level caption:
"A video contains several instances with different classes on the table."
• scene-level caption:
"The scene shows a table with , a bowl and can. There is also a can of spam and a box ."
• object-level caption:
"There are a table with a red bowl, a can of sardines, a can of meat and a box of candy."
DenseFusion
Ours
Ground-Truth
Snapshot
Time
Time

Figure 3. Visualization results of 6-DoF pose tracking on the YCB-Video dataset. We compare our method with the representative baseline (DenseFusion [6]), and the corresponding language captions are generated from our method (left side). To keep in line with DenseFusion, we also project each object shape model to 2D image frame using different colors.

Table 1. Per-category results of our proposed method on the NOCS-REAL275 dataset under different evaluation metrics.

Category	$IoU25$	$IoU50$	$IoU75$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
Bottle	88.3	85.2	76.4	50.9	52.3	68.5	84.7
Bowl	89.2	86.0	76.2	54.7	55.2	66.8	86.9
Camera	83.6	80.1	77.9	43.1	55.7	67.4	88.4
Can	87.5	84.0	75.6	47.3	58.6	70.2	85.0
Laptop	81.6	80.0	74.2	42.4	58.9	70.1	83.8
Mug	88.2	85.1	75.7	47.8	56.5	69.0	84.2
Average	**86.6**	**83.4**	**76.0**	**47.7**	**56.2**	**68.7**	**85.5**

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

Object	SGPA [1]	SPD [5]	CenterSnap [2]	Ours w/o seg.	Ours
002 master chef can	0.16	0.15	0.18	0.16	**0.14**
003 cracker box	0.23	**0.16**	0.17	0.18	**0.16**
006 mustard bottle	0.31	0.55	**0.14**	0.21	0.20
024 bowl	0.10	0.13	0.13	0.11	**0.09**
025 mug	0.14	0.12	**0.07**	0.14	0.11
Average	0.19	0.22	**0.14**	0.16	**0.14**

Table 3. Ablation analysis on the robustness to tracking pose errors. Init. $\times m$ means adding m times train-time errors in pose initialization. All $\times m$ means adding m times errors to all estimated poses in every previous frames.

Dataset	Metric	Orig.	Init.× 1	Init.× 2	All× 1	All× 2
NOCS-REAL275	$5°5cm$ ↑	**56.2**	55.2	53.9	55.0	53.3
	$IoU25$ ↑	**86.6**	85.0	83.8	85.7	82.1
	R_{err} ↓	**5.6**	6.6	7.4	5.8	6.8
	T_{err} ↓	**3.3**	5.2	6.3	4.0	5.5
YCB-Video	ADD ↑	**80.4**	74.1	73.3	75.0	74.2
	ADD-S ↑	**86.1**	83.0	82.4	84.3	82.2

Table 4. Ablation analysis on influence of different configurations for inter-frame embeddings. "non-2D features" means our model without adding image features and "Fusion" indicates cross-attention based cross-coupled fusion module. "Caption" means the language caption embeddings.

Dataset	Metric	Inter-Frame Embeddings (f_{t-1}/f_t)			
		Orig.	non-2D features	w/o Fusion	w/o Caption (f_c)
NOCS-REAL275	$5°5cm$ ↑	**56.2**	55.4	55.0	50.9
	$IoU25$ ↑	**86.6**	84.5	82.2	81.7
	R_{err} ↓	**5.6**	7.8	6.3	9.2
	T_{err} ↓	**3.3**	4.3	5.0	6.8
	CD ↓	**0.08**	0.14	0.10	0.16
YCB-Video	ADD ↑	**80.4**	73.9	75.0	68.1
	ADD-S ↑	**86.1**	80.1	84.3	77.2
	CD ↓	**0.14**	0.18	0.14	0.20

depicted in Tab. 2. We compare with three representative baselines including SGPA [1], SPD [5], CenterSnap [2] and we selected five categories including "002 master chef can", "003 cracker box", "006 mustard bottle", "024 bowl", "025 mug". As one can easily deduce in Tab. 2, we obtain average CD metrics of 0.16 and 0.14 using our complete model and its variant, respectively, and we outperforms the state-of-the-arts also indicates the superiority for zero-shot shape reconstruction.

Table 5. Ablation analysis on the effect of the number of pose hypothesis matrices. The original seting (Orig.) is 5×10^4.

Dataset	Metric	Orig.	1×10^5	2×10^4	5×10^3	2×10^3
NOCS-REAL275	$5°5cm \uparrow$	**56.2**	54.3	53.7	35.4	42.1
	$IoU25 \uparrow$	**86.6**	84.0	85.2	65.7	58.9
	$R_{err} \downarrow$	**5.6**	5.8	6.9	14.8	16.9
	$T_{err} \downarrow$	**3.3**	4.0	5.6	20.3	27.3
YCB-Video	ADD ↑	**80.4**	79.3	75.1	66.1	58.1
	ADD-S ↑	**86.1**	85.8	79.8	68.4	60.2

B.2. Extra Ablation Analyses

Robustness to pose noises. Due to our method needs to be based on the pose from the previous frame or initial pose, we further ablate the robustness of our method against different noise pose inputs on pose accuracy. As described in Tab. 3, we gradually increase the initial pose error from one to two time to examine the robustness to pose initialization, meanwhile, we add one or two times pose error to every previous frames to examine the robustness to tracking pose errors. It can be seen that our method is robust to the noise of the previous or initial poses.

Influence of inter-frame embeddings. To investigate the influence of different configurations of the inter-frame embeddings in our framework, we conduct experiments using three variations of our model: without 2D image features, *i.e., non 2D features*, our model without cross-coupled fusion and our model without adding languange caption embeddings during testing. As shown in Tab. 4, the results demonstrate that the caption embeddings f_c and our proposed cross-coupled fusion module can possess the stronger pairwise representation ability.

Effect of pose hypothesis matrices. In L4D-Track, we use the Assumption to estimate the change in pose to obtain the optimal pose results. So, we also investigate the effect of different choices of the number of hypothesis matrices on pose accuracy. we gradually reduce the number from 1×10^5 to 2×10^3. Tab. 5 concludes the comparative results on botn two public datasets. It shows that the pose result is relatively stable to the original choice and the smaller the number, the greater the impact on pose accuracy when the number is less than 2×10^4.

B.3. Additional Visualizations

We provide extra qualitative visualization results of 6-DoF pose tracking and 3D shape reconstructing on NOCS-REAL275 dataset, as depicted in Fig. 2. Compared to the SGPA baseline that adjusts the prior feature by injecting instance information into the prior feature, our approach can estimate more accurate pose and generate more complete shape with the help of the language captions. It reflects the strong generalized estimation ability of our method.

We also visualize some comparative results made by DenseFusion [6] with iterative refinement (two iterations) and our model. As seen in Fig. 3, the baseline fails to tracking the pose of the bowl and the tuna fish can, whereas our approach remains more robust performance. It also indicates that our method performs well even when the targeted objects are heavily occluded.

References

[1] 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. 2, 3

[2] 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. 3

[3] 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.

[4] 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. 1

[5] 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. 3

[6] 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. 1, 3, 4

[7] 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.

[8] 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. 1