2024「中技社科技獎學金」
2024 CTCI Foundation Science and Technology Scholarship
境外生研究獎學金
Research Scholarship for International Graduate Students

Segmenting 3D Shape Parts via Text by 2D Vision-Language Model Distillation

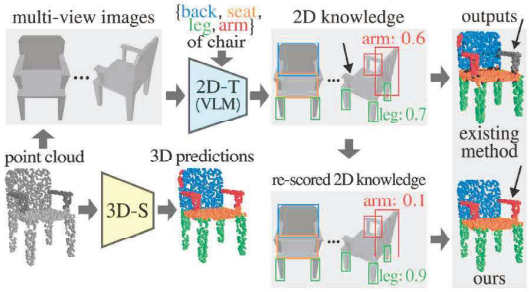
國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

PhD student: Ardian Umam, **Advisors:** Prof. Jen-Hui Chuang & Prof. Yen-Yu Lin
Departement of Electrical Engineering and Computer Science, National Yang Ming Chiao Tung University

Abstract

We propose a framework that transfers 2D knowledge from vision-language models (VLMs) for 3D shape part segmentation. It addresses three challenges: lack of 3D segmentation in invisible regions, inconsistent 2D VLM predictions, and lack of knowledge accumulation. The framework adopts a teacher-student model, where a VLM and a 3D point cloud backbone are employed as the teacher and student networks, respectively. Bidirectional distillation is carried to refine 2D predictions, enhancing 3D segmentation.

Motivations



➤ Goal: Developing a zero-shot/few-shot 3D part segmentation method by Vision-Language Model (VLM) distillation, which resolves 3 major issues:

- ➥ $\mathcal{I}_1$: the lack of 3D segmentation in invisible or undetected regions in the 2D projections
- ➥ $\mathcal{I}_2$: inconsistent 2D predictions by VLMs
- ➥ $\mathcal{I}_3$: the lack of knowledge accumulation across shapes

Contributions

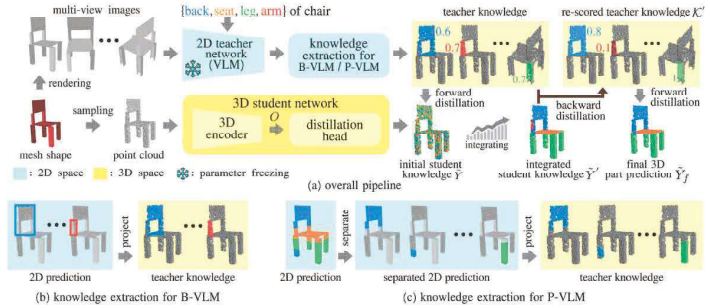
- ➥ Introduce a cross-modal distillation framework, from 2D VLMs to 3D part segmentation model which alleviates the 3 major issues and generalizes to both Bounding-box VLM (B-VLM) and Pixel-wise VLM (P-VLM)
- ➥ Propose a bi-directional distillation which enhances the 2D knowledge sources and subsequently improving the 3D predictions
- ➥ Our method can leverage existing generative models to enrich knowledge source for distillation

Publications of This and Our Previous Works

- ➥ Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. "PartDistill: 3D Shape Part Segmentation by Vision-Language Model Distillation". CVPR. 2024
- ➥ Ardian Umam, Cheng-Kun Yang, Jen-Hui Chuang, and Yen-Yu Lin. "Unsupervised Point Cloud Co-part Segmentation via Co-attended Superpoint Generation and Aggregation". IEEE TMM. 2024.
- ➥ Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. "Point MixSwap: Attentional point cloud mixing via swapping matched structural divisions". ECCV. 2022

Proposed Method

➤ Framework



➤ Forward distillation

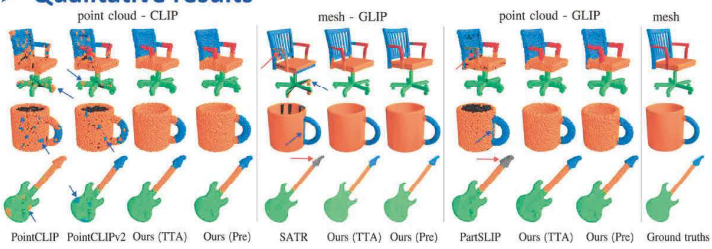
- ➥ Multi-view rendering examples
- ➥ Simultaneously learn incomplete 2D knowledges and 3D features ($\mathcal{I}_1$ ✓)
- ➥ Accumulate knowledges from collection of shapes ($\mathcal{I}_3$ ✓)

➤ Backward distillation

- ➥ Not all knowledges are reliable
- ➥ Refine the knowledges by assigning higher scores to those of high quality and suppressing the low-quality ones ($\mathcal{I}_2$ ✓)

Experimental Results

➤ Qualitative results



➤ Quantitative results

➥ Zero-shot segmentation (% mIoU) on the ShapeNetPart dataset

VLM	Data type	Method	Airplane	Bag	Cap	Chair	Earphone	Overall
CLIP	point cloud	PointCLIP	22.0	44.8	13.4	18.7	28.3	31.0
		PointCLIPv2	35.7	53.3	53.1	51.9	48.1	48.4
		OpenScene	34.4	63.8	56.1	59.8	62.6	52.9
		Ours (TTA)	**37.5**	62.6	55.5	56.4	55.6	**53.8**
		Ours (Pre)	**40.6**	**75.6**	**67.2**	**65.0**	**66.3**	**63.9**
GLIP	point cloud	Ours (TTA)	57.3	62.7	56.2	74.2	45.8	54.7
		Ours (Pre)	69.3	70.1	67.9	86.5	51.2	64.1
	mesh	SATR [1]	32.2	32.1	21.8	25.2	19.4	32.3
		Ours (TTA)	**53.2**	**61.8**	**44.9**	**66.4**	**43.0**	**49.5**
		Ours (Pre)	**64.8**	**64.4**	**51.0**	**67.4**	**48.3**	**56.3**

➥ Leveraging generated data

Distilled data	ShapeNetPart			COSEG	
	Airplane	Chair	Guitar	Chair	Guitar
Train-set (baseline)	69.3	86.2	76.8	96.4	68.0
Gen. data	69.0	85.3	75.6	96.1	67.5
Gen. data & train-set	70.8	88.4	78.3	97.4	70.2

Scan for the code!

財團法人中技社 CTCI FOUNDATION