Computer Vision and Image Understanding 274 (2027) 104990

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

PartDistill++: Distilling multiple foundation models for robust 3D shape part segmentation

Ardian Umam [a,b], Cheng-Kun Yang [c], Min-Hung Chen [d], Jen-Hui Chuang [a,e], Yen-Yu Lin [a,*]

[a] National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[b] Qualcomm, Taipei, Taiwan
[c] MediaTek, Hsinchu, Taiwan
[d] NVIDIA, Taipei, Taiwan
[e] Jimei University, Xiamen, China

ARTICLE INFO

Communicated by Nicu Sebe

Dataset link: https://github.com/ardianumam/PartDistill-plusplus

Keywords:
Point cloud segmentation
Shape part segmentation
Distillation
Multi-modal segmentation

ABSTRACT

Inspired by the success of vision-language models (VLMs) in 2D vision tasks, researchers have extended their application to 3D understanding, including 3D shape part segmentation as addressed in this work. Beyond the challenges of cross-modal transfer from 2D to 3D, the quality of 3D segmentation outcomes depends on the reliability of the VLM predictions. While a VLM can produce satisfactory predictions for certain shape categories, it may struggle with others, and different VLMs may exhibit distinct strengths and weaknesses. Additionally, semantically agnostic vision models (VMs) can generate more precise 2D segmentation boundaries and could be beneficial if integrated properly. The key challenge lies in effectively propagating semantic labels to these semantically agnostic segments and mitigating negative transfer when individual segments encompass multiple semantic categories. In this paper, we introduce PartDistill++, an enhanced cross-modal distillation framework that leverages multiple foundation models (FMs), including semantically aware VLMs and semantically agnostic VMs, to harness their combined strengths while mitigating individual weaknesses. Our proposed Knowledge Enhancer Module enables the distillation of these multiple FMs by purifying noisy knowledge from individual FMs and propagating semantic labels to semantically agnostic VM segments, ensuring that only beneficial knowledge is employed in the distillation process. Extensive experiments demonstrate that PartDistill++ significantly outperforms existing methods, achieving new state-of-the-art results on both the ShapeNetPart and PartNetE datasets, in the absence of 3D ground-truth annotations. The code for this work is available at https://github.com/ardianumam/PartDistill-plusplus.

1. Introduction

Recent advances in foundation models (FMs) (Bommasani et al., 2021), particularly vision-language models (VLMs) (Zhang et al., 2024), have profoundly reshaped the field of computer vision. These models facilitate various downstream tasks, such as image enhancement (Wei et al., 2025), anomaly detection (Zanella et al., 2024), domain adaptation (Yu et al., 2025), action recognition (Osman et al., 2024), referring expression (Ke et al., 2023; Ding et al., 2022), and image generation (Cheng et al., 2022). The success of these applications is attributed to VLMs' extensive knowledge of visual-linguistic relationships, which is acquired by training on large-scale image–text pairs readily available on the internet and can compensate for the lack of ground-truth annotations in downstream tasks.

Inspired by the success of VLMs in 2D vision, researchers have extended their application to 3D understanding tasks, especially those where 3D training data annotations are sparse or absent. Notably, studies in Zhang et al. (2022b), Zhu et al. (2023), M. Liu et al. (2023), Abdelreheem et al. (2023), Mei et al. (2024) utilize VLMs for 3D shape part segmentation, which is crucial for various 3D vision applications, such as robotic manipulation (Aleotti and Caselli, 2012; Ückermann et al., 2013; Aleotti and Caselli, 2011), shape stylization (Michel et al., 2022), and augmented/virtual reality (AR/VR) (Xu et al., 2022). A common paradigm projects a 3D shape into multi-view 2D images. A VLM is applied to these images to obtain 2D predictions. These multi-view predictions are then transferred to 3D space to vote on the final 3D segmentation (Zhang et al., 2022b; Zhu et al., 2023; M. Liu et al., 2023; Abdelreheem et al., 2023).

However, directly transferring 2D VLM predictions to 3D segmentation, as done in Zhang et al. (2022b), Zhu et al. (2023), M. Liu et al. (2023), Abdelreheem et al. (2023), presents three primary challenges: (

* Corresponding author.
E-mail address: lin@cs.nycu.edu.tw (Y.-Y. Lin).

https://doi.org/10.1016/j.cviu.2026.104990
Received 15 November 2025; Received in revised form 14 February 2026; Accepted 23 October 2026
Available online 26 October 2026
1077-3142/© 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

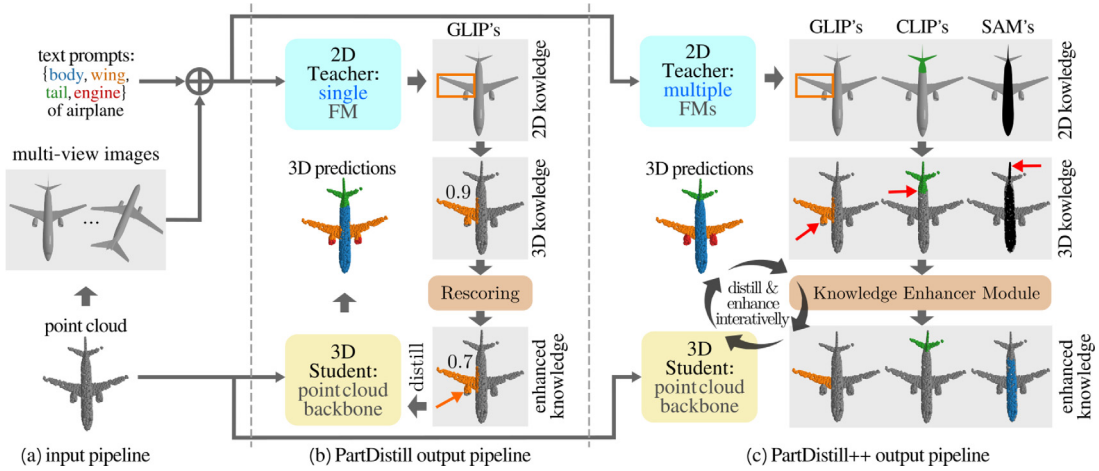


Fig. 1. We present an extended framework, PartDistill++, building upon our preliminary study, PartDistill (Umam et al., 2024), to achieve more robust 3D shape part segmentation by distilling 2D knowledge from multiple foundation models (FMs). Unlike the original PartDistill, which performs a one-time rescoring of each teacher's knowledge based on their qualities (*e.g.*, from 0.9 to 0.7), leaving undesirable portions of the knowledge intact (see the orange arrow), PartDistill++ iteratively learns to extract only the beneficial portions of the teacher's knowledge through the *Knowledge Enhancer Module*, leveraging the strengths of each FM. This iterative process effectively refines the teacher's knowledge over time. In addition, the module can propagate semantic labels to semantically agnostic SAM segments (shown in black), enabling our framework to incorporate SAM segments that produce more precise semantic boundaries.

(I_1) invisible 3D regions in the 2D multi-view images, (I_2) inconsistency of VLM predictions that can lead to negative transfers, and (I_3) the lack of knowledge accumulation across different 3D shapes. To address these challenges, our previous work, PartDistill (Umam et al., 2024) introduces a cross-modal distillation framework, where the student network learns from both the 2D VLM teacher and the 3D geometrical features of the input shapes, effectively leveraging the strengths of both modalities. In our preliminary work (Umam et al., 2024), we observe that while a VLM can produce satisfactory predictions for certain shape categories, it may deliver low-quality results for others, with different VLMs exhibiting varying patterns. In other words, each VLM has its particular strengths and weaknesses. This raises the question: *Can we distill multiple VLMs into a single student network to leverage their combined strengths while mitigating individual weaknesses without using any 3D training data annotations?*

To this end, it is crucial to identify beneficial knowledge from each VLM while avoiding negative transfers that degrade the final 3D segmentation. Negative transfer can arise from unfavorable knowledge of individual VLMs, such as semantic leakage in GLIP (L.H. Li et al., 2022), where box predictions cover pixels from other semantics, or coarse-grained predictions of CLIP, as highlighted by the red arrows in Fig. 1.

In addition to semantically aware VLMs, we also consider a vision model (VM), namely the Segment Anything Model (SAM) (Kirillov et al., 2023), which is designed as a promptable and semantically agnostic segmentation foundation model. SAM aims to decompose images into spatially coherent segments, and is trained using large-scale binary masks without semantic annotations. This design principle makes SAM complementary to VLMs, as it produces spatially coherent image segments with precise boundaries that can later be assigned semantic labels using information from VLMs.

However, integrating such semantically agnostic segments is non-trivial, since semantic labels must be propagated onto SAM segments, and negative transfer may occur when individual segments span multiple semantic entities. One such example is shown in Fig. 1(c), indicated in black, where a SAM segment predominantly covers the body of an airplane while partially including the tail.

Specifically, we propose a ***Knowledge Enhancer Module*** (KEM) to enable the distillation of multiple FMs, including semantically aware VLM and semantically agnostic VM. Importantly, KEM is designed for the realistic case where a predicted segment produced by FMs may cover multiple semantic regions (*i.e.*, mixed semantics), rather than assuming that each segment corresponds to a single semantic label. The module has two key functionalities: $(\mathcal{F}_1)$ *purification* and $(\mathcal{F}_2)$ *semantic propagation*. The former aims to cope with noisy knowledge from individual FMs by filtering out the undesirable portions of the knowledge and retaining the beneficial parts, thus preventing negative transfers to the final 3D segmentation. Meanwhile, the latter targets at propagating reasonable semantic labels to semantically agnostic VM (*i.e.*, SAM) segments, by learning the semantics from VLMs. These two functionalities are achieved by learning a mapping from binary masks associated with each VLM segment to segment targets, estimated using both teacher and student knowledge. We extend the pipeline from our preliminary work (Umam et al., 2024) to incorporate this module, naming it ***PartDistill++***.

Similar to PartDistill, PartDistill++ utilizes a teacher–student model, where FMs serve as the 2D teacher network, and the student network is developed based on a 3D point cloud backbone, as shown in Fig. 1. The student network accepts point clouds of the shape, while the teacher network receives the multi-view images and text prompts, producing 2D predictions as 2D knowledge. For clarity, the figure only shows one 2D prediction from each FM. This 2D knowledge is then transferred to the 3D space as 3D knowledge and passed through the *knowledge enhancer module* to filter out the undesirable portions. For example, the module filters out the semantic leakage in GLIP and purifies the SAM segment that may cover multiple semantics (see the red arrows in Fig. 1). Additionally, it propagates semantic labels to the SAM segments, which are initially semantically agnostic (indicated in back in Fig. 1). Finally, this enhanced knowledge is distilled into the 3D student network, enriching the distillation process.

While PartDistill++ builds upon the same teacher–student distillation framework as PartDistill, it introduces several substantive methodological changes. First, PartDistill++ replaces the confidence-based re-scoring strategy — which uniformly penalizes entire knowledge units and may leave undesirable portions intact — with a learnable module, KEM, that explicitly performs knowledge purification. KEM selectively retains beneficial regions within each teacher prediction while filtering out undesirable portions, enabling PartDistill++ to operate at a finer granularity than PartDistill and more effectively handle noisy and mixed-semantics teacher outputs. Second, PartDistill++ introduces an iterative refinement mechanism, in which the enhanced knowledge produced by KEM and the student predictions are progressively improved

during training. In contrast, PartDistill applies a single-pass re-scoring without feedback, limiting its ability to accumulate and refine knowledge across shapes. Third, PartDistill++ generalizes the framework to incorporate semantically agnostic FMs, such as SAM, via semantic propagation—a capability not supported by PartDistill. This extension enables the effective use of high-quality boundary information and cannot be achieved by the original formulation.

In this work, we extend our preliminary study, PartDistill (Umam et al., 2024), to achieve more robust 3D shape part segmentation. Our key contributions are as follows. First, we present an enhanced framework, ***PartDistill***++, capable of distilling knowledge from multiple FMs—including semantically aware VLMs like CLIP (Radford et al., 2021) and GLIP (L.H. Li et al., 2022), and semantically agnostic VMs such as SAM (Kirillov et al., 2023)—leveraging the strengths of each. Second, we develop a *knowledge enhancer module* to carefully extract the beneficial portions of the teacher's knowledge from each FM and to propagate semantic labels to semantically agnostic SAM segments. Third, extensive experiments show that PartDistill++ outperforms our previous work, PartDistill (Umam et al., 2024), with 8.1% and 9.8% higher mIoUs on the ShapeNetPart (Yi et al., 2016) and PartNetE (M. Liu et al., 2023) datasets, respectively, and surpasses existing methods by a substantial margin.

2. Related work

2.1. Foundation models

A foundation model (FM) is trained on large-scale datasets and can be effectively adapted to a wide range of downstream tasks (Bommasani et al., 2021). FMs can produce direct predictions (Radford et al., 2021; L.H. Li et al., 2022; Kirillov et al., 2023) or generate intermediate representations for further processing (Caron et al., 2021; Oquab et al., 2024). Depending on the data domain, the two primary types of FMs relevant to this work are vision models (VMs) (Kirillov et al., 2023) and vision-language models (VLMs) (Radford et al., 2021; L.H. Li et al., 2022; Zhang et al., 2022). VMs are pre-trained on visual data, while VLMs are pre-trained on both textual and visual data. Consequently, VLM predictions typically exhibit semantic awareness due to their training on textual data, whereas VM predictions are often semantically agnostic. These models can generate various outputs, such as image classifications (Radford et al., 2021; Jia et al., 2021), pixel-wise predictions (B. Li et al., 2022; Kirillov et al., 2023), or bounding box estimations (L.H. Li et al., 2022; Zhang et al., 2022; S. Liu et al., 2023).

In this work, we propose a cross-modal distillation framework that leverages knowledge from multiple complementary 2D FMs to carry out 3D shape part segmentation. Specifically, we exploit the knowledge from both semantically aware VLMs and semantically agnostic VMs, which can generate pixel-wise or bounding box predictions. For simplicity, the terms VM and VLM are used interchangeably with FM throughout the paper.

2.2. 3D part segmentation using foundation models

Several works have explored 3D segmentation without relying on 3D annotations, including unsupervised clustering approaches (Sivri and Sahillioğlu, 2024) and semantic embedding-based methods (Yang et al., 2023; Omata et al., 2019). Without any ground-truth labels, the state-of-the-art methods for 3D part segmentation (M. Liu et al., 2023; Zhu et al., 2023; Abdelreheem et al., 2023; Umam et al., 2024) leverage foundation models (FMs) to transfer knowledge from the 2D domain to 3D space. The pioneering work PointCLIP (Zhang et al., 2022b) and its extended PointCLIPv2 utilize CLIP (Radford et al., 2021) by aggregating its multi-view predictions to the corresponding 3D points through a simple mean operation. PartSLIP (M. Liu et al., 2023) and SATR (Abdelreheem et al., 2023) refine this approach with weighting mechanisms based on the superpoints and face areas covered in the 2D predictions.

A recent work, GeoZe (Mei et al., 2024), introduces a geometrically-driven aggregation scheme that promotes local smoothness and global consistency using hand-crafted geometrical features. Concurrently, our preliminary work, PartDistill (Umam et al., 2024), also incorporates 3D geometrical features to improve 3D segmentations by using learned rather than hand-crafted features. Besides addressing challenges ($\boldsymbol{I_1}$) ~ ($\boldsymbol{I_3}$) discussed in Section 1, one significant advantage of our method over prior approaches (Zhang et al., 2022b; Zhu et al., 2023; M. Liu et al., 2023; Abdelreheem et al., 2023; Mei et al., 2024) is that, by distilling FM knowledge into a lightweight student network, inference can be performed without repeatedly invoking computationally expensive FM teachers. This effectively reduces computational redundancy while maintaining segmentation performance. In this work, we extend PartDistill to PartDistill++ to enable the distillation of knowledge from multiple complementary FMs, harnessing their collective strengths while mitigating individual weaknesses.

2.3. Knowledge distillation

The fundamental work on knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) emphasizes transferring knowledge from a larger model to a smaller one. OpenScene (Peng et al., 2023) applies this principle to distill an FM for 3D understanding, though our focus is on part segmentation while theirs is on indoor/outdoor scene segmentation. A key distinction is that our method can enhance the knowledge sources in the 2D modality through the proposed knowledge enhancer module.

Another line of work, FeatureNeRF (Ye et al., 2023) and MTKD-RL (Yang et al., 2025), explores distilling multiple teacher models. While FeatureNeRF requires labels in its pipeline, MTKD-RL can operate without labels, directly learning to adaptively weight teacher pseudo-labels. This idea aligns with our previous work, PartDistill, which derives weights for each teacher's knowledge. However, beyond weighting, our PartDistill++ focuses on knowledge purification, filtering out the undesirable portions of the knowledge and retaining the beneficial parts. Additionally, PartDistill++ has the ability to perform semantic propagation, allowing it to distill semantically agnostic VMs–approach not feasible in MTKD-RL.

3. Proposed method

We introduce our method with an overview, then detail 2D and 3D knowledge representations for seamless transfer. Next, we describe the knowledge enhancer module for purifying knowledge and propagating semantic labels, followed by the optimization process and implementation details.

3.1. Overview

Given a set of 3D shapes, this work aims to segment each one into R semantic parts without requiring any 3D ground-truth annotations. To achieve this, we transfer knowledge from multiple 2D foundation models (FMs) to the 3D domain via our method ***PartDistill***++. As illustrated in Fig. 2(a), our method takes triplet data as input, which includes the point cloud of the shape with N 3D points, multi-view rendered images from V different viewpoints, and R text prompts, each describing a target semantic part within the 3D shapes.

In the 2D modality, the V multi-view images and the R text prompts are fed into multiple FMs. For each view v, a set of bounding boxes, $B^v = \{b_i^v\}$, is generated by all bounding-box FMs (B-FMs). Similarly, a set of part segments, $S^v = \{s_i^v\}$, is produced by all semantically aware pixel-wise FM (P-FMs), and a set of class-agnostic segments, $X^v = \{x_i^v\}$, by all semantically agnostic P-FMs. In this work, we employ GLIP (L.H. Li et al., 2022) as a B-FM, while CLIP (Radford et al., 2021)

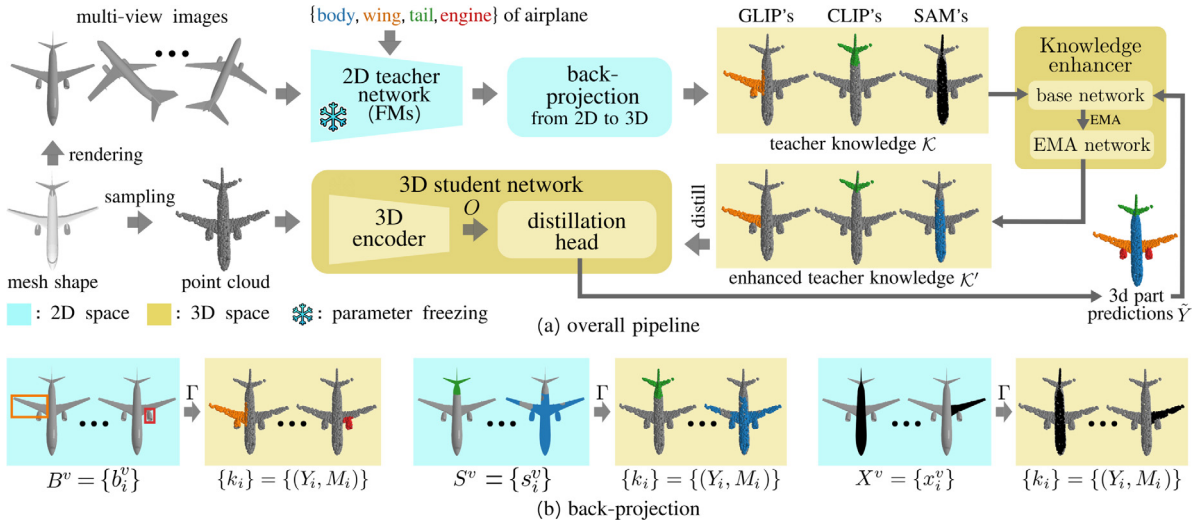


Fig. 2. Overview of the proposed method. (a) The overall pipeline where the knowledge extracted from multiple foundation models (FMs) is distilled into a 3D student network to perform 3D shape part segmentation. Within the pipeline, a *Knowledge Enhancer Module* refines the teacher's knowledge, leading to improved student predictions. These improved predictions are fed back into the module, creating a cyclical process that further refines the teacher's knowledge. This iterative process effectively enhances the final part segmentation. (b) 2D predictions are back-projected into 3D space. The left, middle, and right sub-figures illustrate this process for 2D predictions made by a bounding-box FM (B-FM), a semantically aware pixel-wise FM (P-FM), and a semantically agnostic P-FM, respectively, where Γ denotes the projection operation.

and SAM (Kirillov et al., 2023) are used as semantically aware and semantically agnostic P-FMs, respectively. We then back-project each element of B^v, S^v and X^v into the 3D space, as illustrated in Fig. 2(b). Subsequently, a set of D *teacher knowledge units*, $\mathcal{K} = \{k_d\}_{d=1}^{D}$, is obtained by aggregating the 2D predictions of all FMs from V multi-view images, where each k_d uniquely corresponds to a single 2D prediction, namely b_i, s_i or x_i.

In the 3D modality, each point cloud is fed into the 3D *student* network. This network, equipped with a 3D encoder and a distillation head, produces point-wise part predictions, $\tilde{Y} \in \mathbb{R}^{N\times R}$. In our pipeline, we first improve the teacher knowledge $\mathcal{K}$ by passing it through the proposed *knowledge enhancer module*, undergoing knowledge purification and semantic propagation, yielding enhanced knowledge $\mathcal{K}'$. This enhanced knowledge is then distilled into the student network by aligning $\tilde{Y}$ with $\mathcal{K}'$ via minimizing the proposed loss, $\mathcal{L}_{distill}$ (specified in Section 3.4). The enhanced teacher knowledge $\mathcal{K}'$ leads to better segmentation results, which are then fed back into the knowledge enhancer module, creating a cyclical process that further refines the teacher knowledge. This iterative process effectively improves the final 3D part segmentation. Finally, the segmentation results are obtained by assigning each point to the part category with the highest probability in $\tilde{Y}$.

3.2. From 2D to 3D

To bring the 2D knowledge to the 3D modality, our method renders the 3D shape into V multi-view images $\{I^v \in \mathbb{R}^{H\times W}\}_{v=1}^{V}$ from different V viewpoints. Multiple FMs process these images, along with text prompts of R parts, generating 2D predictions. In the vth image, a set of bounding boxes, $B^v = \{b_i^v\}$, is obtained from all B-FMs, with $b_i^v \in \mathbb{R}^{4+R}$ encoding the box coordinates and the probabilities of the R parts. For the same input image, a set of part probability maps, $S^v = \{s_i^v\}$, is generated from all semantically aware P-FMs, where each $s_i^v \in \mathbb{R}^{H\times W\times R}$ is associated with a segment of a particular class. Meanwhile, a set of class-agnostic segments, $X^v = \{x_i^v\}$, is produced from all semantically agnostic P-FMs, where each $x_i^v \in \{0,1\}^{H\times W}$ corresponds to a single class-agnostic segment.

We back-project each element of B^v, S^v, and X^v into 3D space to derive readily distillable knowledge. Specifically, this operation produces a set of D knowledge units, $\mathcal{K} = \{k_d\}_{d=1}^{D}$, from all FMs across V multi-view images, where each knowledge unit k_d is generated from *exactly one* 2D prediction produced by an FM (either a bounding box b_i, a part probability map s_i, or a class-agnostic segment x_i), via

$$k_d = (Y_d, M_d) = \begin{cases} \Gamma(b_i), & \text{for a B-FM,} \\ \Gamma(s_i), & \text{for a semantically aware P-FM,} \\ \Gamma(x_i), & \text{for a semantically agnostic P-FM,} \end{cases} \tag{1}$$

where Γ denotes the back-projection operation using the camera parameters (Zhang, 2000) employed in the multi-view image rendering. We choose back-projection to obtain point-level distillation targets (Y_d, M_d) that are static for a fixed set of rendered views and teacher predictions, and thus can be precomputed/cached, whereas an alternative formulation that projects the student's evolving 3D predictions to 2D would require repeated rendering throughout training. Each unit k_d comprises two components, Y_d and M_d. $Y_d \in \mathbb{R}^{N\times R}$ represents the point-specific part probabilities, and $M_d \in \{0,1\}^N$ is the mask indicating the 3D points covered by b_i, s_i, or x_i in the 2D space. Note that Y_d is all zeros for semantically agnostic P-FMs at this stage since the corresponding segment is class-agnostic.

Our method aims to transfer the knowledge units $\mathcal{K}$ derived from multiple FMs by distilling them into the 3D student network. First, a 3D encoder, such as Point-M2AE (Zhang et al., 2022a), is applied to the point cloud to extract per-point features, $O \in \mathbb{R}^{N\times E}$, capturing both local and global geometrical information. We then estimate point-wise part predictions, $\tilde{Y} \in \mathbb{R}^{N\times R}$, by feeding the point features O into the distillation head. A straightforward approach to distilling the knowledge is to directly align the part probabilities from the 3D modality, $\tilde{Y}$, with their 2D counterparts, Y. However, as discussed earlier, it is crucial to carefully identify the beneficial knowledge from each FM, maximizing its strengths while avoiding negative transfers. Additionally, direct distillation from SAM is not feasible due to its class-agnostic nature. To address these challenges, we pass the knowledge $\mathcal{K}$ to our knowledge enhancer module.

3.3. Knowledge enhancer module

To distill the knowledge from multiple FMs, the knowledge enhancer module carries out two key functionalities: ($\mathcal{F}_1$) knowledge purification and ($\mathcal{F}_2$) semantic propagation. Purification aims to filter

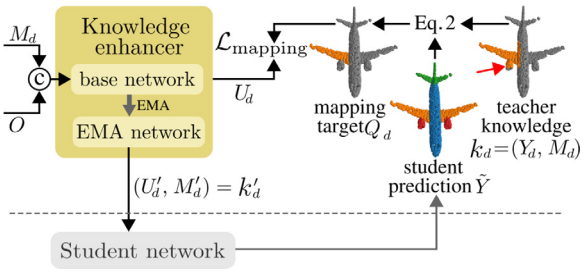


Fig. 3. The knowledge enhancer module refines each teacher knowledge $k_d = (Y_d, M_d)$ by mapping its knowledge mask M_d concatenated with point features O to the target Q_d obtained via Eq. (2). In the figure, © denotes the concatenation operation. This module establishes a cyclical process where its improved output U'_d leads to better student predictions $\tilde{Y}$, and these improved predictions $\tilde{Y}$, in turn, enhance the knowledge U'_d. The figure below the dashed line is provided to better illustrate this cyclical process.

out undesirable portions of each knowledge unit, retaining only beneficial parts, while semantic propagation assigns appropriate semantics to class-agnostic segments derived by SAM. To enable these functions, we learn a mapping function: Given a binary mask M_d from semantically aware FMs (*i.e.*, GLIP and CLIP in this work), the mapping function is trained to identify the beneficial portion of the masked region and assign it a reliable semantic label. To this end, we use the fixed teacher signal Y_d (not learned in this module) together with the student prediction $\tilde{Y}$, and focus on their semantic intersection as the mapping target, since the intersection is more likely to represent high-quality knowledge. Fig. 3 illustrates this process.

Formally, for each knowledge unit $k_d = (Y_d, M_d)$, we compute the semantic intersection with the student prediction $\tilde{Y}$, denoted as Q_d, as

$$Q_d = M_d \odot \left(\underset{r}{\text{argmax}}(Y_d) \Leftrightarrow \underset{r}{\text{argmax}}(\tilde{Y}) \right) \odot \underset{r}{\text{argmax}}(Y_d), \quad (2)$$

where $\odot$ and $\Leftrightarrow$ represent element-wise multiplication and element-wise equality (comparison) operations, respectively. The argmax operation in Eq. (2) is applied row-wise, returning the class with the highest probability for each of the N points. Non-zero elements in $Q_d \in \{0, 1, 2, \ldots, R\}^N$ indicate the R possible intersecting semantics of the corresponding points, while zero values denote the semantically non-intersecting points, including those outside the mask. In our pipeline, knowledge associated with these non-intersecting points is considered undesirable (indicated by the red arrow in Fig. 3) and is mapped to a '*background*' class—a class we aim to exclude. This intersection-based target is intentionally conservative: it retains only regions where the teacher and student agree, while the remaining non-intersecting points are treated as background and excluded from effective supervision. This design reduces the risk of reinforcing large incorrect regions when early student predictions are unstable. Technically, we use label 0s in Q_d to denote the background class.

The knowledge enhancer module takes as input the concatenation of the mask M_d and point features O, generating per-point estimation $U_d \in \mathbb{R}^{N \times (1+R)}$, as illustrated in Fig. 3. The first column of U_d represents the probability of the background class, while the remaining R columns correspond to the R part classes. We then fit every U_d to the mapping target Q_d by minimizing the mapping loss, $\mathcal{L}_{\text{mapping}}$, which is the cross-entropy errors defined by

$$\mathcal{L}_{\text{mapping}} = \sum_{d \in \mathcal{A}} \sum_{n=1}^{N} \sum_{r=0}^{R} \mathbb{1}_{[Q_d(n)=r]} \log(U_d(n, r+1)), \quad (3)$$

where $\mathcal{A}$ represents the index set of all knowledge units from semantically aware FMs, *i.e.*, GLIP and CLIP in this work. $\mathbb{1}_{[\cdot]}$ in Eq. (3) is the indicator function.

By minimizing Eq. (3), the network learns to identify beneficial portions of each teacher's knowledge unit by mapping them to their semantic intersection with the student's knowledge. The remaining portions of each teacher's knowledge unit are treated as undesirable and are mapped to the background class, thus achieving the purification functionality ($\mathcal{F}_1$). We currently utilize knowledge units from GLIP and CLIP, as their semantic maps $\{Y_d\}$ are necessary to obtain the mapping targets $\{Q_d\}$ during the mapping learning. Once this mapping is learned, the network can effectively map all teacher knowledge units, including those from SAM. In this context, a SAM mask M_d obtained via Eq. (1) is concatenated with the point features O and serves as the input to the knowledge enhancer module. An enhanced knowledge U_d, which contains semantic information, is produced, thereby achieving the semantic propagation functionality ($\mathcal{F}_2$).

3.4. Distillation loss and optimization

After purification and semantic propagation via the knowledge enhancer module, the enhanced teacher knowledge from multiple FMs improves the cross-modal knowledge transfer, and can then be distilled into the student network by aligning it with the student's predictions $\tilde{Y}$. However, simultaneously optimizing the student network and the knowledge enhancer module, with each relying on the other's output, can lead to unexpected results. If one network produces unstable outputs, it can negatively impact the other.

To address this issue, we design two sub-networks within the knowledge enhancer module, as shown in Fig. 3. Specifically, we devise a base network and an exponential moving average (EMA) network. The base network directly learns the mapping ability, while the EMA network accumulates the mapping ability by setting its parameters as the exponential moving average of the base network's parameters, *i.e.*,

$$\theta'(t+1) = \gamma\theta'(t) + (1-\gamma)\theta(t), \quad (4)$$

where t denotes the training iteration, γ represents the decay rate, and $\theta'(t)$ and $\theta(t)$ denote the parameter sets of the EMA and base networks at iteration t, respectively. This scheme ensures more stable outputs, as demonstrated in the ablation studies in Section 4.4.

Finally, we conduct the distillation by aligning the student network output $\tilde{Y}$ with the EMA network output $\mathcal{K}' = \{(U'_d, M'_d)\}_{d=1}^{D}$, where $M'_d(n)$ takes value 0 if $\{U'_d(n, r)\}_{r=1}^{1+R}$ has its highest value on the first element, namely the nth point assigned as the background class, and 1 otherwise. The alignment is achieved by minimizing the distillation loss $\mathcal{L}_{\text{distill}}$, defined using a masked cross-entropy formula as,

$$\mathcal{L}_{\text{distill}} = -\sum_{d=1}^{D} \frac{1}{|M'_d|} \sum_{n=1}^{N} \sum_{r=1}^{R} M'_d(n) C_d^n Z_d^{n,r} \log(\tilde{Y}(n, r)), \quad (5)$$

where $|M'_d|$ is the area covered by the mask M'_d, and $C_d^n = \max_r(U'_d(n, r))$ is the confidence score at point n. $Z_d^{n,r}$ takes value 1 if for point n, part r has the highest score in U'_d, and 0 otherwise.

In our pipeline, we simultaneously minimize $\mathcal{L}_{\text{distill}}$ and $\mathcal{L}_{\text{mapping}}$. As we optimize $\mathcal{L}_{\text{mapping}}$, the knowledge enhancer module is improved, leading to progressively better knowledge U'. This enhanced U' positively impacts the distillation process, resulting in better predictions $\tilde{Y}$ from the student network. We then feed these improved $\tilde{Y}$ back into the knowledge enhancer module, creating a cyclical process that further refines the knowledge output U'. This iterative enhancement effectively improves the final part segmentation, as demonstrated in Section 4.4.

3.5. Implementation details

The proposed framework is implemented in PyTorch (Paszke et al., 2019) and trained for 50 epochs. We adopt a warm-up strategy in which KEM is not trained initially, namely the student is optimized only with $\mathcal{L}_{\text{distill}}$ to stabilize its predictions. In principle, KEM can be

Table 1
3D segmentation (mIoU%) on the ShapeNetPart dataset via Foundation Model (FM) knowledge transfer w/o any 3D labels.*

Method	Data type	FM	Airplane	Bag	Cap	Chair	Earphone	Guitar	Knife	Laptop	Mug	Table	Overall
PointCLIPv2 (Zhu et al., 2023)	point cloud	CLIP	33.5	60.4	52.8	51.5	56.5	71.5	66.7	61.6	48.0	61.1	49.5
OpenScene (Peng et al., 2023)			34.4	63.8	56.1	59.8	62.6	69.3	70.1	65.4	51.0	60.4	52.9
GeoZe (Mei et al., 2024)			33.6	70.2	64.7	66.1	63.7	73.6	77.9	74.0	63.2	63.8	56.3
PartDistill (Umam et al., 2024)			40.6	75.6	67.2	65.0	66.3	85.8	79.8	92.6	83.1	68.7	63.9
PartDistill++			40.6	77.1	76.4	66.2	66.4	86.9	79.9	92.5	83.6	69.6	65.0
SATR (Abdelreheem et al., 2023)	mesh	GLIP	32.2	32.1	21.8	25.2	19.4	37.7	40.1	50.4	76.4	22.4	32.3
PartDistill			64.8	64.4	51.0	67.4	48.3	64.8	70.0	83.1	86.5	79.3	56.3
PartDistill++			65.6	65.4	52.9	79.1	50.3	65.7	72.6	83.3	87.1	79.5	58.1
Ensemble	point cloud	CLIP, GLIP	65.2	70.6	71.0	77.6	66.1	86.0	80.6	92.1	85.2	69.5	65.5
MTKD-RL (Yang et al., 2025)			67.1	72.7	74.0	80.1	66.1	87.3	80.8	92.4	89.8	71.5	67.6
PartDistill			66.8	71.8	73.5	81.3	65.8	86.5	81.2	92.8	89.5	70.8	67.5
PartDistill++			67.8	76.0	76.6	84.4	66.4	87.3	83.6	93.5	90.1	72.5	69.7
PartDistill++	point cloud	CLIP, GLIP, SAM	**72.6**	**78.3**	**82.0**	**86.8**	**70.0**	**87.6**	**86.1**	**94.1**	**94.1**	**80.2**	**72.2**
PartDistill++	mesh		**65.9**	**66.3**	**68.3**	**79.9**	**56.9**	**82.8**	**72.6**	**83.6**	**88.8**	**79.7**	**62.8**

*Results for other categories, including those in Table 2 and Table 3, can be found in the supplementary material.

Table 2
3D segmentation (mIoU%) on the PartNetE dataset via Foundation Model (FM) knowledge transfer w/o any 3D labels.

Method	Data type	FM	Bottle	Box	Chair	Display	Kettle	Knife	Lamp	Oven	Suitcase	Table	Overall
PartSLIP (M. Liu et al., 2023)	point cloud	GLIP	76.3	57.5	60.7	43.8	20.8	46.8	37.1	33	40.2	47.7	27.3
PartDistill (Umam et al., 2024)			77.4	69.7	74.1	50.5	24.2	59.2	58.8	34.2	43.2	50.2	39.9
PartDistill++			77.4	72.5	75.4	52.2	38.6	59.1	59.0	36.6	44.6	52.6	41.7
Ensemble	point cloud	CLIP, GLIP	76.6	74.7	76.6	52.9	32.5	57.4	58.3	36.8	43.4	53.3	42.2
MTKD-RL (Yang et al., 2025)			76.6	75.7	77.9	55.0	37.0	58.3	58.1	39.8	45.6	53.7	44.3
PartDistill			76.8	75.4	77.6	55.8	36.7	57.8	58.4	40.1	44.9	53.9	44.5
PartDistill++			77.2	78.5	79.5	57.5	38.4	59.2	58.8	45.2	52.4	55.7	46.7
PartDistill++	point cloud	CLIP, GLIP, SAM	**77.9**	**80.3**	**80.1**	**61.4**	**39.2**	**59.4**	**59.2**	**49.7**	**60.1**	**56.4**	**49.7**

enabled once $\mathcal{L}_{\text{distill}}$ enters a stable regime (*e.g.*, the moving-average loss exhibits only small epoch-to-epoch changes below a preset threshold for several consecutive epochs). In practice, we use 20 warm-up epochs, after which the framework is optimized with both $\mathcal{L}_{\text{distill}}$ and $\mathcal{L}_{\text{mapping}}$. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 and a batch size of 8. Following PartDistill, the student network employs Point-M2AE (Zhang et al., 2022a), pre-trained in a self-supervised manner on the ShapeNet55 dataset (Chang et al., 2015), as the 3D encoder. The encoder's weights are frozen, and only the learnable parameters in the distillation head are updated. The distillation head uses a 4-layer multi-layer perceptron with ReLU activation (Agarap, 2018). Unless further specified, the knowledge enhancer module uses a decay rate of 0.7 for the EMA update, which is performed at the transition between epochs. The module adopts a self-attention network (Vaswani et al., 2017) as its base network, with detailed architectural specifications provided in the supplementary material.

To ensure fair comparisons with competing methods (Zhang et al., 2022b; Zhu et al., 2023; M. Liu et al., 2023; Abdelreheem et al., 2023; Mei et al., 2024), we follow their settings, including the text prompts and the 2D rendering. Each shape is rendered into 10 multi-view images from either a sparse point cloud (Zhang et al., 2022b; Zhu et al., 2023; Mei et al., 2024), a dense point cloud (M. Liu et al., 2023), or a mesh shape (Abdelreheem et al., 2023). Following prior works and our preliminary ablation study (Umam et al., 2024), this number of views is sufficient, as performance typically saturates around this setting. All FMs (CLIP, GLIP, and SAM) are used with their default configurations and prompting strategies following prior works (PointCLIPv2 for CLIP, PartSLIP for GLIP, and SAM's default regular-grid prompting), and no model-specific hyperparameter tuning is performed.

4. Experimental results

4.1. Dataset and evaluation metric

We use two major benchmark datasets, ShapeNetPart (Yi et al., 2016) and PartNetE (M. Liu et al., 2023), to assess the performance of our method. ShapeNetPart has 16 categories with a total of 31,963 shapes, while PartNetE comprises 2266 shapes of 45 categories. The mean intersection over union (mIoU) (Mo et al., 2019) between the segmentation results and the ground-truth labels is adopted as the evaluation metric in the test-set data. Following PartDistill (Umam et al., 2024), we first benchmark our method on transferring 2D knowledge from foundation models for 3D part segmentation without requiring any 3D ground-truth labels. Then, we present experiments when additional few-shot labels are available.

4.2. 3D part segmentation via 2D knowledge transfer

We evaluate the performance of our method, PartDistill++, working with a single or multiple FMs. For a fair comparison, we follow the settings of existing methods (Zhu et al., 2023; Abdelreheem et al., 2023; M. Liu et al., 2023; Peng et al., 2023; Mei et al., 2024) that perform 3D segmentation by transferring 2D FM knowledge, reporting their mIoU performances from their respective papers. For the pixel-wise foundation model (P-FM), we use the protocols of PointCLIPv2 (Zhu et al., 2023), utilizing CLIP (Radford et al., 2021) with a ViT-B/32 (Dosovitskiy et al., 2020) backbone for pixel-wise predictions. For the bounding box foundation model (B-FM), we employ the GLIP-Large model (L.H. Li et al., 2022), aligning with PartSLIP (M. Liu et al., 2023) and SATR (Abdelreheem et al., 2023). Most competing methods report their performance on ShapeNetPart, while PartSLIP evaluates on PartNetE. Additionally, we extend OpenScene (Peng et al., 2023) and MTKD-RL (Yang et al., 2025) for 3D part segmentation working with single and multiple teachers, respectively, with Point-M2AE (Zhang et al., 2022a) used as backbone for a fair comparison.

The major advantage of PartDistill++ lies in its ability to exploit knowledge from multiple FMs, both semantically aware and semantically agnostic, thereby leveraging their combined strengths. In contrast, other benchmark methods are developed to utilize a single FM (Zhu et al., 2023; Abdelreheem et al., 2023; M. Liu et al., 2023; Peng et al., 2023; Mei et al., 2024; Umam et al., 2024) and/or are limited to only semantically aware FMs (Yang et al., 2025). This distinctive capability allows our method to significantly outperform the benchmark methods in most categories of the ShapeNetPart and PartNetE datasets, as shown in Tables 1 and 2, respectively.

To illustrate, we begin by benchmarking an ensemble of two semantically aware teachers — CLIP and GLIP — which together guide

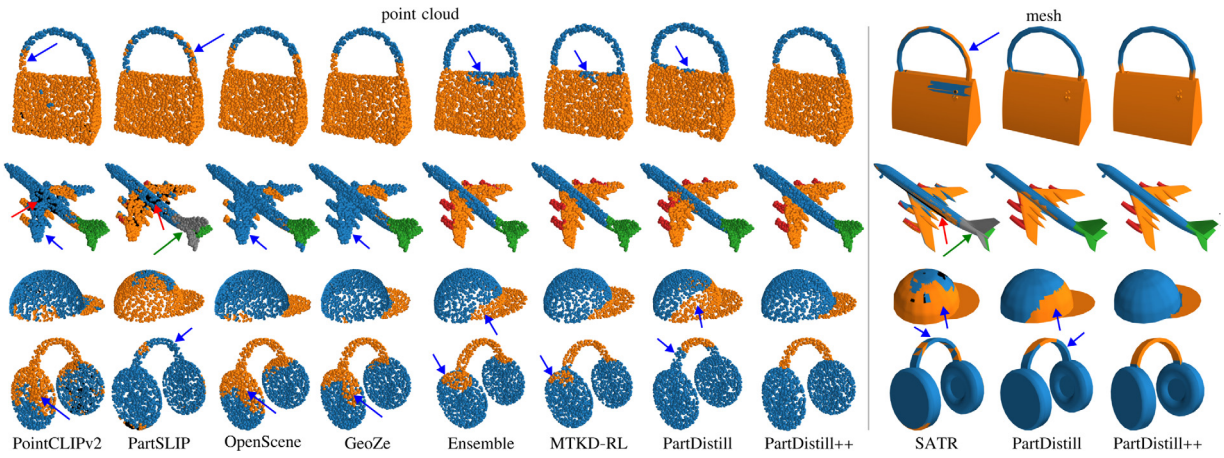


Fig. 4. Visualization of 3D part segmentation results on the ShapeNetPart dataset, where different parts are depicted in different colors. Invisible 3D regions in 2D multi-view images (issue I_1) are shown in black, highlighted by red arrows. Blue and green arrows indicate cases of negative transfers (issue I_2) and lack knowledge accumulation (issue I_3), respectively.

the student network and serve as our baseline, achieving an overall mIoU of 65.5% (referred to as Ensemble in Table 1). MTKD-RL (Yang et al., 2025), which learns adaptive teacher weighting via reinforcement learning, boosts performance to 67.6%, a gain of 2.1%. Similarly, PartDistill applies backward distillation to reweight teacher knowledge, yielding a 2.0% improvement to 67.5%.

Unlike these methods, which focus on reweighting and often leave undesirable portions intact (see red arrows in Fig. 5), PartDistill++ goes further: it actively purifies each teacher's knowledge by excluding undesirable portions while retaining only the beneficial parts. This principled refinement leads to a 4.2% gain over baseline, from 65.5% to 69.7%. Moreover, with its ability to incorporate semantically agnostic FMs, i.e., SAM in our case, into the multi-teacher distillation process, PartDistill++ achieves an even greater gain of 6.7%, boosting from 65.5% to 72.2%. Consistent improvement patterns are also observed on the PartNetE dataset, as shown in Table 2.

We further evaluate PartDistill++ in the single-FM setting. When using CLIP as the sole FM, PartDistill++ achieves an overall mIoU of 65.0% on the ShapeNetPart dataset, 8.7% higher than the best mIoU reported by other methods (GeoZe), demonstrating a substantial improvement. This is primarily because PartDistill++ integrates 3D geometric features, unlike PointCLIPv2, and possesses the ability to refine the source of knowledge, which GeoZe and OpenScene lack. As a result, it better handles inconsistencies in FM predictions and mitigates negative transfers (issue I_2). Compared to our preliminary study PartDistill, PartDistill++ achieves notable better performance due to the iterative enhancement in the knowledge enchanter module, as verified in Section 4.4.

When GLIP serves as the FM alone, PartDistill++ consistently achieves better results compared to other methods. Specifically, it obtains the mIoUs, which are 25.8% and 14.4% higher than those of SATR on the ShapeNetPart dataset (Table 1) and PartSLIP on the PartNetE dataset (Table 2), respectively. In SATR and PartSLIP, the use of GLIP can intensify the issue of uncovered 3D regions (I_1) due to potential undetected areas, and escalate the negative transfer issue (I_2) due to semantic leakages, where box predictions often include points from other semantics.

In addition to the quantitative comparisons, a qualitative comparison of the segmentation results is presented in Fig. 4. It is evident that methods directly transferring 2D predictions to 3D segmentation, namely PointCLIPv2, PartSLIP, and SATR, struggle with incomplete segmentation in occluded regions (issue I_1), shown in black and highlighted by red arrows. These methods may also suffer from a lack of knowledge accumulation across shapes (issue I_3), where erroneous 2D predictions in individual shapes just remain as incorrect 3D segmentation. For example, if GLIP in PartSLIP and SATR fails to detect the airplane tail in several 2D multi-view images, it results in an incorrect absence of prediction in the 3D modality (depicted in gray, highlighted by the green arrows in Fig. 4). Other methods partially mitigate these issues by leveraging geometric features (GeoZe, OpenScene, PartDistill) and/or distilling from multiple teacher networks (Ensemble and MTKD-RL), yet still exhibit negative transfer effects (issue I_2), highlighted by the blue arrows. PartDistill++ goes further with its purification and semantic propagation capabilities, yielding notable improvements as shown in Fig. 4.

4.3. 3D part segmentation with additional few-shot labels

We further evaluate our method when additional few-shot labels are available by adopting the setup used in PartSLIP (M. Liu et al., 2023). Specifically, a fine-tuned GLIP model (L.H. Li et al., 2022) provided by PartSLIP is used, where 8-shot labeled shapes are utilized to fine-tune the model in each category of the PartNetE dataset (Yi et al., 2016). Additionally, we integrate off-the-shelf CLIP and SAM models for the multi-FM setting. Alongside distilling knowledge from the FM teacher, we also ask the student network to minimize the standard cross-entropy loss between its outputs and the 8 labeled shapes.

As demonstrated in Table 3, the few-shot 3D segmentation methods ACD (Gadelha et al., 2020) and Prototype (Zhao et al., 2021) are adapted to the PointNet++ (Zhao et al., 2021) and PointNext (Qian et al., 2022) backbones, respectively, resulting in average performance improvements for these backbones. PartSLIP (M. Liu et al., 2023) and PartSTAD (Kim and Sung, 2024) leverage multi-view FM predictions to enhance 3D segmentation, leading to improved mIoU scores. Under the same FM settings, our methods enhance the sources of knowledge via the rescoring mechanism in PartDistill and the knowledge enhancer module in PartDistill++ and can yield better segmentation results.

PartDistill++ improves PartDistill from 65.9% to 69.6% by leveraging multiple FMs, including semantically aware and semantically agnostic FMs, whereas PartDistill is designed to work with a semantically aware FM. From the table, PartDistill++ performs slightly better than PartDistill with GLIP as the sole FM and shows notable improvements when "SAM" or "CLIP and SAM" are incorporated. In the last two rows of Table 3, with or without CLIP, PartDistill++ reaches similar performance. One reason is that with additional few-shot labels, CLIP with coarse segments might not be very beneficial since the knowledge enhancer module can learn semantic propagation from the few-shot labels and GLIP segments. Additionally, we present results from fine-tuning Point-M2AE with the few-shot labels, which exhibit lower performance compared to our method, underscoring the significant impact of our distillation framework.

Table 3
3D segmentation (mIoU%) on the PartNetE dataset via FM knowledge transfer with additional few-shot labels.

Method	FM	Bottle	Box	Chair	Display	Kettle	Knife	Lamp	Oven	Suitcase	Table	Overall
PointNet++ (Qi et al., 2017)	Non-FM-based	27.0	24.5	42.2	30.2	28.6	22.2	10.5	19.4	3.3	7.3	20.4
PointNext (Qian et al., 2022)		67.6	69.4	65.1	53.7	60.6	59.7	55.4	36.8	14.5	22.1	40.6
ACD (Gadelha et al., 2020)		22.4	21.1	39.0	29.2	40.2	39.6	13.7	8.9	13.2	13.5	23.2
Prototype (Zhao et al., 2021)		60.1	68.8	70.8	67.3	62.7	50.4	38.2	36.5	35.5	25.7	44.3
Point-M2AE (Zhang et al., 2022a)		72.4	79.0	83.4	74.3	64.3	68.0	57.6	53.3	57.5	33.6	56.4
PartSLIP (M. Liu et al., 2023)	GLIP	83.4	84.5	85.3	84.8	77.0	65.2	60.0	**73.5**	70.4	42.4	59.4
PartDistill (Umam et al., 2024)		84.6	87.9	88.4	87.4	78.6	71.4	69.2	72.8	73.4	63.3	65.9
PartDistill++		84.5	88.5	88.2	87.6	78.9	71.6	70.1	72.6	73.2	64.3	66.5
PartSTAD (Kim and Sung, 2024)	GLIP, SAM	83.6	81.1	85.3	82.3	84.2	63.8	68.4	71.9	68.3	47.8	65.0
PartDistill++		84.5	90.8	87.5	87.6	78.6	72.6	72.8	71.9	73.8	**65.6**	69.5
PartDistill++	CLIP, GLIP, SAM	**84.6**	**92.8**	**88.9**	**88.6**	**79.5**	**73.8**	**73.4**	73.2	**74.5**	65.4	**69.6**

4.4. Ablation studies

Proposed components. We conduct ablation studies on the proposed components, and the mIoU scores computed between the 3D segmentation output and ground-truth labels of the ShapeNetPart dataset are presented in Experiments (1) ~ (9) of Table 4. First, we compare the segmentation performances obtained via direct voting and cross-modal distillation. In Experiment (1), direct voting is used to obtain the 3D segmentation results, where part labels are assigned by voting from multi-view box predictions made by GLIP. In contrast, in Experiment (2), the segmentation results are achieved through cross-modal distillation, which minimizes $\mathcal{L}_{\text{distill}}$, employing 2D GLIP and a 3D encoder as the teacher and student networks, respectively, leveraging the strengths of both modalities.

Based on the mIoU scores in Experiments (1) and (2), it is evident that 3D segmentation obtained via cross-modal distillation achieves a substantial performance improvement, with an mIoU increase of 10.5% compared to direct voting. Direct voting faces several issues, including ($\boldsymbol{I_1}$) uncovered 3D regions in 2D multi-view images, ($\boldsymbol{I_2}$) negative transfers due to inconsistency in FM predictions, and ($\boldsymbol{I_3}$) lack of knowledge accumulation across different 3D shapes. Cross-modal distillation alleviates these issues by simultaneously learning 2D features from the FM teacher network and 3D geometrical features from the shape collections by the 3D student network.

From Table 1, it can be observed that using individual FM produces satisfactory segmentation on several categories but lower quality results on others, with different FMs exhibiting varying patterns. For instance, using GLIP produces satisfactory segmentation on the chair category but lower quality on the earphone category, while using CLIP yields favorable segmentation on the earphone category but less favorable results for the airplane category. We then investigate the complementary impact of using multiple FMs in cross-modal distillation. In Experiments (2) and (3), we present the segmentation performances obtained by distilling individual GLIP and CLIP models, respectively. In Experiment (4), we jointly employ GLIP and CLIP, distilling the predictions from both into the 3D student network, resulting in a notable improvement of 4.6% higher mIoU compared to the best result of using individual FMs in Experiments (2) and (3).

However, distilling multiple FMs by naively merging predictions from them, such as GLIP and CLIP in Experiment (4), may hinder optimal performance. Not all knowledge from individual FMs is beneficial, and incorporating unfavorable knowledge can degrade the final segmentation results, preventing the achievement of optimal performance. To address this, we incorporate our *Knowledge Enhancer Module* (KEM) in Experiment (5). The KEM performs purification ($\mathcal{F}_1$), filtering out undesirable portions of the knowledge from GLIP and CLIP, while retaining only the beneficial parts. As shown in Experiment (5), a substantially higher mIoU of 4.3% is achieved compared to Experiment (4). By purifying each FM's knowledge, the KEM enhances the knowledge source, subsequently improving the final 3D segmentation. The visualization of this knowledge purification, particularly from GLIP and CLIP, is depicted in the left part of Fig. 5.

Table 4
Ablation studies on the proposed method.

No	Employed FM			Distill. ($\mathcal{L}_{\text{distill}}$)	KEM ($\mathcal{L}_{\text{mapping}}$)	Re-scoring	mIoU
	GLIP	CLIP	SAM				
1	✓						50.3
2	✓			✓			60.8
3		✓		✓			60.1
4	✓	✓		✓			65.4
5	✓	✓		✓	✓		69.7
6	✓	✓		✓		✓	67.5
7	✓	✓	✓	✓	✓		**72.2**
8	✓	✓	✓	✓	w/o SAM purification		71.3
9	✓	✓	✓	✓	one-time		69.5

Under the same setting where GLIP and CLIP are employed as the FMs, we adopt the re-scoring mechanism introduced in our preliminary work, PartDistill, which updates the score of each FM's knowledge unit based on its quality. The result provided in Experiment (6) shows notably lower performance compared to (5), which uses the KEM. This is because the re-scoring mechanism enhances the knowledge only once through a one-time, and penalizes each piece of knowledge as a whole, leaving undesirable portions intact (see the red arrows in Fig. 5). In contrast, the KEM iteratively enhances the knowledge, extracting the beneficial portions and filtering out the undesirable parts.

Unlike PartDistill's re-scoring mechanism, which is limited to accepting semantically aware FMs, the KEM can work with semantically agnostic FMs. SAM, a semantically agnostic FM, produces segments with more precise boundaries (Fig. 5 shows two examples of such segments in black) and can potentially improve the segmentation results. In addition to performing knowledge purification ($\mathcal{F}_1$), the KEM also conducts semantic propagation ($\mathcal{F}_2$), assigning appropriate semantics to the SAM's knowledge units as it learns semantic mapping via Eq. (3) using CLIP and GLIP. The visualization of this semantic propagation is shown in the right-hand side of Fig. 5. Integrating SAM through the KEM achieves a significantly higher mIoU of 72.2%, compared to 69.7% without SAM, as shown in Experiments (7) and (5), respectively. This improvement highlights a capability that PartDistill does not support: semantic propagation that assigns part semantics to class-agnostic SAM segments, enabling their effective use in the distillation process.

We then ablate the KEM under different settings. First, we turn off knowledge purification for SAM, *i.e.*, only propagating semantics and using their original segments as they are. Note that turning off the knowledge purification for GLIP and CLIP is equivalent to the result in Experiment (4). Turning off knowledge purification for SAM in the KEM, as presented in Experiment (8), leads to a notable degradation in mIoU compared to Experiment (7). This highlights the importance of both ($\mathcal{F}_1$) knowledge purification and ($\mathcal{F}_2$) semantic propagation in SAM to reach better segmentation results.

Next, we make the KEM perform one-time enhancement instead of iterative enhancement. Specifically, we pre-train the KEM and use it in the distillation pipeline without further updating its parameters. In this setting, as shown in Experiment (9), the mIoU experiences a

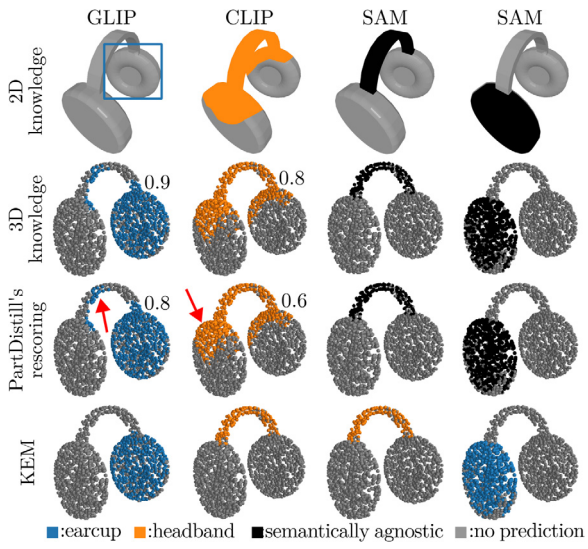


Fig. 5. Visualization of the enhanced knowledge using the PartDistill rescoring mechanism and our Knowledge Enhancer Module (KEM) in PartDistill++. Given a GLIP's bounding box and CLIP's segment that locate the earcup and headband of the earphone shape, respectively, beyond the desired semantics, our KEM performs knowledge purification ($\mathcal{F}_1$), which filters out the undesired parts. In contrast, PartDistill's rescoring only updates the confidence scores, leaving the undesired segment intact (as indicated by the red arrows). For SAM, given semantically agnostic segments, our KEM conducts semantic propagation ($\mathcal{F}_2$), assigning appropriate semantics. However, in PartDistill, these segments remain semantically agnostic since it lacks the capability to handle such segments.

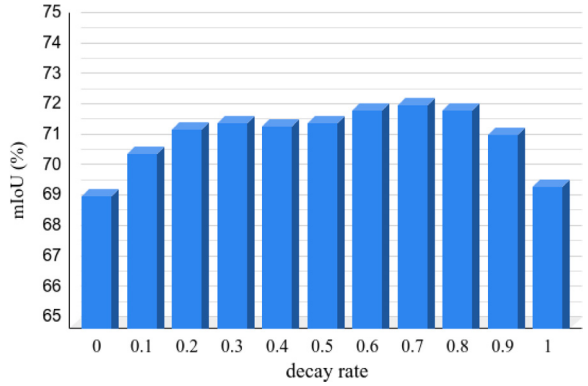


Fig. 6. Ablation study on different EMA decay rate values on the ShapeNetPart dataset.

notable drop, highlighting the importance of iterative enhancement in the KEM.

EMA decay rate. As discussed in Section 3.4 in the main paper, to stabilize the simultaneous optimization of the KEM parameters and the student network parameters in our pipeline, we take the exponential moving average of base network parameters to update the EMA network. This stabilization is crucial for maximizing the effect of iterative enhancement in the KEM. Specifically, we ablate the EMA decay rate γ from 0 to 1 in Eq. (4), where a higher value represents greater stability but results in slower updates regarding the enhancement ability from the base network, and vice versa. Under different settings of the EMA decay rate γ, mIoU scores are collected from the ShapeNetPart dataset and presented in Fig. 6.

From the figure, our method achieves satisfactory segmentation performance across a broad range of decay rate values, specifically from $\gamma = 0.2$ to $\gamma = 0.8$, with the best performance at $\gamma = 0.7$, resulting in an mIoU of 72.2%. When γ is 0.1 or lower, the mIoU performance decreases, indicating that the learned enhancement ability in the base network has not been stable enough to effectively contribute to updating the EMA network. At $\gamma = 0$, it is equivalent to directly using the base network, resulting in the instability and a suboptimal mIoU of 69.2%. For $\gamma = 0.9$ or higher, the mIoU also degrades because the updates from the base network are too slow. Setting γ to 1 is equivalent to the one-time enhancement presented in Experiment (9) of Table 4 in the main paper, where the KEM uses the initial parameters of the network, from the training right after the warm-up epochs, and does not get updated on the remaining optimization epochs.

Table 5
Inference time and number of parameters comparison with benchmark methods, run in NVIDIA A5000.

Method	Inference time (ms)			Numb. of params.	
	Rendering	2D	3D	3D	2D
PointCLIPv2 (Zhu et al., 2023)	232	44	4	–	149.6 M
GeoZe (Mei et al., 2024)	232	44	29	–	149.6 M
SATR (Abdelreheem et al., 2023)	2.0 K	11.2 K	649.6 K	–	430.4 M
PartSLIP (M. Liu et al., 2023)	9.8 K	3.4 K	41.9 K	–	430.4 M
Ours	–	–	25	26.9 M	–

4.5. Computational comparison

Table 5 presents a comparative analysis of inference latency and parameter count across several benchmark methods, evaluated on an NVIDIA A5000 GPU. In contrast to prior approaches that rely on heavy 2D backbones and rendering pipelines, our method is entirely distillation-based. Once the distillation process is complete, inference proceeds directly on raw point clouds, eliminating the need for image rendering and 2D teacher inference.

This architectural shift yields substantial computational savings. Methods such as SATR and PartSLIP suffer from high latency due to their dependence on large 2D teachers (*e.g.*, 430.4M parameters from GLIP) and computationally intensive 3D voting mechanisms–such as geodesic reweighting in SATR and superpoint generation in PartSLIP. In contrast, our distilled model completes inference in just 25 ms, using only 26.9M parameters from a lightweight 3D student network. While PointCLIPv2 and GeoZe offer faster runtimes than SATR and PartSLIP, they still rely on 2D backbones (149.6M parameters from CLIP) and rendering stages, resulting in overall latency more than an order of magnitude higher than ours.

By decoupling from the 2D teacher and optimizing a lightweight 3D student with our framework, the proposed method not only achieves desirable inference speed but also delivers state-of-the-art segmentation performance. This makes it particularly well-suited for real-time deployment scenarios, where latency and memory constraints are critical. The omission of rendering and 2D inference stages is a deliberate architectural choice—made possible through our framework.

4.6. Real-world demo on scanned point clouds

To demonstrate the robustness and real-world applicability of our method, we conduct experiments on real-world scanned point clouds provided by PartSLIP. These point clouds are captured using an iPhone 12 Pro Max equipped with a LiDAR sensor and contain realistic challenges such as sensor noise and non-uniform point density. Note that this real-world data does not provide ground-truth labels.

We apply our method to these scans by first uniformly sampling 2048 points from each input point cloud and performing part segmentation on this sparse representation. The predicted labels are then transferred to the original dense point cloud via k-nearest neighbor voting with $k = 20$. Fig. 7 shows qualitative results on two representative object categories, display and chair. From left to right, we visualize the raw input point cloud, the segmentation results on the 2048-point



Fig. 7. Part segmentation on real-world scanned point clouds. From left to right: the raw input point cloud, the segmentation on the 2048-point subsampled cloud, and the segmentation transferred to the original dense scan.

subsampled cloud, and the segmentation transferred to the original dense scan. As shown in the figure, our method produces coherent and semantically meaningful part segmentation on real-world scans, and the transferred predictions preserve meaningful part structures on the dense point cloud.

One reasonable explanation for this strong real-world generalization is that our FM teachers are trained on large-scale datasets, including real-world data, which provide strong semantic priors and improve generalizability across domains. These additional experiments demonstrate that our method is not limited to clean synthetic benchmarks and can be effectively applied to real-world scanned point clouds.

5. Conclusion

In this work, we extend our preliminary study, PartDistill, to develop a more robust framework for 3D shape part segmentation, named PartDistill++. By leveraging the strengths of multiple foundation models (FMs), including semantically aware and semantically agnostic FMs, PartDistill++ effectively addresses the challenges of transferring 2D FM predictions to 3D segmentation. The introduction of the knowledge enhancer module, with its purification and semantic propagation functionalities, enables the careful extraction of beneficial knowledge, harnessing the strengths of each FM while mitigating individual weaknesses. This module ensures the enhanced knowledge is distilled into the 3D student network, resulting in significant improvements over our previous work. Extensive experiments demonstrate that PartDistill++ significantly surpasses existing methods on major benchmark datasets, including ShapeNetPart and PartNetE.

CRediT authorship contribution statement

Ardian Umam: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Cheng-Kun Yang:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Min-Hung Chen:** Writing – review & editing, Validation, Methodology, Formal analysis. **Jen-Hui Chuang:** Writing – review & editing, Validation, Investigation, Funding acquisition. **Yen-Yu Lin:** Writing – review & editing, Validation, Methodology, Funding acquisition, Conceptualization.

Funding

This work was supported by the National Science and Technology Council (NSTC) under grants 114-2221-E-A49-038-MY3, 115-2221-E-A49-119-MY3, and 115-2634-F-A49-011, as well as MediaTek Inc.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2026.104990.

Data availability

The code and data for this work is available at https://github.com/ardianumam/PartDistill-plusplus.

References

Abdelreheem, A., Skorokhodov, I., Ovsjanikov, M., Wonka, P., 2023. SATR: Zero-shot semantic segmentation of 3D shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Agarap, A.F., 2018. Deep learning using rectified linear units (ReLU). ArXiv Preprint.

Aleotti, J., Caselli, S., 2011. Part-based robot grasp planning from human demonstration. In: Proceedings of the IEEE International Conference on Robotics and Automation.

Aleotti, J., Caselli, S., 2012. A 3D shape segmentation approach for robot grasping by parts. Robot. Auton. Syst..

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. ArXiv Preprint.

Buciluǎ, C., Caruana, R., Niculescu-Mizil, A., 2006. Model compression. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al., 2015. ShapeNet: An information-rich 3D model repository. ArXiv Preprint.

Cheng, Q., Wen, K., Gu, X., 2022. Vision-language matching for text-to-image synthesis via generative adversarial networks. IEEE Trans. Multimed..

Ding, H., Liu, C., Wang, S., Jiang, X., 2022. VLT: Vision-language transformer and query generation for referring segmentation. IEEE Trans. Pattern Anal. Mach. Intell..

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations.

Gadelha, M., RoyChowdhury, A., Sharma, G., Kalogerakis, E., Cao, L., Learned-Miller, E., Wang, R., Maji, S., 2020. Label-efficient learning on point clouds using approximate convex decompositions. In: Proceedings of the European Conference on Computer Vision.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. Adv. Neural Inf. Process. Syst. Work..

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the International Conference on Machine Learning.

Ke, J., Wang, J., Chen, J.C., Jhuo, I.H., Lin, C.W., Lin, Y.Y., 2023. CLIPREC: Graph-based domain adaptive network for zero-shot referring expression comprehension. IEEE Trans. Multimed..

Kim, H., Sung, M., 2024. PartSTAD: 2D-to-3D part segmentation task adaptation. In: Proceedings of the European Conference on Computer Vision.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. ArXiv Preprint.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R., 2022. Language-driven semantic segmentation. In: Proceedings of the International Conference on Learning Representations.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al., 2022. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al., 2023. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. ArXiv Preprint.

Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H., 2023. PartSLIP: Low-shot part segmentation for 3D point clouds via pretrained image-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Mei, G., Riz, L., Wang, Y., Poiesi, F., 2024. Geometrically-driven aggregation for zero-shot 3D point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Michel, O., Bar On, R., Liu, R., Benaim, S., Hanocka, R., 2022. Text2Mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H., 2019. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Omata, K., Furuya, T., Ohbuchi, R., 2019. Annotating 3D models and their parts via deep feature embedding. In: International Conference on Multimedia & Expo Workshops.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2024. DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res..

Osman, N., Camporese, G., Ballan, L., 2024. Multi-modal transformer with language modality distillation for early pedestrian action anticipation. Comput. Vis. Image Underst..

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst..

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al., 2023. Openscene: 3D scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst..

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B., 2022. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. Adv. Neural Inf. Process. Syst..

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning.

Sivri, T.T., Sahillioğlu, Y., 2024. A data-centric unsupervised 3D mesh segmentation method. Vis. Comput..

Ückermann, A., Haschke, R., Ritter, H., 2013. Realtime 3D segmentation for human-robot interaction. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.

Umam, A., Yang, C.K., Chen, M.H., Chuang, J.H., Lin, Y.Y., 2024. PartDistill: 3D shape part segmentation by vision-language model distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems.

Wei, Y., Zhang, Y., Li, K., Wang, F., Tang, S., Zhang, Z., 2025. Leveraging vision-language prompts for real-world image restoration and enhancement. Comput. Vis. Image Underst..

Xu, X., Ruan, Y., Sridhar, S., Ritchie, D., 2022. Unsupervised kinematic motion detection for part-segmented 3d shape collections. In: Proceedings of the ACM SIGGRAPH Conference.

Yang, Y., Hayat, M., Jin, Z., Zhu, H., Lei, Y., 2023. Zero-shot point cloud segmentation by semantic-visual aware synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Yang, C., Yu, X., Yang, H., An, Z., Yu, C., Huang, L., Xu, Y., 2025. Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Ye, J., Wang, N., Wang, X., 2023. FeatureNeRF: Learning generalizable NeRFs by distilling foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L., 2016. A scalable active framework for region annotation in 3D shape collections. ACM Trans. Graph..

Yu, Q., Irie, G., Aizawa, K., 2025. Open-set domain adaptation with visual-language foundation models. Comput. Vis. Image Underst..

Zanella, L., Liberatori, B., Menapace, W., Poiesi, F., Wang, Y., Ricci, E., 2024. Delving into CLIP latent space for video anomaly recognition. Comput. Vis. Image Underst..

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell..

Zhang, R., Guo, Z., Gao, P., Fang, R., Zhao, B., Wang, D., Qiao, Y., Li, H., 2022a. Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. Adv. Neural Inf. Process. Syst..

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

Zhang, W., Wu, L., Zhang, Z., Yu, T., Ma, C., Jin, X., Yang, X., Zeng, W., 2024. Unleash the power of vision-language models by visual attention prompt and multi-modal interaction. IEEE Trans. Multimed..

Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J., 2022. GLIPv2: Unifying localization and vision-language understanding. Adv. Neural Inf. Process. Syst..

Zhao, N., Chua, T.S., Lee, G.H., 2021. Few-shot 3D point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhu, X., Zhang, R., He, B., Zeng, Z., Zhang, S., Gao, P., 2023. PointCLIP V2: Adapting clip for powerful 3D open-world learning. Proc. the IEEE/CVF Int. Conf. Comput. Vis..