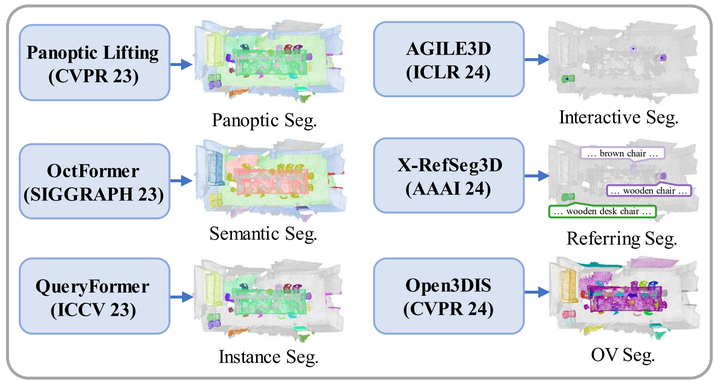
A Unified Framework for 3D Scene Understanding

Wei Xu*, Chunsheng Shi*, Sifan Tu, Xin Zhou, Dingkang Liang, Xiang Bai†

Huazhong University of Science and Technology

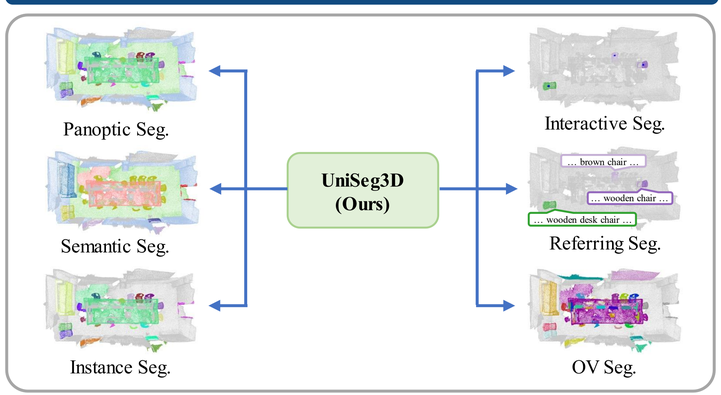
*Equal contribution. †Corresponding author.

Problem & Motivation



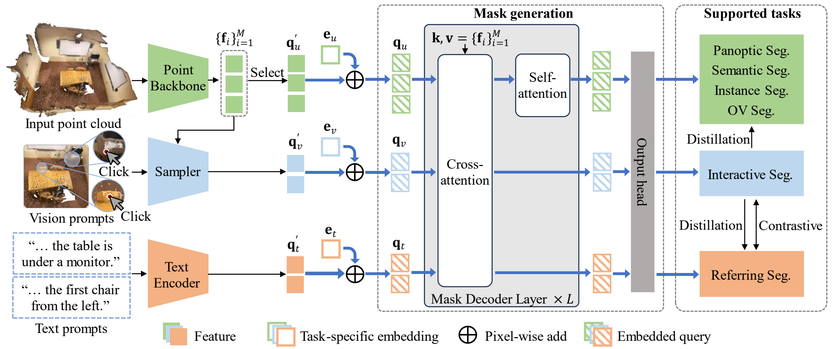
- Approaches limit their 3D scene understanding to task-specific perspectives.
- Unifying multiple tasks within a single model can reduce computation consumption and benefit real-world applications.
- Simply integrating separated methods into a single architecture faces challenges balancing the customized optimizations.

Contribution



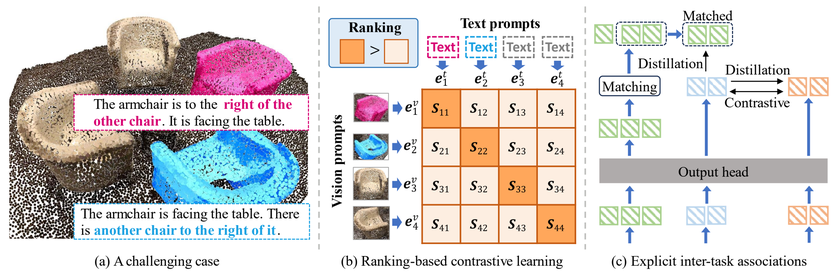
- UniSeg3D achieves six 3D segmentation tasks in one inference by a single model.
- UniSeg3D is a flexible and efficient framework, which can be easily extended to more tasks.

Methodology



Overview: UniSeg3D mainly consists of three modules: a point cloud backbone, prompt encoders, and a mask decoder.

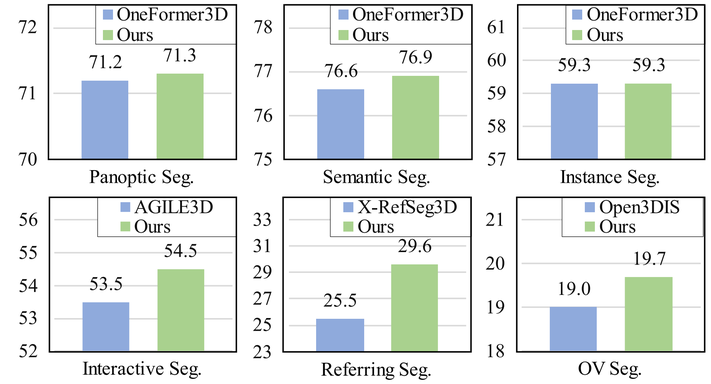
- **Query representation:** We use queries to unify representations of the input information.
- **Task-specific embedding:** Add task-specific embedding to mask decoder for digging task-specific information.
- **Shared mask decoder and output head:** We use the same mask decoder and output head for all tasks without task-specific specialized modules.



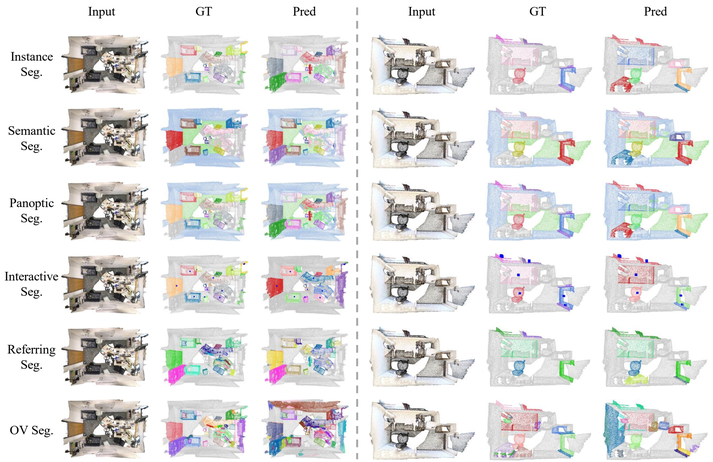
(a) A challenging case (b) Ranking-based contrastive learning (c) Explicit inter-task associations

We establish inter-task associations to facilitate more comprehensive 3D scene understanding: (a) A challenging case requiring the distinction of textual positional information within the expressions. (b) A contrastive learning matrix for vision-text pairs, where a ranking rule is employed to suppress incorrect pairings. (c) Knowledge distillation across multi-task predictions.

Experiments



Our method contains no task-customized modules, while consistently outperforming specialized SOTA solutions, demonstrating a desirable potential to be a solid unified baseline.



Visualizations of multi-task segmentation results, indicating effectiveness in a subjective way.

References

[1] OneFormer3D: One Transformer for Unified Point Cloud Segmentation. CVPR 24.

[2] AGILE3D: Attention Guided Interactive Multi-object 3D Segmentation. ICLR 24.

[3] X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks. AAAI 24.

[4] Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance. CVPR 24.