Pattern Recognition 179 (2026) 113678

Contents lists available at ScienceDirect

Pattern Recognition

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

UNIT3D: Unified Instance-relative Transformer for indoor 3D object detection and segmentation

Xinrun Liu [a], Weikun Liu [b], Bin Fan [a], Hongmin Liu [a,*]

[a] The School of Artificial Intelligence, University of Science and Technology Beijing, Beijing, 100083, China
[b] The School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, 100083, China

ARTICLE INFO

Keywords:
Unified model
3D object detection
3D segmentation
Indoor scene understanding

ABSTRACT

In this paper, we present UNIT3D, the first fully unified 3D scene understanding framework that bridges this gap by integrating object detection, semantic, instance, and panoptic segmentation tasks within a single model. We first examine the limits of current designs by adding task-specific heads to strong baselines, finding that simple multi-task extensions perform poorly on added tasks and even degrade the original ones. This reveals a fundamental architectural conflict between the geometric precision required for detection and the grouping adaptivity needed for segmentation. To address this, we propose the Unified Instance-relative Transformer, which replaces task-specific components with a shared, conflict-free query interaction mechanism. We discard the restrictive mask attention used in prior work and introduce an instance-relative position encoding that retains the benefits of soft mask attention while restoring the rigorous spatial cues necessary for box regression. Our decoder incorporates Spatial Aware Self Attention to encode instance centers for scene-level context and Vertex Guided Cross Attention to encode instance vertices for fine-grained details. This global-to-local design allows UNIT3D to directly output masks and bounding boxes in a fully end-to-end manner. Crucially, this strict one-to-one matching strategy eliminates the reliance on Non-Maximum Suppression (NMS), avoiding heuristic post-processing errors and ensuring robust detection even in crowded 3D scenes. Experiments on ScanNet, ScanNet200, and S3DIS demonstrate that UNIT3D achieves state-of-the-art results in 3D object detection while remaining competitive on segmentation tasks, proving that cross-paradigm unification can effectively facilitate mutual enhancement across domains. Code is available at github.com/liuxinrun/unit3d.

1. Introduction

The perception and comprehension of indoor 3D scenes are foundational technologies for next-generation applications such as autonomous robotics, augmented reality (AR), and intelligent environments. Among the myriad of perception tasks, 3D object detection (localizing bounding boxes) and 3D segmentation (assigning a semantic or instance label to every point) represent two fundamental axes of scene understanding. Historically, these tasks have been addressed by disparate, specialized model architectures [1,2]. This "one-model-per-task" paradigm not only introduces redundancy in research and development but, more critically, it artificially severs the intrinsic geometric and semantic correlations between detection and segmentation. This partition impedes the development of a holistic and comprehensive scene representation. While recent advances have produced powerful unified segmentation frameworks like OneFormer3D [3] and UniSeg3D [4], which adeptly consolidate multiple segmentation subtasks, a central limitation persists: these state-of-the-art (SOTA) frameworks remain confined to the domain of segmentation. They have not successfully integrated the equally vital task of 3D object detection, thereby limiting their utility in applications that demand complete object-level understanding.

In 3D point cloud perception, conventional approaches typically employ task-specific models optimized for individual objectives. Current 3D unified segmentation frameworks, exemplified by OneFormer3D [3], UniSeg3D [4], and Uni3DL [5], are predominantly built upon a **Mask Attention** mechanism. Consequently, the scope of these methods is strictly confined to unification within the segmentation domain. In contrast, contemporary 3D detection methods completely avoid mask attention, as demonstrated by V-DETR [6], UniDet3D [7], and SPGroup3D [1]. This architectural divergence creates a fundamental barrier: while detection and segmentation tasks can mutually benefit each other within a unified Transformer-based architecture, existing approaches cannot be directly extended across tasks due to their fundamentally different design principles.

* Corresponding author.
E-mail addresses: D202110362@xs.ustb.edu.cn (X. Liu), u202242348@xs.ustb.edu.cn (W. Liu), bin.fan@ieee.org (B. Fan), hmliu_82@163.com (H. Liu).

https://doi.org/10.1016/j.patcog.2026.113678
Received 20 November 2025; Received in revised form 15 February 2026; Accepted 3 April 2026
Available online 6 April 2026

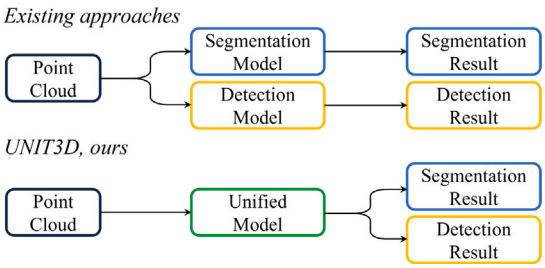


Fig. 1. Traditional 3D point cloud detection and segmentation approaches employ distinct task-specific models to achieve best performance. We propose UNIT3D, a unified 3D framework that tackles detection, semantic, instance, and panoptic segmentation tasks with a multi-task train-once design.

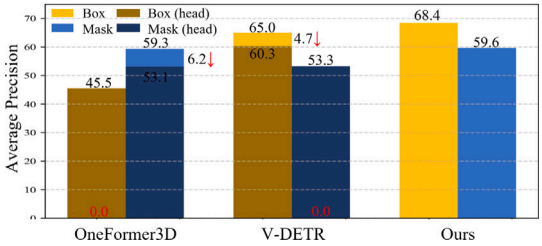


Fig. 2. Performance degradation when adding task-specific heads to existing models. Our method outperforms current SOTA approaches using a single unified architecture.

Why is bridging this gap critical? We argue that unifying detection and segmentation is essential for four decisive reasons. First, regarding **efficiency**, real-world agents require both bounding boxes for planning and masks for interaction; a unified model halves the inference overhead compared to maintaining separate pipelines. Second, regarding **synergy**, geometric constraints from detection act as strong structural priors that regularize segmentation, specifically improving the understanding of countable objects ("Things") that are critical for interaction. Third, for **universality**, resolving the conflict between regression and classification within a single query mechanism is a prerequisite for developing general-purpose 3D perception systems. Finally, regarding **performance comparability**, we empirically demonstrate that under fair training conditions, unifying these tasks does not degrade segmentation. On the contrary, detection and segmentation are mutually promoting, yielding highly competitive and balanced performance across all domains.

This challenge contrasts with the relative success of unified approaches in 2D vision, where methods like Transformer-based architectures such as Mask DINO [8] have successfully unified detection and segmentation tasks. However, direct translation of these 2D methods to 3D is non-trivial due to the sparse, unstructured nature of 3D point clouds, necessitating fundamentally different architectural considerations.

To validate this architectural limitation, we conducted systematic experiments by adding task-specific heads to existing SOTA models. As shown in Fig. 2, simply adding a detection head to the segmentation-centric OneFormer3D (which relies on mask attention) or incorporating a segmentation head into the detection-centric V-DETR (which relies on global positional encoding) yields suboptimal performance on the additional tasks. More critically, such naive multi-task extensions even degrade the performance of the original tasks. These failures reveal a fundamental architectural conflict: Mask attention excels at grouping irregular points for segmentation but dilutes the precise spatial structural information required for bounding box regression; conversely, strong global positional encodings facilitate box localization but lack the local adaptivity for fine-grained mask delineation. This dilemma prompts two questions: *(1) Why do existing unified architectures fail to reconcile spatial structure with instance adaptivity?* and *(2) Can we design a unified representation that transcends these task-specific trade-offs?*

3D detection is a region-level regression task demanding global contextual reasoning and precise geometric centers, whereas 3D segmentation is a point-level classification task emphasizing fine-grained local details. OneFormer3D's reliance on mask attention precludes its extension to detection, preventing it from being a fully unified framework despite its success in unifying various segmentation tasks (semantic, instance, panoptic). To bridge this gap, as illustrated in Fig. 1, we propose **UNIT3D** (Unified Instance-relative Transformer).

Unlike prior works that rely on disjoint branches or are limited to a single domain, UNIT3D breaks the boundaries of 3D perception by adhering to three core Design Principles: **Conflict-free Query Interaction:** We move beyond disjoint multi-head designs by employing a shared query mechanism. This architecture effectively resolves the feature conflicts inherent in simultaneous detection and segmentation, allowing the two tasks to mutually enhance — rather than impede — each other within a unified representation. **Instance-Relative Geometry:** To restore the rigorous spatial cues necessary for detection without losing segmentation flexibility, we replace restrictive, spatially-agnostic mask attention with a novel Instance-Relative Position Encoding. By prioritizing geometric fidelity via SASA and VGCA, we bridge the gap between the rigid structural requirements of box regression and the grouping adaptivity of mask prediction. **Global-to-Local Attention:** Our architecture follows a hierarchical perception flow that enables a fully end-to-end, NMS-free pipeline. It first establishes scene-level understanding through instance centers (Global) and then explicitly refines fine-grained details through instance vertices (Local). This ensures robust, direct output of both high-precision bounding boxes and fine-grained masks across varying granularities.

By integrating these principles, UNIT3D establishes a new standard for unified 3D scene understanding, moving from fragmented task-specific designs to a holistic perception interface.

Guided by these principles, the primary contributions of this work are summarized as follows:

- We propose UNIT3D, the first truly unified framework that breaks the boundary between 3D object detection and segmentation. Unlike prior segmentation-only unified models (e.g., OneFormer3D), UNIT3D concurrently handles precise box regression and mask prediction within a single Transformer architecture.
- To realize this unification, we implement the design principles via two novel components: **Spatial Aware Self Attention** (for center-based global context) and **Vertex Guided Cross Attention** (for vertex-based local refinement). This implementation effectively resolves the architectural incompatibilities between tasks.
- Compared with single-task models, UNIT3D achieves comparable inference time while executing multiple tasks concurrently, and significantly simplifies the pipeline by removing the reliance on NMS.
- Extensive experiments demonstrate that our method achieves SOTA performance on 3D object detection while maintaining competitive results on segmentation, validating the effectiveness of our unified design.

The remainder of this paper is organized as follows: Section 2 describes recent studies related to our work. Section 3 introduces the crucial components of our UNIT3D. Extensive experimental results on various datasets are reported in Section 4, and the conclusion of this paper is in Section 5.

2. Related work

In this section, we briefly review three relevant literature in the fields of indoor 3D object detection, 3D segmentation, and unified perception frameworks. We examine the evolution from task-specific architectures to integrated systems that seek to provide a holistic understanding of 3D environments.

2.1. 3D object detection

3D object detection aims to localize and classify instances within a scene using oriented bounding boxes. We first discuss point-based 3D object detection approaches, followed by voxel-based methods, and finally, transformer-based methods, all of which are capable of directly processing point clouds.

2.1.1. Point-based methods

Point-based detectors operate directly on raw points using PointNet++ [9] to learn local geometric cues. VoteNet [10] casts detection as deep Hough voting, where seed points predict offsets toward object centers and aggregate local features to form proposals. To stabilize vote clustering and reduce category confusion, MLCVNet [11] inject semantic priors and multi-level context. To improve initial proposals and orientations, H3DNet [12] uses hybrid and overcomplete geometric primitives. Voting is further refined by BRNet [13], which back-traces from voted centers to retrieve representative seeds and refine surface cues. RepSurf-U [14] designs surface-oriented representations to capture very local geometry and thin structures.

Typical pipelines include farthest point sampling (FPS), local neighborhood grouping, per-point MLPs or point convolutions, vote generation, vote clustering, and a refinement head for box parameterization. These methods are strong at modeling fine geometry and small or thin objects, but face sampling overhead, the quadratic complexity of FPS, sensitivity to sparse/noisy inputs, and weaker global context modeling than grid- or transformer-based backbones.

2.1.2. Voxel-based 3D object detection

To avoid the quadratic cost of FPS [9], voxelization converts irregular points into sparse grids for efficient sparse 3D convolutions. Early dense-voxel CNNs waste compute on empty space. Replacing dense operators with sparse convolutions in an encoder–decoder topology markedly improves scalability and robustness to sparsity [15]. Building on sparse backbones, FCAF3D [16] removes anchors and proposes a data-driven oriented box parameterization, achieving a strong accuracy-speed trade-off. TR3D [17] simplifies the architecture and heads for real-time indoor detection. CAGroup3D [18] performs class-aware local grouping on object-surface voxels and adds a two-stage refinement for better localization. DSPDet3D [19] focuses compute via dynamic spatial pruning to improve recall for small or occluded objects. Further, optimized label assignment [20] reduces matching ambiguity, and open-vocabulary, two-stage designs extend to open sets [21].

Key design choices include voxel size (accuracy-efficiency trade-off), submanifold sparse convolutions to preserve features, hierarchical down/up-sampling with skip connections, and oriented box regression heads. Sparse-convolution pipelines scale well with active voxels and encode global context, but may lose fine details due to quantization and require careful voxelization and hyperparameter tuning.

2.1.3. Transformer-based 3D object detection

Most systems use a sparse 3D backbone to extract scene features, initialize a small set of object queries (learnable or point-anchored), and iteratively refine box parameters with a transformer decoder supervised by bipartite matching. GroupFree3D [22] couples transformer encoder–decoder but still relies on Non-Maximum Suppression (NMS), while 3DETR [23] adopts bipartite matching to remove NMS and shows competitive results with minimal 3D-specific bias. To encode spatial relations, V-DETR [6] introduces vertex-relative position encoding to bias attention toward object vertices or centers, improving AP and convergence. Objformer [24] introduces instance-wise interaction modules to explicitly model the relationships between 3D instances, HRNet [25] and LGNet [26] enhance point cloud representations via multi-scale feature aggregation. For queries, Uni3DETR [27] mixes learnable and data-driven query points to fuse local and global cues across scenes. Efficiency-oriented models based on state-space decoders [28] target linear-time updates of scene and query features to improve late-layer gains and runtime.

These methods offer unified set prediction without hand-crafted anchors or proposals, strong global reasoning, and clean objectives. Challenges remain in query initialization, positional encoding for 3D geometry, data efficiency, and memory/time cost at high resolution.

2.2. 3D segmentation

While detection provides coarse geometric localization through bounding boxes, 3D segmentation tasks seek a more granular, per-point understanding of the scene. In the following, we discuss the current state of 3D segmentation tasks, including semantic segmentation, instance segmentation, and panoptic segmentation.

2.2.1. Semantic segmentation

Semantic segmentation operates on points or voxels. Point-based pipelines learn per-point features via neighborhood aggregation (PointNet++ [9]) and transformer blocks (Point Transformer [29], Point-BERT-style designs [30]) to capture local-to-global context, often with multi-scale grouping, residual MLPs, and attention. Voxel-based pipelines convert points into regular or sparse grids and apply dense or sparse 3D CNNs [31], commonly using U-Net-like designs; sparse convolutional U-Nets are efficient and strong on large indoor scenes. Typical trade-offs include point precision vs. voxel quantization, memory vs. resolution, and pretraining vs. training from scratch.

2.2.2. Instance segmentation

Top-down (proposal-based) methods detect 3D regions or boxes first and then predict per-region masks, leveraging box priors and ROI-aligned local features, as in GSPN [32], and other proposal-based designs [33]. These approaches benefit from detection signals but can degrade with imperfect proposals and under heavy occlusion.

Bottom-up (grouping-based) methods learn per-point embeddings, center offsets, or energy fields, followed by clustering and refinement, as in DyCo3D [34], SSTNet [35], and SoftGroup [36]. Techniques include dynamic convolutions for instance-aware feature aggregation, semantic superpoint trees, and soft grouping with confidence-driven refinement. These methods handle shape diversity well but can be sensitive to radii and thresholds.

Transformer-based single-shot methods decode a fixed set of instance queries over sparse scene features to directly predict masks and classes without proposals or clustering. Mask3D [2] uses mask attention over voxel features, Query Refinement Transformer [37] improves query quality over iterations, and SPFormer [38] pools points into superpoints and uses a query decoder with superpoint cross-attention and bipartite matching on superpoint masks to train end-to-end without intermediate aggregation or NMS, reaching high accuracy and fast inference.

2.2.3. Panoptic segmentation

3D panoptic segmentation remains underexplored. Many methods lift 2D panoptic predictions into 3D or fuse multi-view RGB masks with 3D geometry, then aggregate and reconcile labels across views [39–42]. These pipelines are often validated on ScanNet and depend on multi-view RGB, which can add complexity, introduce projection errors, and lose information. Recent fully 3D designs seek to learn panoptic masks directly on point clouds.

2.3. Unified perception frameworks

Recent advancements in specialized detection and segmentation models have naturally sparked interest in developing integrated systems. In the 2D domain, numerous unified perception frameworks

have emerged to consolidate diverse tasks into a single architecture. However, in the 3D domain, such unified frameworks remain scarce.

To reduce task-specific pipelines and enable knowledge sharing, unified frameworks have emerged [43]. In 2D, Mask DINO [8] unifies detection and segmentation with a transformer and mask classification, using denoising and anchor refinement. Following this trend, Omg-seg [44] explores the limits of multi-task generalization by supporting over ten distinct segmentation tasks (including open-vocabulary settings) with a single model. In the video domain, Tube-link [45] establishes a universal video segmentation framework, introducing a flexible cross-tube mechanism to link temporal tubes for consistent tracking and segmentation.

In 3D segmentation, OneFormer3D [3] uses a shared transformer decoder with task-conditioned queries to jointly predict semantic, instance, and panoptic outputs, and UniSeg3D [4] generalizes this to six tasks by treating inputs as queries, sharing a mask decoder and heads, and transferring knowledge via distillation and contrastive learning. Beyond vision-only modeling, Uni3DL [5] unifies 3D vision and language: a query transformer lets latent and text queries attend to 3D features to produce semantic and mask outputs, and a task router composes heads for object classification, mask prediction, grounding, captioning, and text-3D retrieval in a single architecture that operates directly on point clouds. PUPS [46] presents a unified point cloud panoptic segmentation approach, employing a dynamic convolution mechanism to merge semantic and instance branches efficiently. Bridge3D [47] significantly enhances the quality of 3D scene representation learning by leveraging text and 2D features for high-precision self-supervised learning, effectively bridging the gap between 3D, 2D, and textual object-level features. Similarly, MTP [48] utilizes powerful visual foundation models to align 3D Transformer structures with region-level information, thereby improving the effectiveness of knowledge distillation for self-supervised 3D scene understanding. Both Bridge3D and MTP operate as pre-training frameworks; during the fine-tuning stage, the decoders used for pre-training are discarded, and task-specific decoders are introduced to adapt the models to various downstream tasks.

Despite these advances, a gap remains in effectively unifying high-precision 3D detection with segmentation. 2D unified methods (like Mask DINO) rely on dense priors not applicable to sparse 3D data. While pre-training frameworks such as Bridge3D and MTP improve representation learning, they still rely on task-specific decoders during fine-tuning, failing to achieve a truly unified inference architecture. Crucially, 3D unified frameworks like OneFormer3D focus primarily on segmentation tasks (semantic, instance, panoptic). They struggle to incorporate detection because their core mechanism (Mask Attention) is spatially agnostic, creating a trade-off where one must choose between segmentation quality (via masks) or detection accuracy (via boxes). Our work addresses this by replacing task-specific attention with a generalized instance-relative encoding that serves both tasks simultaneously without compromise. To the best of our knowledge, UNIT3D is the first framework to fully unify 3D detection and segmentation tasks within a single architecture, successfully resolving the fundamental conflict between geometric precision and grouping adaptivity.

3. Approach

In this section, we first formulate the problem of 3D perception and describe our main idea. Subsequently, we detail the UNIT3D framework for unified indoor 3D perception task.

3.1. Overview

Problem Definition. To unify 3D detection and segmentation within a single framework, we define a joint output space that maps unstructured point clouds to consistent representations. Let an input 3D scene be represented by a point cloud $\mathcal{P} = \{p_1, \dots, p_N\}$, where each point $p_i \in \mathbb{R}^d$ contains 3D coordinates (x, y, z) and optional attributes such as color and normals ($d \geq 3$). Our unified perception task aims to jointly predict four interrelated outputs:

3D Object Detection: Predict K object predictions $\mathcal{B} = \{b_1, \dots, b_K\}$, where each $b_k = (c_k, \beta_k)$ comprises a class label c_k and 3D bounding box β_k.

Semantic Segmentation: Generate per-point semantic labels $\mathcal{S} = \{s_1, \dots, s_N\}$, where $s_i \in C$ represents the semantic class of point p_i.

Instance Segmentation: Output instance masks $\mathcal{I} = \{m_1, \dots, m_K\}$ corresponding to detected objects.

Panoptic Segmentation: Produce per-point panoptic labels $\mathcal{Q} = \{q_1, \dots, q_N\}$, where each $q_i = (s_i, z_i)$ contains a semantic label s_i and instance identifier z_i.

Objective: Develop a unified framework that generates these four outputs in a single forward pass, leveraging shared representations to ensure consistency and computational efficiency.

Our approach builds upon the OneFormer3D [3] framework with strategic modifications to enable unified 3D object detection and segmentation. The core innovation lies in our **instance relative position encoding** mechanism, which replaces the restrictive mask attention with two complementary attention components. As illustrated in Fig. 3, our framework maintains the original OneFormer3D architecture (gray components) while introducing our novel attention mechanisms (blue components) and a dedicated detection branch.

The key insight of our design is the recognition that detection and segmentation tasks require different levels of spatial understanding. Detection benefits from global scene-level context (instance center positions), while segmentation requires fine-grained local details (relative instance positions). Our instance relative position encoding addresses these needs through a dual-stage approach, enabling seamless task unification without architectural conflicts.

3.2. Spatial aware self attention

To overcome the limitation of traditional self-attention — which relies solely on feature similarity and lacks the explicit geometric anchors necessary to distinguish between spatially distant but visually similar instances — we propose Spatial Aware Self Attention (SASA). Traditional self-attention mechanisms in Transformers compute attention weights based solely on feature similarity:

$$Att = \text{softmax}(q \cdot k^T) \cdot v \tag{1}$$

where q, k, and v denote the query, key, and value, respectively. In this scheme, spatial relationships are implicitly learned from the data, which is suboptimal for 3D scene understanding, where explicit geometric relationships are crucial for instance-level reasoning.

Our SASA addresses this limitation by explicitly incorporating the **center positions** of instances within the global scene context. Unlike traditional approaches, SASA integrates spatial awareness directly into the attention computation, enhancing the model's scene-level understanding essential for both 3D detection and segmentation tasks.

Using queries derived through SuperPoint pooling with superpoint centers as reference points, we compute the relative position vector $r = c_{box} - c_{sp}$ between iteratively updated 3D bounding box center points c_{box} and the current query reference points c_{sp}. The spatial positional encoding is formulated as:

$$p = \rho(\varphi(r)), \tag{2}$$

where φ represents a transformation function (such as Sinusoidal) for dimension alignment, and ρ denotes a MLP that introduces learnable parameters for adaptive spatial representation. Our enhanced self-attention then becomes:

$$SASA = \text{softmax}((q + p) \cdot (k + p)^T) \cdot v \tag{3}$$

This design enables each query to be spatially aware of its center position within the scene, thereby providing the global context necessary for accurate object detection while maintaining the spatial coherence essential for segmentation tasks.

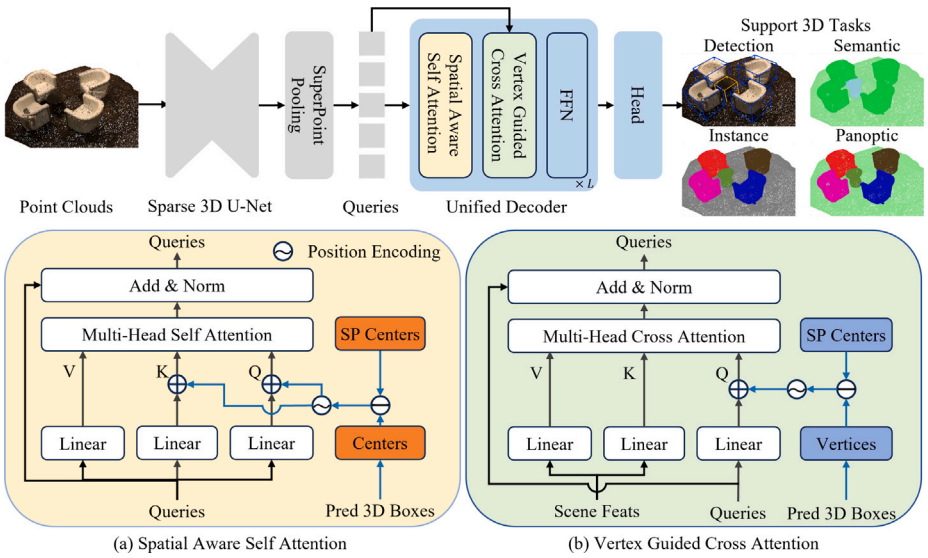


Fig. 3. Overview of our unified perception framework. The model processes the input point cloud through a shared Sparse 3D U-Net, followed by a SuperPoint Pooling and a Instance relative position encoding mechanism that enriches queries with geometric information for both detection and segmentation tasks. The outputs include 3D bounding boxes, semantic, instance, and panoptic mask.

3.3. Vertex guided cross attention

While SASA provides a global sense of instance locations, it lacks the fine-grained geometric detail required to precisely delineate object boundaries. To resolve the "spatial agnosticism" of standard mask attention, which fails to capture the localized vertex-level cues necessary for accurate box regression and mask refinement, we introduce Vertex Guided Cross Attention (VGCA). Traditional cross-attention mechanisms face a fundamental limitation when attempting to unify detection and segmentation:

$$Att_d = \text{softmax}((q+p)\cdot k^T)\cdot v \tag{4}$$

$$Att_s = \text{softmax}(q\cdot k^T\cdot \text{mask})\cdot v \tag{5}$$

Detection methods (Att_d) rely on positional encoding (p) for geometric awareness, while segmentation methods (Att_s) utilize hard mask attention for foreground-background distinction. These architectural differences prevent direct task unification.

Our VGCA overcomes this limitation by introducing a **soft mask attention** mechanism that encodes the **relative positions** of instance vertices. This design provides the fine-grained spatial guidance required for segmentation while maintaining the geometric flexibility essential for detection, effectively bridging the architectural gap between the two tasks.

VGCA transforms 3D detection boxes $B \in \mathbb{R}^7$ into vertex-based positional encoding. Given eight vertex coordinates v_i, where $i \in \{1,2,\dots,8\}$, derived from box B, we compute relative position vectors between each vertex and the current query point c_{sp}:

$$\begin{aligned} p_{vg} &= \rho\left(\varphi\left([c_{sp}-v_1,\dots,c_{sp}-v_8]\right)\right) \\ VGCA &= \text{softmax}((q+p_{vg})\cdot k^T)\cdot v \end{aligned} \tag{6}$$

where $[\cdot]$ denotes concatenation. Unlike the previous hard mask attention, which distinguishes foreground from background in a binary manner, our vertex-guided position encoding (p_{vg}) provides **soft spatial guidance** that naturally adapts to both detection (through geometric box constraints) and segmentation (through fine-grained vertex relationships). This soft attention mechanism eliminates the need for task-specific attention designs and enables mutual enhancement between tasks.

Global-Local Instance Understanding The synergy between SASA and VGCA realizes our unified instance-relative transformer through a global-to-local perception hierarchy:

(1) Global Stage (SASA): Encodes instance center positions for scene-level context, enabling the model to understand object relationships and spatial layout, which are crucial for detection.

(2) Local Stage (VGCA): Encodes relative instance positions through vertex relationships, providing fine-grained spatial guidance essential for precise segmentation.

This dual-stage design allows our unified framework to simultaneously predict detection bounding boxes (benefiting from global context) and segmentation masks (supported by local vertex guidance). The soft attention mechanism facilitates knowledge transfer between tasks. Beyond the specific implementation of UNIT3D, we posit that this global-to-local instance modeling serves as a generalizable design principle for 3D scene understanding. By explicitly decoupling scene-level structural reasoning from instance-level local refinement, this hierarchical paradigm offers a scalable foundation that can inspire future unified architectures to reconcile the inherent conflicts in multi-granularity perception tasks.

3.4. Training

End-to-end training of our Transformer-based method requires three essential components: a cost function to measure similarity between predictions and ground truth objects, a matching strategy for assignment, and a loss function for optimization.

Cost Function. To support both detection and segmentation within our unified framework, we extend the OneFormer3D [3] cost function by incorporating a bounding box regression term:

$$C_{ik} = -\beta_c\cdot p_{i,c_k} + \beta_m\cdot C_{ik}^{\text{mask}} + \beta_b\cdot \text{MPDIoU}(b_i, b_k^{\text{gt}}) \tag{7}$$

where p_{i,c_k} is the predicted classification probability of the ith proposal for ground truth class c_k. The terms b_i and b_k^{gt} denote predicted and ground truth bounding boxes. The mask matching cost C_{ik}^{mask} combines binary cross-entropy and Dice loss with Laplace smoothing. Bounding box similarity is measured using MPDIoU loss. We set $\beta_c = 0.5$, $\beta_m = 1$, and $\beta_b = 2$.

Matching Strategy. Unlike methods using computationally intensive Hungarian matching, we adopt the efficient disentangled matching

from OneFormer3D [3]. Each query is initialized from superpoint features, creating explicit correspondence between queries and superpoints. Since each superpoint belongs to a single ground truth object, this establishes clear mapping between superpoints, queries, and ground truth instances.

To address the inconsistency between segmentation ($k = 1$ positive match) and detection ($k > 1$ positive matches), we enforce consistent one-to-one matching where only the lowest-cost proposal is selected as the positive sample. The matching criterion is:

$$\hat{C}_{ik} = \begin{cases} C_{ik} & \text{if the } i\text{-th superpoint} \in k\text{-th object} \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

Loss Function. The comprehensive loss function combines multiple task objectives:

$$\mathcal{L} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \mathcal{L}_{bce} + \mathcal{L}_{dice} + \mathcal{L}_{sem} + \mathcal{L}_{det} \tag{9}$$

where $\mathcal{L}_{cls}$ is the cross-entropy classification loss with $\lambda_{cls} = 0.5$, $\mathcal{L}_{bce}$ and $\mathcal{L}_{dice}$ are the mask losses, $\mathcal{L}_{sem}$ is the semantic segmentation loss using binary cross-entropy, and $\mathcal{L}_{det}$ is the MPDIoU loss for bounding box regression.

4. Experiment

4.1. Experimental setup

4.1.1. Datasets

We evaluted the proposed method on ScanNet [49], ScanNet200 [50] and S3DIS [51].

- ScanNet: This benchmark comprises 1613 reconstructed indoor scenes, with 1201 scenes for training, 312 for validation, and 100 for testing; it annotates 18 object categories with axis-aligned 3D bounding boxes. Following common practice [20], we report mean Average Precision at IoU thresholds of 0.25 and 0.5 (AP25/AP50) for evaluation in 18 categories with large size variance.
- ScanNet200: Built upon ScanNet, this benchmark expands the label space from 18 to about 200 categories (198 instance-level and 2 stuff classes) using the same source data. We follow the ScanNet200 benchmark [50], with 1201 training scenes and 312 validation scenes.
- S3DIS: This dataset [51] spans six large-scale areas with 272 scenes and provides annotations for 13 semantic categories; we adopt the standard protocol that evaluates on Area-5 while training on the remaining areas.

4.1.2. Metric

For 3D object detection, we report the mean Average Precision (mAP) at IoU thresholds of 0.25 and 0.5, following the standard evaluation protocol. For semantic segmentation, we use the mean Intersection over Union (mIoU) metric. For instance segmentation, we report the Average Precision (AP) for each category which is an average of scores obtained with IoU thresholds set from 0.5 to 0.95, with a step size of 0.05. For panoptic segmentation, we use the Panoptic Quality (PQ) metric [52]; we also report PQ_{th} and PQ_{st}, estimated for thing and stuff categories, respectively.

4.1.3. Implementation details

We implement our model based on the OneFormer3D [3] framework and MMDetection3D [53]. The model is trained on two NVIDIA 3090 GPU with a batch size of 8. We use the AdamW optimizer with an initial learning rate of 0.0002, weight decay of 0.05, and polynomial scheduler with a base of 0.9 for 1024 epochs. We apply the standard augmentations: horizontal flipping, random rotations around the z-axis, and random scaling. During the training, we assign a ground truth object to the nearest superpoints. No test-time augmentations are applied during the inference time. We follow OneFormer3D report the best results at the validation set.

4.2. Quantitative evaluation

4.2.1. Evaluation on the ScanNet dataset

On the ScanNet dataset, we list the quantitative results of UNIT3D and other 25 methods in Table 1. On the ScanNet validation set, our unified model delivers strong and balanced performance across detection, semantic/instance/panoptic segmentation metrics. For detection, it reaches 78.4 mAP@25 and 67.4 mAP@50, surpassing DETR-based baselines such as UniDet3D (77.5/66.1), especially at the stricter IoU=0.50 threshold, which reflects more accurate localization. Semantic quality is also high with 75.7 mIoU, indicating robust scene understanding that supports instance and panoptic heads. Instance AP attains 58.1, on par with top transformer-based mask decoders, while panoptic quality reaches 71.7 PQ, evidencing good alignment between thing and stuff predictions. Importantly, augmenting the inputs with surface normals consistently improves all tasks: the surface-normal variant attains 79.4/68.4 in detection, 76.9 mIoU in semantic segmentation, 59.6 in instance AP, and 72.9 PQ in panoptic segmentation. These gains confirm that the unified architecture scales well in both accuracy and consistency, and that simple geometric cues such as normals translate into tangible benefits without sacrificing throughput.

4.2.2. Evaluation on the ScanNet200 dataset

Table 2 lists the quantitative results of UNIT3D and other methods on the ScanNet200 dataset. ScanNet200 greatly increases category granularity and includes many small or fine-grained classes, making detection and segmentation more challenging. Under this setting, our model remains robust and outperforms strong baselines. For detection, it achieves 36.6 mAP@25 and 30.4 mAP@50, for semantic segmentation reaches 30.4 mIoU, showing stable discrimination over a long-tailed label space, and instance AP is 30.6, surpassing several dedicated instance systems. Panoptic performance is also competitive, with 29.0 PQ, reflecting consistent improvements when scaling to many categories and small objects. As on ScanNet, adding surface normals further increases performance to 38.9/31.7 for detection and 30.3 PQ for panoptic segmentation. Overall, the results suggest that the unified training objective and shared representation enable effective generalization to larger ontologies and object scale variance, while the normal-augmented variant offers additional robustness in challenging conditions.

Table 2 reports the results on the ScanNet200 dataset, which features a challenging long-tailed distribution with fine-grained categories. Our method achieves state-of-the-art performance in detection with 38.9 mAP@25 and 31.7 mAP@50, significantly outperforming specialist detectors like DLLA and TR3D. Regarding segmentation, we observe that OneFormer3D marginally outperforms our method (e.g., 31.2 PQ vs. 30.3 PQ). This slight gap is expected: OneFormer3D employs pure mask attention, which allows for highly flexible feature grouping ideal for irregular segmentation boundaries but lacks the geometric rigidity required for box regression. In contrast, UNIT3D incorporates explicit spatial constraints (instance-relative position encoding) to enable high-precision detection. Despite the minor trade-off ($< 1\%$ drop in PQ), UNIT3D offers a critical advantage: it is the only framework that provides simultaneous high-precision detection and segmentation. While OneFormer3D fails to predict bounding boxes entirely, UNIT3D unlocks comprehensive scene understanding capabilities with a single, simplified architecture, making it a more versatile choice for real-world robotic applications that require both object localization and pixel-wise parsing.

4.2.3. Evaluation on the S3DIS dataset

Table 3 presents the comparison on the S3DIS Area-5. Our model demonstrates a commanding lead in detection, achieving 77.8 mAP@25 and 65.3 mAP@50, surpassing the previous transformer-based detector UniDet3D by large margins (+2.6 mAP@25, +4.5 mAP@50). In terms of segmentation, while OneFormer3D exhibits higher overall metrics

Table 1
Comparison of our UNIT3D with SOTA methods on ScanNet Val. The best results are highlighted in **bold**, and the second-best results are underscored. "*" denotes using surface normal.

Methods	Presented at	Detection		Semantic	Instance	Panoptic		
		mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
FCAF3D [16]	ECCV'22	71.5	57.3					
TR3D [17]	ICIP'23	72.9	59.3					
Uni3DETR [27]	NIPS'24	71.7	58.3					
SPGroup3D [1]	AAAI'24	74.3	59.6					
DLLA [20]	TCSVT'25	73.8	60.2					
CAGroup3D [18]	NIPS'22	75.1	61.3					
V-DETR [6]	ICLR'24	77.4	65.0					
UniDet3D [7]	AAAI'25	77.5	66.1					
DEST [28]	ICLR'25	78.5	66.6					
PointTransformer [29]	ICCV'21			70.6				
PointMetaBase [54]	CVPR'23			72.8				
PointTransformerV2 [30]	NIPS'22			75.4				
Mask3D [2]	ICRA'23				55.2			
SPFormer [38]	AAAI'23				56.3			
QueryFormer [37]	ICCV'23				56.5			
MAFT [55]	ICCV'23				57.8			
MAFT* [55]	ICCV'23				**59.6**			
SceneGraphFusion [39]	CVPR'21					31.5	30.2	43.4
TUPPer-Map [40]	IROS'21					50.2	47.8	71.5
Panoptic Lifting [41]	CVPR'23					58.9		
PanopticNDT [42]	IROS'23					59.2		
OneFormer3D [3]	CVPR'24			76.6	59.3	71.2	69.6	86.1
UniSeg3D [4]	NIPS'24			76.3	59.1	71.3	69.0	86.2
Ours		78.4	67.4	75.7	58.1	71.7	70.1	85.9
Ours*		**79.4**	**68.4**	**76.9**	**59.6**	**72.9**	**71.4**	**86.4**

Table 2
Comparison of our UNIT3D with SOTA methods on ScanNet200 Val. The best results are highlighted in **bold**, and the second-best results are underscored. This FCAF3D and TR3D results are reported by DLLA. "*" denotes using surface normal.

Methods	Presented at	Detection		Semantic	Instance	Panoptic		
		mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
TR3D [17]	ICIP'23	29.2	21.3					
FCAF3D [16]	ECCV'22	30.4	22.9					
DLLA [20]	TCSVT'25	32.0	23.9					
MinkUNet [31]	CVPR'19			25.0				
MinkUNet + LGround [50]	ECCV'22			28.9				
TD3D [33]	WACV'24				23.1			
Mask3D [2]	ICRA'23				27.4			
OneFormer3D [3]	CVPR'24			**30.1**	**30.6**	**31.2**	**30.7**	**78.6**
Ours		36.6	30.4	28.1	28.7	29.0	28.5	77.0
Ours*		**38.9**	**31.7**	29.3	30.4	30.3	29.8	75.7

Table 3
Comparison of our UNIT3D with SOTA methods on S3DIS Area-5 validation. The best results are highlighted in **bold**, and the second-best results are underscored. "*" denotes additional data is used for pre-training.

Methods	Presented at	Detection		Semantic	Instance	Panoptic		
		mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
FCAF3D [16]	ECCV'22	66.7	45.7					
Uni3DETR [27]	NIPS'23	70.1	48.0					
TR3D [17]	ICIP'23	74.5	51.7					
SPGroup3D [1]	AAAI'24	69.2	47.2					
UniDet3D [7]	AAAI'25	75.2	60.8					
MinkUNet [31]	CVPR'19			65.4				
PointTransformer [29]	ICCV'21			70.4				
PointTransformerV2 [30]	NIPS'22			71.6				
TD3D [33]	WACV'24				48.6			
Mask3D* [2]	ICRA'23				57.8			
OneFormer3D* [3]	CVPR'24			**72.4**	**58.7**	**62.2**	58.4	**65.5**
Ours		**77.8**	**65.3**	68.8	53.0	59.7	**60.0**	59.6

(e.g., mIoU and PQ), two critical factors contextualize this result. First, OneFormer3D benefits from extensive additional pre-training data (is pretrained on Structured3D and ScanNet, marked with *), whereas UNIT3D is trained exclusively on the target dataset. More importantly, a granular analysis of Panoptic Quality reveals a strategic architectural trade-off. Notably, UNIT3D surpasses OneFormer3D in Thing Panoptic Quality (PQ_{th}) (60.0 vs. 58.4). This result validates our core hypothesis: by integrating detection objectives and instance-relative encodings, our model learns superior object-level geometric representations ("Things"). Although this geometric strictness leads to a slight compromise in "Stuff" segmentation (lower PQ_{st}) compared to the spatially-agnostic mask attention of OneFormer3D, UNIT3D effectively balances the two tasks. It delivers the best "Thing" understanding (highest Detection mAP and highest PQ_{th}) among all

Table 4
Comparison of NMS dependency between different 3D detection methods. Our UNIT3D method demonstrates significantly lower dependency on NMS compared to UniDet3D, maintaining competitive performance even when NMS is removed.

Methods	Detection	
	mAP@25	mAP@50
UniDet3D	77.0	65.9
UniDet3D w/o NMS	35.7	32.8
UNIT3D	79.4	68.4
UNIT3D w/o NMS	77.5	67.5

methods, establishing it as the preferred solution for instance-centric 3D perception.

4.2.4. Quantitative analysis

Our framework unifies 3D object detection and segmentation within a single architecture, leading to clear empirical gains across multiple datasets. In particular, the enhanced detection cost with MPDIoU improves box regression stability and accuracy, yielding consistent boosts in AP under both IoU@0.25 and IoU@0.5 settings, and demonstrating that joint optimization across tasks can strengthen object-level reasoning without sacrificing mask quality. These improvements are consistent across benchmarks, underscoring the practical value of unified perception over task-specific pipelines.

To enable this synergy, we introduce two instance-relative encoding mechanisms: SASA and VGCA, that preserve and propagate spatial cues across tasks. SASA encodes instance centers to capture global context, while VGCA leverages instance vertices to refine local geometry. Together, they provide soft spatial guidance that does not depend on hard masks, allowing the model to predict boxes and masks in a coordinated manner. This design supports accurate region-level localization and point-level labeling in a single pass, improving efficiency and reducing reliance on post hoc conversions.

Compared with contemporary approaches, our position encoding is simpler, more precise, and more adaptive to diverse point clouds. V-DETR relies on 3D vertex relative position encoding inside a DETR-style decoder; while effective, it introduces sensitivity to query box initialization and scene sampling density that can affect stability across domains. MAFT abandons mask attention and instead adds an auxiliary center regression task with contextual relative position encoding; although this alleviates the low-recall issue of early masks, it remains dependent on centroid-oriented priors and iterative refinement, making performance sensitive to center estimation and matching losses. In contrast, our instance-relative encodings compute direct relative positions to instances and their vertices, offering mask-independent soft guidance that unifies detection and segmentation. This reduces computational overhead associated with hard masking or elaborate positional tables, improves precision by conditioning attention on instance geometry, and scales robustly across datasets. Collectively, these choices establish a practical and effective paradigm for multi-task 3D perception.

4.2.5. NMS dependency analysis

Furthermore, our approach demonstrates significantly reduced dependency on NMS post-processing compared to existing detection models. Through experimental validation presented in Table 4, we compare UniDet3D and our method under NMS removal conditions, revealing compelling performance characteristics. When NMS is eliminated, UniDet3D experiences a dramatic performance degradation of over 30 percentage points, whereas our method exhibits a minimal decline of merely 1–2 percentage points. This remarkable resilience underscores the inherent strength of our one-to-one matching strategy, which fundamentally enhances the model's detection robustness and reduces reliance on traditional post-processing techniques. By mitigating the need for complex suppression mechanisms, our approach not only simplifies the detection pipeline but also demonstrates superior matching precision across different detection scenarios.

4.2.6. Inference efficiency

In this comprehensive performance analysis, as show in Table 5, we compare the inference efficiency of our proposed method against prominent 3D detection and segmentation approaches, including OneFormer3D [3], UniDet3D [7], V-DETR [6], and UNIT3D, utilizing the ScanNet validation dataset. To ensure a fair evaluation, we measured latency by averaging the inference time on the ScanNet validation set batchsize to 1. All experiments were conducted on a unified platform with an Intel Xeon Gold 5218 CPU and a single NVIDIA 3090 GPU, allowing for a detailed analysis of the time consumption and performance of individual components. Our empirical investigation reveals a pivotal innovation: by introducing minimal computational overhead, we have successfully developed a unified model that seamlessly integrates 3D detection and segmentation tasks, simultaneously achieving superior performance metrics. This approach not only demonstrates remarkable computational efficiency but also highlights the potential of a single, versatile architecture to transcend traditional task-specific limitations in 3D scene understanding.

4.2.7. Architectural comparison

To further elucidate the performance gains and efficiency of UNIT3D, we provide a schematic comparison with existing state-of-the-art methods in Table 6. Compared to specialized models such as UniDet3D, V-DETR (detection-only), and OneFormer3D (segmentation-only), UNIT3D is the first to achieve fully unified task coverage across both domains. Notably, while UniDet3D and V-DETR rely on Non-Maximum Suppression (NMS) for deduplication, UNIT3D implements an NMS-free pipeline via one-to-one bipartite matching. This is facilitated by our instance-relative encoding, which provides sufficient geometric discriminability to distinguish between overlapping queries.

Regarding inference overhead, despite incorporating a sophisticated global-to-local attention mechanism, UNIT3D maintains a highly competitive post-processing latency of 11 ms. This efficiency is primarily attributed to our NMS-free detection design, which only necessitates mask matching for the segmentation components. In contrast, V-DETR suffers from a significant post-processing delay (53 ms) due to its reliance on complex per-category bounding box decoding. These results demonstrate that achieving a unified 3D perception architecture does not necessitate a compromise in computational efficiency.

4.3. Qualitative evaluation

4.3.1. Attention map visualization

Fig. 4 illustrates the attention maps generated by our Vertex Guided Cross Attention (VGCA) mechanism on ScanNet scenes. The visualization reveals the sophisticated spatial reasoning capabilities of our approach: our VGCA demonstrates precise 3D bounding box localization by accurately identifying object boundaries. By leveraging the relative vertex positions of each instance, the mechanism dynamically enhances the internal regions within detected bounding boxes, effectively capturing the spatial context of objects. Moreover, our approach implements a nuanced soft attention mechanism that maintains contextual awareness beyond object boundaries, thereby ensuring robust generalizability across detection and segmentation tasks. This innovative attention strategy enables flexible feature aggregation, transcending the limitations of traditional hard masking techniques and providing a more adaptable representation learning framework for 3D perception.

4.3.2. Visualization of detection and segmentation results

To provide an intuitive understanding of the correlation between detection and segmentation metrics and their practical performance, we present comprehensive visualizations of original and segmented point clouds from three prominent indoor datasets: ScanNet [49] (Fig. 5), ScanNet200 [50] (Fig. 6), and S3DIS [51] (Fig. 7). Each visualization includes the original point cloud, ground-truth annotations, detector

Table 5
The inference time and instance segmentation accuracy on the ScanNet validation split. We show comparable inference time to OneFormer3D, UniDet3D, and V-DETR, being significantly more accurate than all existing methods.

Method	Component	Device	Component time, ms	Total time, ms	mAP@25	mAP@50	AP
OneFormer3D [3]	Superpoint extraction	CPU	168	220	–	–	69.3
	Backbone	GPU	29				
	Superpoint pooling	GPU	2				
	Query decoder	GPU	14				
	Post process	GPU	7				
UniDet3D [7]	Superpoint extraction	CPU	168	217	77.5	66.1	–
	Backbone	GPU	28				
	Superpoint pooling	GPU	1				
	Query decoder	GPU	10				
	Post process	GPU	10				
V-DETR [6]	Backbone	GPU	85	432	77.4	65.0	–
	Farthest point sampling	GPU	114				
	Query decoder	GPU	180				
	Post process	GPU	53				
UNIT3D	Superpoint extraction	CPU	168	237	79.4	68.4	69.6
	Backbone	GPU	28				
	Superpoint pooling	GPU	1				
	Query decoder	GPU	29				
	Post process	GPU	11				

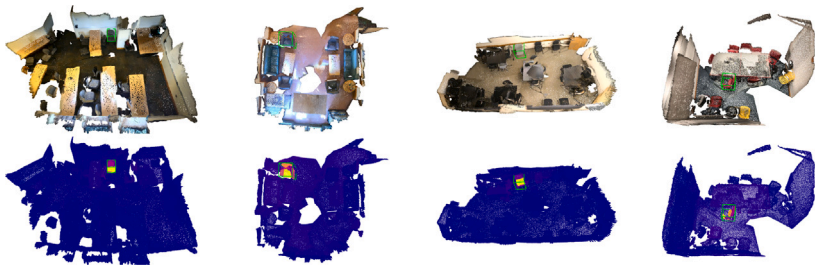


Fig. 4. Illustration of the attention maps learned by our Vertex Guided Cross Attention on ScanNetV2 scenes. The first row shows the input point clouds, and the second row shows the attention maps. The attention maps highlight the regions of interest in the point clouds, demonstrating the model's ability to focus on relevant areas for 3D scene understanding tasks.

Table 6
Architectural comparison. UNIT3D achieves a unified representation for both detection and segmentation by replacing task-specific attention with instance-relative encoding, enabling an NMS-free pipeline.

Feature	UniDet3D [7]	V-DETR [6]	OneFormer3D [3]	UNIT3D (Ours)
Task Scope	Detection	Detection	Segmentation	**Unified (Det.+Seg.)**
Attention Type	Self Attention	Vertex-Relative	Mask Attention	**Instance-Relative**
Positional Enc.	None	Vertex	None	**Global-to-Local**
NMS Dependency	Yes	Yes	Mask Matching	**No (NMS-Free)**
Post Process	10 ms	53 ms	7 ms	11 ms

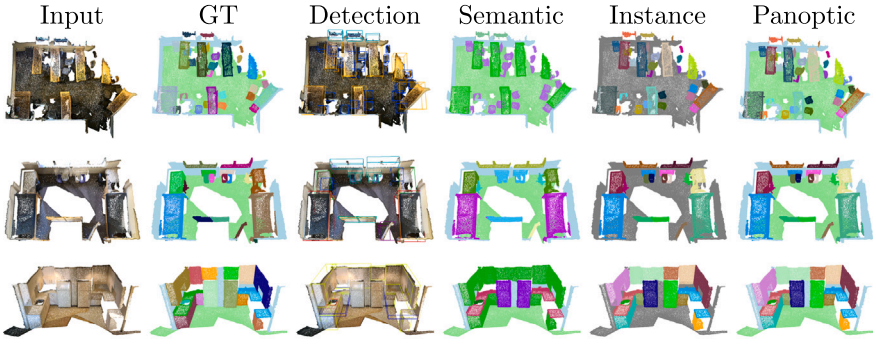


Fig. 5. Visualization of results on ScanNet Val set.

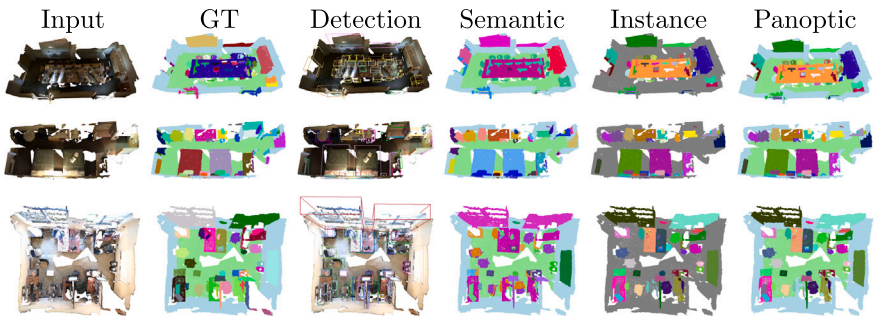


Fig. 6. Visualization of results on ScanNet200 Val set.

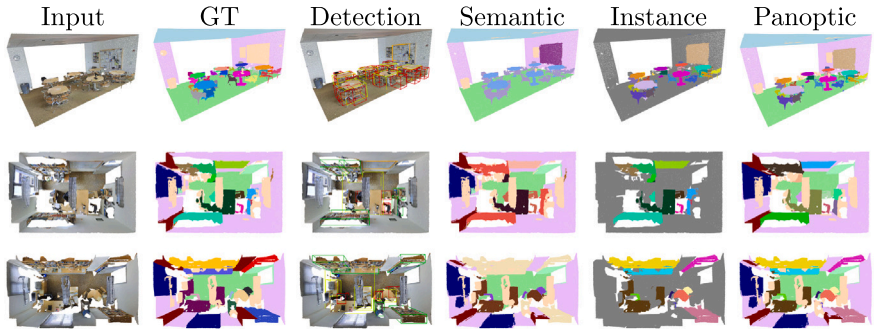


Fig. 7. Visualization of results on S3DIS Area-5.

Table 7
Effectiveness of the proposed components on ScanNet Val. OneFormer3D is the baseline model. The three components are added incrementally to constitute our UNIT3D.

Components	Detection		Semantic	Instance	Panoptic		
	mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
OneFormer3D [3]			76.6	59.3	71.2	69.6	86.1
+ Detection Head	62.4	45.5	73.8	53.1	70.2	68.4	86.8
+ Spatial Aware Self Attention	77.8	66.5	75.5	57.2	71.6	70.0	86.0
+ Vertex Guided Cross Attention	78.4	67.4	75.7	58.1	71.7	70.1	85.9

outputs, and the results of semantic, instance, and panoptic segmentation. These comprehensive panels bridge quantitative scores and visual interpretation, clarifying how gains in metrics such as mAP and PQ manifest in practice through cleaner object boundaries, fewer spurious predictions, improved small-object recall, and more consistent instance grouping across diverse indoor scenes.

4.4. Ablation studies

To assess the effectiveness of various components in our method, we perform an incremental ablation study on the ScanNet validation set. The results are shown in Table 7. The baseline "OneFormer3D" model achieves a moderate performance across segmentation tasks.

4.4.1. Detection head

The integration of a dedicated detection head in the baseline "OneFormer3D" model, along with corresponding detection cost and loss functions, highlights its importance in enhancing object detection capabilities. This module allows the model to directly output detection results without the need for mask post-processing. However, this addition adversely affects the performance of the original tasks.

4.4.2. Spatial aware self attention

Replacing standard self-attention with spatial-aware self-attention enables the model to better capture spatial relationships between points. This enhancement not only further boosts detection performance (mAP@25 to 77.8) but also benefits segmentation performance, as evidenced by the improvement in mIoU, AP, and PQ.

4.4.3. Vertex guided cross attention

Incorporating vertex-guided cross-attention allows the model to effectively leverage geometric information encoded in vertices. This component contributes to a more refined understanding of the 3D scene, leading to the best overall performance across most tasks.

The incremental incorporation of the detection head, spatial aware self attention, and vertex guided cross attention underscores the effectiveness of each component. The final model achieves significant improvements in detection, with results of 78.4 mAP@25 and 67.4 mAP@50. Additionally, it attains competitive results across various segmentation tasks, demonstrating strong performance in scene understanding. These results validate our unified architectural design by addressing the commonalities and limitations of 3D detection and segmentation tasks, thereby enhancing the comprehensive understanding of 3D scenes.

4.4.4. Position encoding ablation

We ablate two factors in the position encoding (PE): the transformation (Fourier vs. Sinusoidal) and the projection layer (MLP). Any PE substantially improves over no PE, confirming the need for positional cues. Sinusoidal PE is slightly stronger than Fourier; adding an MLP

Table 8
Ablation studies at different modules in the Position Encoding of the UNIT3D. The better options are marked in bold.

Transformation	Projection	Detection		Semantic	Instance	Panoptic		
		mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
None	None	62.4	45.5	73.8	53.1	70.2	68.4	**86.8**
Fourier	None	77.5	65.0	75.7	57.9	71.5	69.9	86.3
Fourier	MLP	76.8	65.7	74.6	56.8	70.4	68.7	85.8
Sinusoidal	None	77.7	65.3	75.0	57.0	70.1	68.2	86.7
Sinusoidal	MLP	**78.4**	**67.4**	**75.7**	**58.1**	**71.7**	**70.1**	85.9

Table 9
Ablation studies at different matching strategies of the UNIT3D. The better options are marked in bold.

Topk	Detection		Semantic	Instance	Panoptic		
	mAP@25	mAP@50	mIoU	AP	PQ	PQ_{th}	PQ_{st}
6	77.8	66.0	75.0	56.6	70.5	68.7	85.8
3	78.0	66.2	75.4	57.0	70.8	69.1	85.8
1	**78.4**	**67.4**	**75.7**	**58.1**	**71.7**	**70.1**	**85.9**

projection to Sinusoidal yields the best overall results, with detection mAP@25 and mAP@50 are 78.4 and 67.4, instance AP is 58.1, and panoptic PQ is 71.7, while semantic mIoU stays high at about 75.7 (Table 8). In contrast, Fourier+MLP offers limited gains over Fourier alone and is weaker on some detection and instance metrics. We also observe higher PQ_{th} and comparable PQ_{st} versus the no-PE baseline, indicating better thing localization without harming stuff quality.

4.4.5. *Matching strategy ablation*

We investigate the matching strategy by varying the **Topk** hyperparameter. Prior baselines adopt divergent settings: UniDet3D employs a looser one-to-many assignment (Topk=6), whereas OneFormer3D relies on strict one-to-one matching (Topk=1). As shown in Table 9, the UniDet3D-style setting (Topk=6) results in inferior performance across all metrics, with mAP@25 dropping to 77.8 compared to 78.4 for Topk=1. Although UniDet3D utilizes Topk=6 to encourage faster convergence and potential recall benefits during early training stages, we observe that this one-to-many assignment introduces ambiguity during inference that harms both localization and classification. By switching to **Topk=1**, similar to OneFormer3D, we enforce a strict bipartite matching constraint. This eliminates the need for Non-Maximum Suppression (NMS) and encourages the model to focus on the single best prediction for each ground truth, which is critical for the pixel-wise accuracy required in panoptic segmentation. Consequently, Topk=1 achieves the best overall results, securing the highest detection mAP@25 (**78.4**) and mAP@50 (**67.4**), as well as superior instance AP (**58.1**) and panoptic PQ (**71.7**).

5. Conclusion

In this paper, we present UNIT3D, a unified instance-relative Transformer that jointly tackles 3D object detection and semantic/instance/panoptic segmentation within a single architecture. By addressing inherent task conflicts, we establish a hierarchical global-to-local modeling paradigm realized through Spatial-Aware Self-Attention (Global) and Vertex-Guided Cross-Attention (Local). This design enables a fully end-to-end, NMS-free pipeline via strict one-to-one matching, ensuring robust predictions independent of heuristic tuning, marking a significant step towards general-purpose 3D perception. Crucially, our results demonstrate that this unification is not a compromise; rather, detection and segmentation mutually promote each other, with geometric constraints actively enhancing instance-level understanding under fair comparative settings. On ScanNet, ScanNet200, and S3DIS, our model delivers state-of-the-art detection performance and competitive segmentation results, achieving inference speeds comparable to single-task models. Ablation studies confirm that our unified encoding effectively bridges the gap between region-level regression and pixel-level classification. Limitations include reduced robustness under extreme sparsity and generalization beyond indoor datasets; future work will explore stronger multi-scale reasoning and extensions to open-vocabulary scenarios.

CRediT authorship contribution statement

Xinrun Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation. **Weikun Liu:** Visualization, Validation, Software, Investigation, Data curation. **Bin Fan:** Supervision. **Hongmin Liu:** Writing – review & editing, Supervision.

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.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant U2441251 and Grant U22B2055.

Data availability

Data will be made available on request.

References

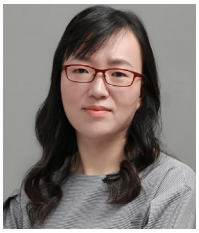
[1] Y. Zhu, L. Hui, Y. Shen, J. Xie, SPGroup3D: Superpoint grouping network for indoor 3D object detection, in: AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 7811–7819.
[2] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, B. Leibe, Mask3d: Mask transformer for 3d semantic instance segmentation, in: International Conference on Robotics and Automation, 2023, pp. 8216–8223.
[3] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, D. Rukhovich, OneFormer3D: One Transformer for Unified Point Cloud Segmentation, in: Computer Vision and Pattern Recognition, 2024, pp. 20943–20953.
[4] W. Xu, C. Shi, S. Tu, X. Zhou, D. Liang, X. Bai, A unified framework for 3d scene understanding, in: Neural Information Processing Systems, vol. 37, 2024, pp. 59468–59490.
[5] X. Li, J. Ding, Z. Chen, M. Elhoseiny, Uni3dl: Unified model for 3d and language understanding, in: European Conference on Computer Vision, 2023, pp. 1–16.
[6] Y. Shen, Z. Geng, Y. Yuan, Y. Lin, Z. Liu, C. Wang, H. Hu, N. Zheng, B. Guo, V-DETR: DETR with Vertex Relative Position Encoding for 3D Object Detection, in: International Conference on Learning Representations, 2024, pp. 1–19.
[7] M. Kolodiazhnyi, A. Vorontsova, M. Skripkin, D. Rukhovich, A. Konushin, Unidet3d: Multi-dataset indoor 3d object detection, in: AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 4365–4373.
[8] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L.M. Ni, H.-Y. Shum, Mask dino: Towards a unified transformer-based framework for object detection and segmentation, in: Computer Vision and Pattern Recognition, 2023, pp. 3041–3050.
[9] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, in: Neural Information Processing Systems, 2017, pp. 5099–5108.

[10] C.R. Qi, O. Litany, K. He, L.J. Guibas, Deep hough voting for 3d object detection in point clouds, in: European Conference on Computer Vision, 2019, pp. 9277–9286.
[11] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu, J. Wang, Mlcvnet: Multi-level context votenet for 3d object detection, in: Computer Vision and Pattern Recognition, 2020, pp. 10447–10456.
[12] Z. Zhang, B. Sun, H. Yang, Q. Huang, H3dnet: 3d object detection using hybrid geometric primitives, in: European Conference on Computer Vision, 2020, pp. 311–329.
[13] B. Cheng, L. Sheng, S. Shi, M. Yang, D. Xu, Back-tracing representative points for voting-based 3d object detection in point clouds, in: Computer Vision and Pattern Recognition, 2021, pp. 8963–8972.
[14] H. Ran, J. Liu, C. Wang, Surface representation for point clouds, in: Computer Vision and Pattern Recognition, 2022, pp. 18942–18952.
[15] J. Gwak, C. Choy, S. Savarese, Generative sparse detection networks for 3d single-shot object detection, in: European Conference on Computer Vision, 2020, pp. 297–313.
[16] D. Rukhovich, A. Vorontsova, A. Konushin, Fcaf3d: Fully convolutional anchor-free 3d object detection, in: European Conference on Computer Vision, 2022, pp. 477–493.
[17] D. Rukhovich, A. Vorontsova, A. Konushin, Tr3d: Towards real-time indoor 3d object detection, in: International Conference on Image Processing, 2023, pp. 281–285.
[18] H. Wang, S. Dong, S. Shi, A. Li, J. Li, Z. Li, L. Wang, et al., Cagroup3d: Class-aware grouping for 3d object detection on point clouds, in: Neural Information Processing Systems, vol. 35, 2022, pp. 29975–29988.
[19] X. Xu, Z. Sun, Z. Wang, H. Liu, J. Zhou, J. Lu, DSPDet3D: 3D small object detection with dynamic spatial pruning, in: European Conference on Computer Vision, 2025, pp. 355–373.
[20] X. Liu, L. Zhao, B. Fan, J. Lu, H. Liu, Dynamic learnable label assignment for indoor 3D object detection, Trans. Circuits Syst. Video Technol. 35 (10) (2025) 10134–10147.
[21] Z. Sun, X. Xu, B. Fan, J. Lu, H. Liu, OV-GT3D: A generalizable open-vocabulary two-stage 3D detector with dual path distillation, Pattern Recognit. (ISSN: 0031-3203) 171 (2025) 112156–112168.
[22] Z. Liu, Z. Zhang, Y. Cao, H. Hu, X. Tong, Group-free 3d object detection via transformers, in: European Conference on Computer Vision, 2021, pp. 2949–2958.
[23] I. Misra, R. Girdhar, A. Joulin, An end-to-end transformer model for 3d object detection, in: European Conference on Computer Vision, 2021, pp. 2906–2917.
[24] M. Tao, C. Zhao, M. Tang, J. Wang, Objformer: Boosting 3D object detection via instance-wise interaction, Pattern Recognit. 146 (2024) 110061–110070.
[25] B. Lu, Y. Sun, Z. Yang, R. Song, H. Jiang, Y. Liu, Hrnet: 3D object detection network for point cloud with hierarchical refinement, Pattern Recognit. 149 (2024) 110254–110265.
[26] J. Ma, Y. Huang, C. Qian, J. Kang, J. Liu, H. Zhang, W. Hong, Lgnet: Local and global point dependency network for 3D object detection, Pattern Recognit. 154 (2024) 110585–110595.
[27] Z. Wang, Y.-L. Li, X. Chen, H. Zhao, S. Wang, Uni3detr: Unified 3d detection transformer, in: Neural Information Processing Systems, vol. 36, 2024, pp. 39876–39896.
[28] C. Wang, W. Yang, X. Liu, T. Zhang, State space model meets transformer: A new paradigm for 3D object detection, in: International Conference on Learning Representations, 2025, pp. 1–24.
[29] H. Zhao, L. Jiang, J. Jia, P.H. Torr, V. Koltun, Point transformer, in: European Conference on Computer Vision, 2021, pp. 16259–16268.
[30] X. Wu, Y. Lao, L. Jiang, X. Liu, H. Zhao, Point Transformer V2: Grouped Vector Attention and Partition-Based Pooling, in: Neural Information Processing Systems, vol. 35, 2022, pp. 33330–33342.
[31] C. Choy, J. Gwak, S. Savarese, 4D spatio-temporal convnets: Minkowski convolutional neural networks, in: Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.
[32] L. Yi, W. Zhao, H. Wang, M. Sung, L.J. Guibas, Gspn: Generative shape proposal network for 3d instance segmentation in point cloud, in: Computer Vision and Pattern Recognition, 2019, pp. 3947–3956.
[33] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, D. Rukhovich, Top-down beats bottom-up in 3d instance segmentation, in: Conference on Applications of Computer Vision, 2024, pp. 3566–3574.
[34] T. He, C. Shen, A. Van Den Hengel, Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution, in: Computer Vision and Pattern Recognition, 2021, pp. 354–363.
[35] Z. Liang, Z. Li, S. Xu, M. Tan, K. Jia, Instance segmentation in 3d scenes using semantic superpoint tree networks, in: European Conference on Computer Vision, 2021, pp. 2783–2792.
[36] T. Vu, K. Kim, T.M. Luu, T. Nguyen, C.D. Yoo, Softgroup for 3d instance segmentation on point clouds, in: Computer Vision and Pattern Recognition, 2022, pp. 2708–2717.
[37] J. Lu, J. Deng, C. Wang, J. He, T. Zhang, Query refinement transformer for 3d instance segmentation, in: European Conference on Computer Vision, 2023, pp. 18516–18526.
[38] J. Sun, C. Qing, J. Tan, X. Xu, Superpoint Transformer for 3D Scene Instance Segmentation, in: AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 2393–2401.
[39] S.-C. Wu, J. Wald, K. Tateno, N. Navab, F. Tombari, SceneGraphFusion: Incremental 3D Scene Graph Prediction from RGB-D Sequences, in: Computer Vision and Pattern Recognition, 2021, pp. 7515–7525.
[40] Z. Yang, C. Liu, Tupper-Map: Temporal and Unified Panoptic Perception for 3D Metric-Semantic Mapping, in: International Conference on Intelligent Robots and Systems, 2021, pp. 1094–1101.
[41] Y. Siddiqui, L. Porzi, S.R. Bulò, N. Müller, M. Nießner, A. Dai, P. Kontschieder, Panoptic lifting for 3D scene understanding with neural fields, in: Computer Vision and Pattern Recognition, 2023, pp. 9043–9052.
[42] D. Seichter, B. Stephan, S.B. Fischedick, S. Müller, L. Rabes, H.-M. Gross, Panopticndt: Efficient and robust panoptic mapping, in: International Conference on Intelligent Robots and Systems, 2023, pp. 7233–7240.
[43] X. Li, H. Ding, H. Yuan, W. Zhang, J. Pang, G. Cheng, K. Chen, Z. Liu, C.C. Loy, Transformer-based visual segmentation: A survey, Trans. Pattern Anal. Mach. Intell. 46 (2024) 10138–10163.
[44] X. Li, H. Yuan, W. Li, H. Ding, S. Wu, W. Zhang, Y. Li, K. Chen, C.C. Loy, Omg-seg: Is one model good enough for all segmentation? in: Computer Vision and Pattern Recognition, 2024, pp. 27948–27959.
[45] X. Li, H. Yuan, W. Zhang, G. Cheng, J. Pang, C.C. Loy, Tube-link: A flexible cross tube framework for universal video segmentation, in: International Conference on Computer Vision, 2023, pp. 13923–13933.
[46] S. Su, J. Xu, H. Wang, Z. Miao, X. Zhan, D. Hao, X. Li, PUPS: Point cloud unified panoptic segmentation, in: AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 2339–2347.
[47] Z. Chen, L. Jing, Y. Li, B. Li, Bridging the domain gap: Self-supervised 3d scene understanding with foundation models, in: Neural Information Processing Systems, vol. 36, 2023, pp. 79226–79239.
[48] Z. Chen, L. Yang, Y. Li, L. Jing, B. Li, Sam-guided masked token prediction for 3d scene understanding, in: Neural Information Processing Systems, vol. 37, 2024, pp. 82962–82981.
[49] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.
[50] D. Rozenberszki, O. Litany, A. Dai, Language-grounded indoor 3d semantic segmentation in the wild, in: European Conference on Computer Vision, 2022, pp. 125–141.
[51] I. Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3D semantic parsing of large-scale indoor spaces, in: Computer Vision and Pattern Recognition, 2016, pp. 1534–1543.
[52] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: Computer Vision and Pattern Recognition, 2019, pp. 9404–9413.
[53] M. Contributors, MMDetection3D: OpenMMLab next-generation platform for general 3D object detection, 2020, https://github.com/open-mmlab/mmdetection3d.
[54] H. Lin, X. Zheng, L. Li, F. Chao, S. Wang, Y. Wang, Y. Tian, R. Ji, Meta architecture for point cloud analysis, in: Computer Vision and Pattern Recognition, 2023, pp. 17682–17691.
[55] X. Lai, Y. Yuan, R. Chu, Y. Chen, H. Hu, J. Jia, Mask-attention-free transformer for 3d instance segmentation, in: European Conference on Computer Vision, 2023, pp. 3693–3703.

Xinrun Liu received the B.Eng. degree from the School of Advanced Engineering, University of Science and Technology Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, University of Science and Technology Beijing, China. His research interests include 3D scene understanding and autonomous driving perception.

Weikun Liu is currently pursuing the B.Sc. degree with the School of Mathematics and Physics, University of Science and Technology Beijing, China. His research interests include 3D scene understanding.

Bin Fan received the B.S. degree from the Beijing University of Chemical Technology, Beijing, China, in 2006, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently a Professor with School of Artificial Intelligence and the Institute of Artificial Intelligence, University of Science and Technology Beijing, China. He has wide research interests in computer vision, pattern recognition, image processing, and multimedia.



Hongmin Liu received the B.S. degree from Xidian University, Xi'an, China, in 2004 and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009. She is currently a Professor with the School of Artificial Intelligence, University of Science and Technology Beijing, China. Her research intersts include computer vision and smart sensing.