GEOMETRY-AWARE 2D-3D MULTIMODAL FUSION FOR 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal 2D-3D segmentation leverages complementary 2D semantic priors and 3D geometric structures for scene understanding. However, geometric inconsistency across the 2D-3D fusion pipeline remains a key challenge. To address this challenge, we propose GAMFusion, a geometry-aware 2D-3D multimodal fusion framework designed to preserve geometric consistency throughout the 2D-to-3D feature flow. First, we construct an extensible parallel fusion topology that decouples heterogeneous modality pathways and introduces pixel-wise adaptive interaction to mitigate feature interference caused by modality discrepancies. Second, we introduce a corrective feature consistency alignment module that restores local surface structure before voxelization using geometric consistency constraints. Finally, we design a normal-aware 3D contextual fusion mechanism that incorporates surface normal constraints into neighborhood aggregation to reduce cross-surface feature leakage caused by Euclidean distance-based retrieval. Extensive experiments on ScanNet200, S3DIS, and AI2THOR demonstrate the effectiveness of the proposed method, with strong improvements over existing approaches, particularly on boundary-sensitive metrics.

1 INTRODUCTION

Multimodal 2D-3D segmentation exploits the complementarity between dense image and irregular 3D geometry for scene understanding. Images provide rich appearance and semantic priors, while point clouds and reconstructed surfaces preserve metric geometry for spatial reasoning. Prior 2D-3D fusion methods have shown that image features can enhance sparse 3D representations, especially when geometry alone lacks texture or fine-grained semantic cues.

Despite this promise, the advantages of 2D information are not always fully realized in existing 3D segmentation pipelines. Strong native 3D methods remain highly competitive on benchmarks such as ScanNet and ScanNet200, suggesting that simply lifting or fusing 2D cues does not guarantee stronger 3D representations. We argue that the bottleneck lies in how heterogeneous semantic and geometric cues are aligned under irregular 3D structures. This makes geometric consistency across the 2D-3D fusion pipeline a key challenge.

Prior work has mainly improved how image information is transferred into 3D. Projection-based methods such as SnapNet first apply 2D segmentation networks to rendered views and back-project pixel predictions to point clouds (Boulch et al., 2018). Instead of transferring final pixel predictions, feature-level methods such as MVPNet and DeepViewAgg aggregate multi-view image features into 3D points, addressing point-pixel correspondence and view selection (Jaritz et al., 2019). Unified frameworks such as ODIN further interleave within-view 2D processing with cross-view 3D fusion, reducing the separation between image-based and point-cloud-based reasoning (Jain et al., 2024). Despite this progress, feature transfer alone does not ensure that semantic cues remain aligned with local 3D surfaces and boundaries after entering the 3D space. This motivates us to preserve geometric consistency throughout the 2D-3D fusion pipeline.

To this end, we propose GAMFusion, a geometry-aware 2D-3D multimodal fusion framework for 3D segmentation. Rather than merely transferring 2D features into 3D space, GAMFusion organizes the feature flow around geometric consistency across multimodal encoding, 2D-to-3D alignment, and 3D context aggregation. In the 2D encoding stage, an Extensible Parallel Fusion Topology (EPFT) with a Pixel-wise Interaction Block (PIB) decouples heterogeneous modality pathways and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

adaptively integrates complementary cues to reduce early cross-modal interference. Before voxelization, a Corrective Feature Consistency Alignment (CFCA) module refines lifted features with depth and normal consistency to preserve local surface-aware associations. During 3D context modeling, a normal-aware aggregation mechanism augments Euclidean proximity with surface orientation consistency, reducing cross-surface feature leakage near object boundaries.

In summary, our main contributions are three-fold:

- **Reliability-aware heterogeneous modality fusion.** We propose an Extensible Parallel Fusion Topology (EPFT) with a Pixel-wise Interaction Block (PIB) to decouple heterogeneous modality pathways and adaptively integrate complementary RGB-X cues. This design mitigates early cross-modal interference caused by differences in physical meaning, numerical distribution, and sensor reliability.

- **Pre-voxelization geometry-consistent feature alignment.** We introduce a Corrective Feature Consistency Alignment (CFCA) module before voxelization to refine lifted 2D features using depth continuity and surface normal consistency. It preserves local surface-aware associations during 2D-to-3D feature transition and alleviates topology distortion caused by discrete lifting.

- **Surface-aware 3D context aggregation.** We design a normal-aware 3D context aggregation mechanism that augments Euclidean neighborhood retrieval with surface normal consistency. This encourages feature aggregation along coherent local surfaces and reduces cross-surface feature leakage near object boundaries.

2 Related Work

2.1 Multimodal 2D–3D Segmentation

Multimodal 2D–3D segmentation associates image appearance with 3D geometry through explicit cross-dimensional correspondence (Armeni et al., 2017). Early methods transfer either predictions or intermediate features: SnapNet back-projects 2D predictions, while 3DMV, MVPNet, and 3D-SIS fuse multi-view image features with 3D representations (Boulch et al., 2018; Dai & Nießner, 2018; Jaritz et al., 2019; Hou et al., 2019). Later approaches strengthen cross-modal interaction through bidirectional projection, visibility-aware aggregation, distillation, or alternating 2D–3D reasoning (Hu et al., 2021; Robert et al., 2022; Yan et al., 2022; Jain et al., 2024). Across these designs, the common objective is to establish reliable correspondence between dense image observations and irregular 3D elements so that appearance cues can complement geometric representations. The fusion location ranges from prediction space to intermediate features, and the interaction direction ranges from one-way transfer to repeated exchange between the two domains. These methods improve correspondence and information transfer, but do not explicitly ensure that lifted features remain consistent with local 3D surfaces.

Recent work further exploits pretrained 2D representations. OpenScene transfers image-language features for open-vocabulary understanding (Peng et al., 2023); CLIP2Scene and PLA transfer vision-language supervision to 3D (Chen et al., 2023; Ding et al., 2023); OpenMask3D and Region-PLC extend this direction to open-vocabulary instance or regional understanding (Takmaz et al., 2023; Yang et al., 2024); DITR injects frozen DINO features into 3D segmentation (Knaebel et al., 2025); Concerto learns joint 2D–3D representations (Zhang et al., 2026); and SegDINO3D uses both dense image features and object-level queries (Qu et al., 2026). Their success highlights the value of strong 2D priors, while leaving local surface consistency during lifting and subsequent 3D propagation comparatively underexplored. Stronger image semantics do not by themselves determine how nearby lifted features should interact after entering 3D space: around object contacts or depth discontinuities, semantically informative features can still be mixed across different surfaces if neighborhood construction is governed only by spatial proximity. This distinction is important for GAMFusion: the objective is not to replace stronger image representations, but to regulate how multimodal features are transferred and propagated once they enter irregular 3D space.

2.2 Geometry-Aware Representation for 3D Segmentation

Local aggregation in 3D is commonly organized by metric neighborhoods or spatial partitions. PointNet++, KPConv, RandLA-Net, and PointNeXt build point representations from local neighborhoods (Qi et al., 2017; Thomas et al., 2019; Hu et al., 2020; Qian et al., 2022), whereas Point Transformer variants extend local interaction with attention, stratified neighborhoods, partitioning, or serialized spatial ordering (Zhao et al., 2021; Lai et al., 2022; Wu et al., 2022; 2024). Voxel-based sparse convolution networks such as MinkowskiNet provide an alternative efficient backbone for unordered point-cloud data (Choy et al., 2019). Despite different operators, local connectivity is still largely determined by coordinates or structures derived from spatial proximity. This is a practical choice for irregular data, but coordinate distance measures spatial closeness rather than physical surface connectivity; two points separated by an object boundary can therefore become immediate neighbors even when their local orientations are inconsistent.

Geometric and boundary cues provide complementary information. PointGroup improves instance separation through offset-shifted grouping (Jiang et al., 2020), BFANet explicitly models boundary-sensitive features (Zhao et al., 2025), and HSurf-Net estimates robust surface normals from local hypersurfaces (Li et al., 2022). These studies indicate that coordinates alone do not fully characterize local surface structure. In particular, spatially close points may lie on different surfaces near object contacts or sharp transitions, motivating surface-consistent neighborhood construction. Surface normals describe local orientation independently of absolute position, while depth continuity provides a direct cue for visible foreground--background transitions; their complementary use is therefore more informative than either cue alone. This motivates the two geometric interventions used later in GAMFusion: one before voxelization, when point-level geometry is still available, and one during 3D aggregation, when neighborhoods determine subsequent feature flow.

2.2.1 Bottom-up 3D Instance Segmentation

Bottom-up instance segmentation first computes per-point semantic and geometric features and then groups points into object instances. Early representative approaches include SGPN which builds feature similarity matrix for point grouping (Wang et al., 2018). OccuSeg leverages occupancy signals for graph-based clustering to generate instance masks (Han et al., 2020). HAIS extends PointGroup with hierarchical aggregation to absorb fragmented instance components (Chen et al., 2021). SoftGroup (Vu et al., 2022) adopts soft semantic grouping to mitigate hard label error propagation and introduces a top-down refinement branch for false-positive suppression. These grouping-based approaches heavily rely on quality of per-point semantic prediction; semantic errors will propagate and degrade final instance outputs.

2.2.2 Top-down / Query-based 3D Instance Segmentation

Top-down approaches generate object proposals or instance queries then refine mask predictions. GSPN generates shape-based 3D proposals followed by point-net refinement (Yi et al., 2019). 3D-BoNet directly regresses bounding boxes without anchor design (Yang et al., 2019). More recently, Mask3D, SPFormer, and OneFormer3D adopt query-based Transformer decoding for 3D masks (Schult et al., 2023; Sun et al., 2023; Kolodiazhnyi et al., 2024), reducing reliance on hand-crafted center voting and geometric grouping.

Benchmark datasets are essential for evaluating indoor 3D perception algorithms. ScanNet provides large-scale annotated indoor RGB-D reconstructions (Dai et al., 2017), S3DIS delivers multi-area indoor scene annotations (Armeni et al., 2017), and AI2-THOR provides synthetic interactive indoor environments for simulation-based evaluation (Kolve et al., 2017). Our work is evaluated across these widely-adopted benchmarks.

3 Method

Given an RGB image I_{RGB} and spatially aligned auxiliary modalities $\mathcal{A} = \{A_1, \ldots, A_M\}$, GAMFusion predicts semantic and instance labels for the corresponding 3D scene. Figure 1 shows the four-stage pipeline: EPFT with PIB constructs a reliability-aware 2D representation; CFCA corrects lifted point features before voxelization; Normal-Aware 3D Contextual Fusion constrains

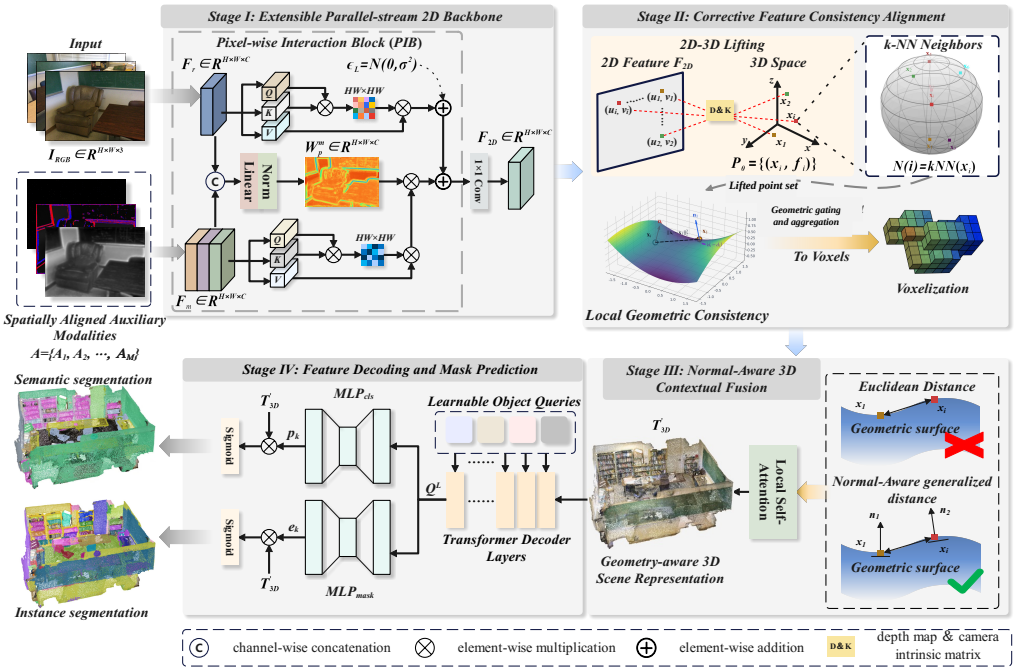


Figure 1: **Overview of GAMFusion.** Stage I performs reliability-aware RGB--auxiliary interaction with EPFT and PIB. Stage II lifts the fused 2D features into 3D and applies CFCA before voxelization. Stage III constructs normal-aware neighborhoods for local 3D context aggregation. Stage IV uses a shared Transformer decoder for semantic and instance prediction.

post-voxelization neighborhood aggregation; and a query-based decoder produces the final predictions. Stages II and III lie on opposite sides of voxelization, respectively regulating point-level mixing before discretization and context exchange afterwards. This separation preserves surface consistency across the 2D--3D feature flow while distinguishing cross-dimensional alignment from 3D contextual reasoning. The placement is deliberate: once voxel pooling merges multiple point features, their original point-level associations cannot be fully recovered, while a one-time pre-voxel correction alone cannot prevent later neighborhood operators from reconnecting spatially close but physically distinct surfaces.

3.1 Extensible Parallel-stream 2D Backbone

Different sensors exhibit distinct feature distributions and local reliability. Instead of stacking all inputs, EPFT uses RGB as the reference stream and lets each auxiliary modality interact with it independently. Let $\mathbf{F}^r, \mathbf{F}^m \in \mathbb{R}^{H\times W\times C}$ denote the RGB and m-th auxiliary feature maps, where H, W, and C are height, width, and channel dimension, respectively, M is the number of auxiliary modalities, and $m \in \{1, \ldots, M\}$. At spatial position p, PIB estimates

$$w_p^m = \mathrm{Norm}\big(\mathrm{Linear}(\mathbf{F}_p^r \oplus \mathbf{F}_p^m)\big)\,, \tag{1}$$

where $\oplus$ denotes channel concatenation, $\mathrm{Linear}(\cdot)$ is a learned projection, $\mathrm{Norm}(\cdot)$ normalizes the response, and w_p^m controls the local contribution of modality m. RGB and auxiliary streams generate projected query, key, and value vectors $\mathbf{Q}_{r,p}, \mathbf{K}_{r,p}, \mathbf{V}_{r,p}$ and $\mathbf{Q}_{m,p}, \mathbf{K}_{m,p}, \mathbf{V}_{m,p}$. Their interaction is

$$\mathbf{Z}_{r,p} = (\mathbf{Q}_{r,p}\mathbf{K}_{r,p}^\top)\mathbf{V}_{r,p}, \quad \mathbf{Z}_{m,p} = w_p^m(\mathbf{Q}_{m,p}\mathbf{K}_{m,p}^\top)\mathbf{V}_{m,p}. \tag{2}$$

Here p indexes one location of the $H \times W$ grid, while the projected query, key, and value vectors lie in the same d-dimensional interaction space; therefore each query--key product in Eq. equation 2 is a scalar compatibility response that modulates the corresponding value vector. A small perturbation $\epsilon_L \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is added to the RGB pathway during fusion, where σ_ϵ is its standard deviation and

the sampled perturbation is broadcast to the compatible RGB feature entries. The m-th enhanced branch is $\widetilde{\mathbf{F}}^m = \mathbf{Z}_r + \epsilon_L + \mathbf{Z}_m$, with the operations applied over all spatial positions. The branch outputs are finally integrated as

$$\mathbf{F}_{2D} = \mathrm{Conv}_{1\times1}\Big([\widetilde{\mathbf{F}}^1, \ldots, \widetilde{\mathbf{F}}^M]\Big), \tag{3}$$

where $[\cdot]$ denotes channel-wise concatenation and $\mathrm{Conv}_{1\times1}$ maps the concatenated features to the common 2D representation $\mathbf{F}_{2D}$. Reliability is therefore estimated before heterogeneous branches are merged, consistent with Stage I in Fig. 1. The weight w_p^m determines how much evidence from modality m is trusted at location p, while EPFT combines the independently filtered RGB--auxiliary branches. With one auxiliary modality the design reduces to an RGB--X dual stream; additional sensors only add parallel interaction branches and do not change Stages II--IV. Because every auxiliary branch is conditioned on the same RGB reference before branch fusion, modality-specific corruption can be suppressed locally without forcing direct low-level correspondence among heterogeneous sensors. The perturbation ϵ_L is confined to this fusion stage and does not affect the subsequent geometric modules.

3.2 Corrective Feature Consistency Alignment

Stage II transfers $\mathbf{F}_{2D}$ into continuous 3D space and corrects local associations before voxelization. Given depth map $D \in \mathbb{R}^{H\times W}$ and camera intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, pixel $p_i = (u_i, v_i)$ is lifted as

$$\mathbf{x}_i = D(u_i, v_i)\mathbf{K}^{-1}[u_i, v_i, 1]^\top, \qquad \mathbf{f}_i = \mathbf{F}_{2D}(p_i), \tag{4}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ is the 3D coordinate, $\mathbf{f}_i \in \mathbb{R}^C$ is its lifted feature, and u_i, v_i are pixel coordinates. The lifted set is $\mathcal{P}_0 = \{(\mathbf{x}_i, \mathbf{f}_i)\}_{i=1}^N$, where N is the number of valid lifted points. Lifting preserves the pixel--point correspondence but removes the regular 2D grid adjacency, so local relations must be reconstructed in 3D. For each point i, Euclidean k-NN provides a candidate neighborhood $\mathcal{N}(i)$, where k is the number of retrieved neighbors. CFCA then assigns each candidate $j \in \mathcal{N}(i)$ a geometric-consistency weight

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma_x^2}\right)\exp\left(-\frac{1 - \mathbf{n}_i^\top \mathbf{n}_j}{\sigma_n^2}\right)\mathbb{I}(|d_i - d_j| < \tau_d), \tag{5}$$

where $\mathbf{n}_i$ and $\mathbf{n}_j$ are unit surface normals, d_i and d_j are depth values, σ_x and σ_n control spatial and normal sensitivity, τ_d is the depth-continuity threshold, and $\mathbb{I}(\cdot)$ is the indicator function. The three factors favor spatial locality, similar surface orientation, and depth continuity. Here $\|\cdot\|_2$ denotes the Euclidean norm and $\exp(\cdot)$ the exponential kernel; thus w_{ij} is nonnegative, decreases with spatial or orientation inconsistency, and becomes zero across large depth discontinuities. Euclidean k-NN only supplies candidates, while their contributions are determined by joint surface compatibility. The corrected feature and voxel representation are

$$\widehat{\mathbf{f}}_i = \sum_{j\in\mathcal{N}(i)} w_{ij}\mathbf{f}_j, \qquad \mathbf{T}_{3D,t} = \mathrm{Mean}\{\widehat{\mathbf{f}}_i \mid \mathbf{x}_i \in \mathcal{V}_t\}, \tag{6}$$

where $\widehat{\mathbf{f}}_i$ is the corrected point feature, t indexes occupied voxels, $\mathcal{V}_t$ is the t-th occupied voxel, and $\mathbf{T}_{3D,t}$ is its pooled feature. Applying CFCA before voxelization uses point-level coordinates and surface cues before many-to-one voxel pooling makes their original associations difficult to recover. It is therefore different from post-voxel boundary refinement: no separate boundary map is predicted; existing geometry directly weights feature transport, and $\mathbf{T}_{3D,t}$ is pooled from corrected rather than raw lifted features. The three cues are intentionally complementary near object contacts. Spatial distance restricts the candidate range but cannot separate touching objects by itself; normal consistency is effective at orientation changes but can be ambiguous for nearly parallel surfaces; and depth continuity rejects visible foreground--background jumps. Their product therefore forms a conservative gate in which large weights are assigned only to candidates that are jointly local, orientation-consistent, and depth-consistent, matching the local geometric consistency operation in Stage II of Fig. 1.

3.3 Normal-Aware 3D Contextual Fusion

CFCA constrains pre-voxelization mixing, but spatially close features can again interact during 3D context aggregation. Stage III therefore replaces purely Euclidean retrieval with the generalized

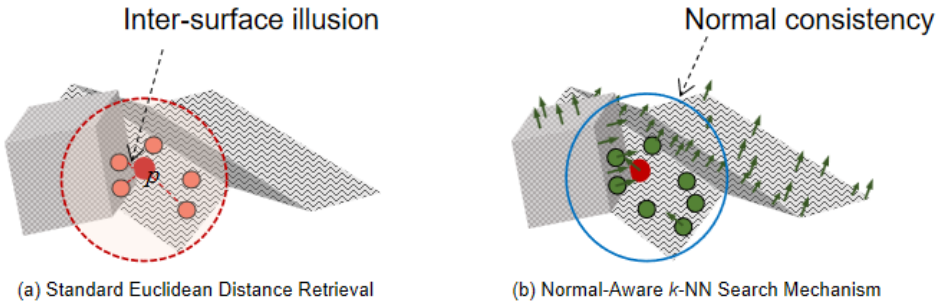


Figure 2: **Euclidean versus Normal-Aware k-NN retrieval.** Euclidean search may connect nearby points across a surface boundary, whereas the proposed distance penalizes normal disagreement and favors geometrically consistent neighbors.

distance

$$D_G(i,j) = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma_x^2} + \lambda \frac{\|\mathbf{n}_i - \mathbf{n}_j\|_2^2}{\sigma_n^2}, \qquad \mathcal{N}_G(i) = \mathrm{kNN}_{D_G}(i), \tag{7}$$

where $D_G(i,j)$ is the Normal-Aware distance, $\lambda \geq 0$ balances spatial proximity and normal discrepancy, and $\mathcal{N}_G(i)$ contains the k candidates with the smallest D_G. Here i and j index voxelized 3D elements; $\mathbf{x}_i$ and $\mathbf{n}_i$ are the representative coordinate and surface normal of element i, with analogous definitions for j, and σ_x, σ_n retain their CFCA meanings. When $\lambda = 0$, the metric reduces to normalized Euclidean retrieval; larger λ gives orientation greater influence on neighbor ranking. Unlike concatenating normals to features, Eq. equation 7 changes the neighborhood topology itself. Appending $\mathbf{n}_i$ to $\mathbf{T}_i$ can enrich the feature representation, but it cannot undo an Euclidean neighborhood that has already connected two different surfaces. By altering retrieval before aggregation, the proposed metric changes which elements are allowed to exchange information. As shown in Fig. 2, spatially close but orientation-inconsistent elements can be excluded before local attention, while coherent-surface elements remain favored.

Given $\mathcal{N}_G(i)$, local self-attention updates the voxel feature as

$$\mathbf{T}'_i = \mathbf{T}_i + \sum_{j \in \mathcal{N}_G(i)} \alpha_{ij} \mathbf{v}_j, \quad \alpha_{ij} = \frac{e^{\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_h}}}{\sum_{l \in \mathcal{N}_G(i)} e^{\mathbf{q}_i^\top \mathbf{k}_l / \sqrt{d_h}}}, \tag{8}$$

where $\mathbf{T}_i$ and $\mathbf{T}'_i$ are the input and updated 3D features; $\mathbf{q}_i = \mathbf{W}_Q \mathbf{T}_i$, $\mathbf{k}_j = \mathbf{W}_K \mathbf{T}_j$, and $\mathbf{v}_j = \mathbf{W}_V \mathbf{T}_j$ are query, key, and value vectors; $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learned projections; d_h is the query/key dimension; and α_{ij} is the normalized attention weight. The attention operator is standard, while geometry enters through its support $\mathcal{N}_G(i)$. CFCA and Normal-Aware aggregation are complementary: the former uses lifted points and depth before discretization, whereas the latter modifies the attention neighborhood after voxelization.

3.4 Feature Decoding and Mask Prediction

Stage IV uses a standard query-based Transformer decoder. Let $\mathbf{Q}^0 = \{\mathbf{q}_s^0\}_{s=1}^{K_q}$ denote K_q learnable object queries, where s indexes queries, and let $\mathbf{T}'_{3D} = \{\mathbf{T}'_i\}$ denote the geometry-aware scene representation. At decoder layer l,

$$\mathbf{Q}^{l+1} = \mathrm{DecoderLayer}(\mathbf{Q}^l, \mathbf{T}'_{3D}), \qquad l = 0, \ldots, L-1, \tag{9}$$

where L is the number of decoder layers and $\mathbf{Q}^l$ is the query set after layer l. Consistent with Figure 1, the s-th final query $\mathbf{q}_s^L$ is mapped by two heads,

$$\mathbf{p}_s = \mathrm{MLP}_{cls}(\mathbf{q}_s^L), \qquad \mathbf{e}_s = \mathrm{MLP}_{mask}(\mathbf{q}_s^L), \tag{10}$$

and matched with the scene representation to obtain semantic and instance responses,

$$\mathbf{M}_s^{sem} = \sigma(\mathbf{p}_s {\mathbf{T}'_{3D}}^\top), \qquad \mathbf{M}_s^{ins} = \sigma(\mathbf{e}_s {\mathbf{T}'_{3D}}^\top). \tag{11}$$

Table 1: Comparison on ScanNet200, AI2-THOR, and S3DIS.

(a) ScanNet200 instance segmentation

Method	mAP_{50}	Head	Common	Tail
TD3D	32.0	50.1	26.4	16.4
Mask3D	38.8	54.2	35.7	23.7
CTF	41.5	57.4	37.0	27.4
SegDINO3D	45.4	58.7	45.3	29.6
ODIN	45.3	50.7	37.5	23.7
Ours	**48.2**	**58.6**	**46.7**	**39.3**

(b) ScanNet200 semantic segmentation

Method	mIoU	Head	Common	Tail
CeCo	34.0	55.1	24.7	18.1
PonderV2	34.6	55.2	27.0	17.5
BFANet	36.0	55.3	29.3	19.3
ODIN	36.8	56.2	29.7	20.7
ALS	41.4	61.0	32.2	27.1
DITR	**44.9**	**62.9**	39.2	28.9
Ours	44.6	51.2	**42.8**	**39.9**

(c) AI2-THOR

Method	mAP	mAP_{50}	mAP_{25}	mIoU
Mask3D	60.6	70.8	76.6	–
ODIN	64.3	73.7	78.6	71.4
Ours	**67.4**	**76.2**	**80.5**	**74.7**

(d) S3DIS

Method	mAP	mAP_{50}	mAP_{25}	mIoU
Mask3D	40.7	54.6	64.2	–
MVPNet	–	–	–	62.4
ODIN	43.0	56.4	70.0	68.6
Ours	**44.0**	**57.3**	**72.1**	**69.8**

Here MLP_{cls} and MLP_{mask} are the prediction heads; $\mathbf{p}_s$ and $\mathbf{e}_s$ are query embeddings projected to the channel dimension of $\mathbf{T}'_{3D}$; $\mathbf{M}_s^{sem}$ and $\mathbf{M}_s^{ins}$ are semantic and instance response maps; and $\sigma(\cdot)$ is sigmoid. Eq. equation 11 computes query-to-scene responses over encoded 3D elements. The shared decoder is standard, so GAMFusion changes the representation delivered to it rather than the task-specific decoding mechanism.

4 Experiments

Experimental setup. GAMFusion is evaluated on ScanNet200, S3DIS, and AI2-THOR for 3D semantic and instance segmentation using mAP, $mAP_{25/50}$, and $mIoU$; Head/Common/Tail results are also reported on ScanNet200. Swin-B is the 2D backbone with 11 sampled 512×512 keyframes. Training uses AdamW (1×10^{-4}), batch size 3, standard geometric/image augmentation, and random view dropping on one RTX PRO 6000. Ghost points and high-resolution sub-sampling are retained. The voxel size is 2 cm, $k = 16$, and $\lambda = 0.5$; the decoder uses 100 queries and 6 layers. Surface normals are pre-computed with HSurf-Net (Li et al., 2022). mAP_{25} and mAP_{50} use IoU thresholds 0.25 and 0.50, while $mIoU$ evaluates semantic overlap. DeLiVER is used for multimodal-fusion ablations because it provides RGB, Depth, Event, and LiDAR, whereas the three indoor 3D benchmarks evaluate the complete pipeline. The same geometric settings are used across datasets unless stated otherwise.

4.1 Comparisons with State-of-the-Art

Table 1 compares GAMFusion with representative 3D and multimodal methods (Schult et al., 2023; Jain et al., 2024). On ScanNet200 instance segmentation, GAMFusion reaches 48.2% mAP_{50}, exceeding SegDINO3D (45.4%) and ODIN (45.3%); its Tail score of 39.3% is 9.7 points above SegDINO3D. For semantic segmentation, 44.6% overall mIoU is comparable to DITR (44.9%), while Common and Tail achieve the best 42.8% and 39.9%. The stronger Tail result, together with a lower Head score, suggests that geometry-consistent features provide complementary evidence particularly for less frequent categories rather than uniformly shifting all classes.

The gains transfer across datasets: on AI2-THOR, GAMFusion obtains 67.4% mAP, 76.2% mAP_{50}, and 74.7% mIoU versus 64.3%, 73.7%, and 71.4% for ODIN; on S3DIS, it reaches 44.0% mAP and 69.8% mIoU. The consistent trend across synthetic and real indoor scenes indicates that the benefit is not tied to a single acquisition setting. Figure 3 further shows cleaner separation around contacting furniture, thin structures, and cluttered boundaries. These cases are precisely where spatial proximity is least reliable as a proxy for surface connectivity, making them useful qualitative evidence for the proposed geometry-consistent propagation strategy.

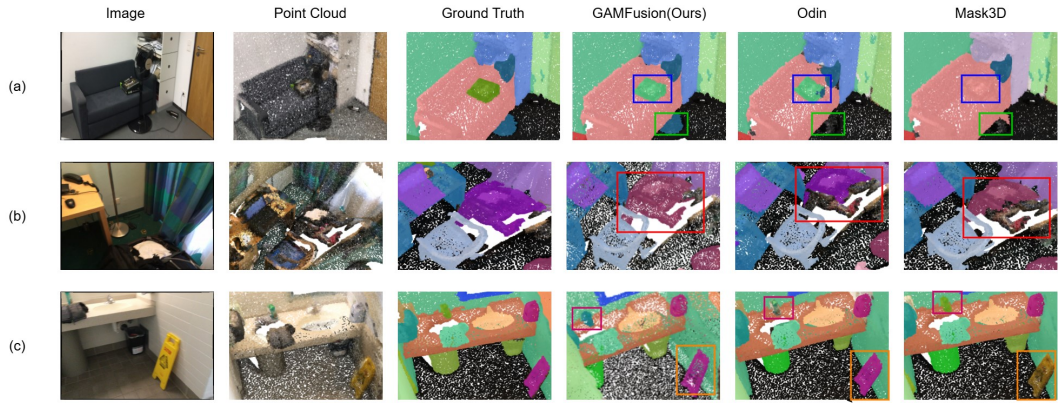


Figure 3: Qualitative comparison on challenging indoor scenes. GAMFusion better preserves boundaries and separates spatially adjacent structures than ODIN and Mask3D, especially in the highlighted regions.

Table 2: Multimodal fusion on DeLiVER ($mIoU$ %).

Input	Concat	PIB+EPFT	Δ
RGB	52.1	52.1	–
RGB+D	63.8	66.5	2.7
RGB+E	55.6	58.4	2.8
RGB+L	55.4	58.9	3.5
RGB+D+E	63.6	66.7	3.1
RGB+D+L	63.4	66.8	3.4
RGB+E+L	56.1	59.8	3.7
RGB+D+E+L	63.5	66.9	3.4

Qualitative results follow the same trend. Competing methods often fail around touching furniture, narrow gaps, and small structures, where Euclidean proximity poorly reflects surface connectivity. GAMFusion keeps these structures separated without uniformly smoothing predictions, supporting the role of CFCA and Normal-Aware retrieval in restricting cross-surface exchange.

4.2 Ablation Studies

Multimodal 2D fusion. Table 2 evaluates EPFT and PIB with RGB, Depth (D), Event (E), and LiDAR (L). PIB+EPFT improves every multimodal setting by 2.7--3.7% mIoU. Direct concatenation drops from 63.8% for RGB+D to 63.5% for RGB+D+E+L, whereas parallel interaction reaches 66.9%, supporting modality-specific filtering before joint fusion. Figure 4 shows lower responses in corrupted depth regions while reliable structures remain active, consistent with the pixel-wise role of w_p^m in Eq. equation 1.

CFCA and Normal-Aware aggregation. Table 3 shows that CFCA raises mAP from 29.8% to 31.4% and mAP_{50} from 43.5% to 44.9%, while mAP_{25} changes from 54.2% to 53.6%. Adding Normal-Aware aggregation improves the metrics to 33.9%, 48.2%, and 56.8%, including a further 3.3-point mAP_{50} gain over CFCA alone. Together the modules gain 4.1 points in mAP and 4.7 points in mAP_{50} over the RGB-D baseline, supporting the complementary Stage-II/Stage-III design in Fig. 1.

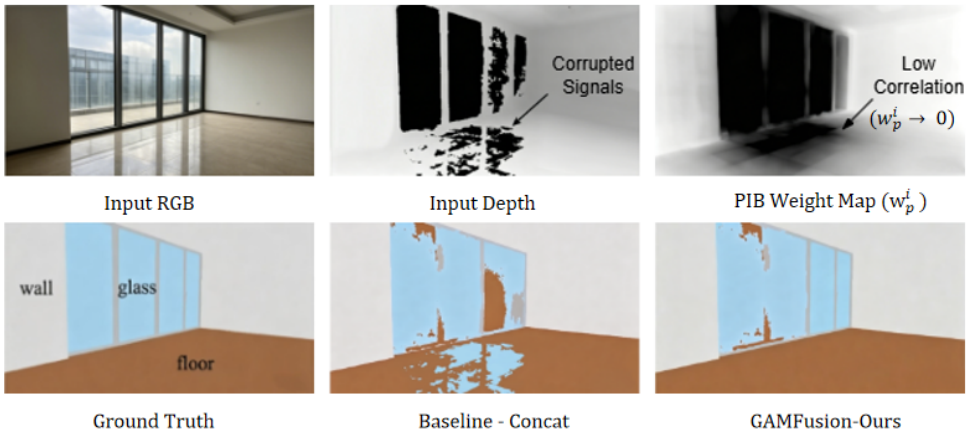


Figure 4: PIB responses under locally corrupted auxiliary observations. Lower responses suppress unreliable regions while preserving useful complementary cues.

Table 3: Ablation of geometry-aware components.

Config	mAP	mAP_{50}	mAP_{25}
Baseline (RGB-D)	29.8	43.5	54.2
+ CFCA	31.4	44.9	53.6
+ CFCA + Normal-Aware	**33.9**	**48.2**	**56.8**
Total gain (pts.)	+4.1	+4.7	+2.6

Table 4: Effect of normal sources and inference cost.

Source	mAP_{50}	mIoU	Total ms
None	44.9	41.5	28.4
PCA	45.3	42.1	30.6
Online	46.8	43.2	41.2
HSurf-Net	**48.2**	**44.9**	40.6
GT mesh	48.5	45.2	28.6

Normal estimation. Table 4 shows that HSurf-Net normals improve mAP_{50} from 44.9% to 48.2% and mIoU from 41.5% to 44.9%, close to the GT-normal upper bound. PCA gives only a small gain, confirming the value of accurate surface orientation. Pre-computed HSurf-Net normals keep segmentation inference at 28.6 ms; including about 12 ms preprocessing gives 40.6 ms total latency, slightly below online joint prediction (41.2 ms). The small gap to ground-truth mesh normals also indicates that learned normals are sufficiently accurate for neighborhood construction. Remaining limitations are the additional normal-preprocessing cost and the use of fixed geometric hyperparameters across local regions with different surface scales; adaptive bandwidth selection is a natural direction for future work.

5 Conclusion

This paper proposes GAMFusion, a geometry-aware multimodal 2D--3D fusion framework for semantic and instance segmentation. EPFT and PIB reduce heterogeneous modality interference, while CFCA and Normal-Aware aggregation preserve surface consistency before and after voxelization. Experiments on ScanNet200, AI2-THOR, S3DIS, and DeLiVER validate the complementary effects of these components, particularly on boundary-sensitive metrics and less frequent categories. Overall, the results support geometry-consistent feature propagation for robust 3D scene understanding.

REFERENCES

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding, 2017.

Alexandre Boulch, Joris Guerry, Bertrand Le Saux, and Nicolas Audebert. Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks. *Computers & Graphics*, 71:189–198, 2018.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7030, 2023.

Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, pp. 15467–15476. IEEE, 2021.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks, 2019.

Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *ECCV*, pp. 458–474. Springer, 2018.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7010–7019, 2023.

Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation, 2020.

Ji Hou, Angela Dai, and Matthias Niessner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4421–4430, 2019.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11108–11117, 2020.

Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, pp. 14368–14377. IEEE, 2021.

Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: a single model for 2d and 3d segmentation. In *CVPR*, pp. 3564–3574. IEEE, 2024.

Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *ICCV Workshops*, 2019.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation, 2020.

Karim Knaebel, Kadir Yilmaz, Daan De Geus, Alexander Hermans, David Adrian, Timm Linder, and Bastian Leibe. Dino in the room: Leveraging 2d foundation models for 3d segmentation, 2025.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20943–20953, 2024.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *ArXiv*, abs/1712.05474, 2017.

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *CVPR*, pp. 8490–8499, 2022.

Qing Li, Yu-Shen Liu, Jin-San Cheng, Cheng Wang, Yi Fang, and Zhizhong Han. HSurf-Net: Normal estimation for 3D point clouds by learning hyper surfaces. In *NeurIPS*, volume 35, pp. 4218-4230, 2022.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies, 2023.

Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23192–23204, 2022.

Jinyuan Qu, Hongyang Li, Xingyu Chen, Shilong Liu, Yukai Shi, Tianhe Ren, Ruitao Jing, and Lei Zhang. Segdino3d: 3d instance segmentation empowered by both image-level and object-level 2d features. In *AAAI*, volume 40, pp. 8649–8658, 2026.

Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *CVPR*, pp. 5565–5574. IEEE, 2022.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pp. 2220–2227. IEEE, 2023.

Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2393–2401, 2023.

Ayca Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68367–68390, 2023.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds, 2019.

Thang Vu, Kookhoi Kim, Tung M Luu, Xuan Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *CVPR*, pp. 2688–2697, 2022.

Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation, 2018.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33330–33342, 2022.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, pp. 4840–4851. IEEE, 2024.

Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*, pp. 677–695. Springer, 2022.

Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds, 2019.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19823–19832, 2024.

Li Yi, Wang Zhao, Wang He, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation, 2019.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In *NeurIPS*, volume 38, pp. 69498–69522, 2026.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pp. 16259–16268, 2021.

Weiguang Zhao, Rui Zhang, Qiufeng Wang, Guangliang Cheng, and Kaizhu Huang. Bfanet: Revisiting 3d semantic segmentation with boundary feature analysis, 2025.