This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Insightful Instance Features for 3D Instance Segmentation

Wonseok Roh[1] Hwanhee Jung[1] Giljoo Nam[2] Dong In Lee[1]
Hyeongcheol Park[1] Sang Ho Yoon[3] Jungseock Joo[4] Sangpil Kim[1*]

[1]Korea University [2]Meta [3]KAIST [4]UCLA

Abstract

*Recent 3D Instance Segmentation methods typically encode hundreds of instance-wise candidates with instance-specific information in various ways and refine them into final masks. However, they have yet to fully explore the benefit of these candidates. They overlook the valuable cues encoded in multiple candidates that represent different parts of the same instance, resulting in fragments. Also, they often fail to capture the precise spatial range of 3D instances, primarily due to inherent noises from sparse and unordered point clouds. In this work, to address these challenges, we propose **IKNE**, a novel instance-wise knowledge enhancement approach. We first introduce an Instance-wise Knowledge Aggregation (IKA) to associate scattered single instance details by optimizing correlations among candidates representing the same instance. Moreover, we present an Instance-wise Structural Guidance (ISG) to enhance the spatial understanding of candidates using structural cues from ambiguity-reduced features. Here, we utilize a simple yet effective truncated singular value decomposition algorithm to minimize inherent noises of 3D features. In our extensive experiments on large-scale datasets, ScanNetV2, ScanNet200, S3DIS, and STPLS3D, IKNE outperforms existing works. We validate the effectiveness of our modules in both kernel-based and transformer-based architectures.*

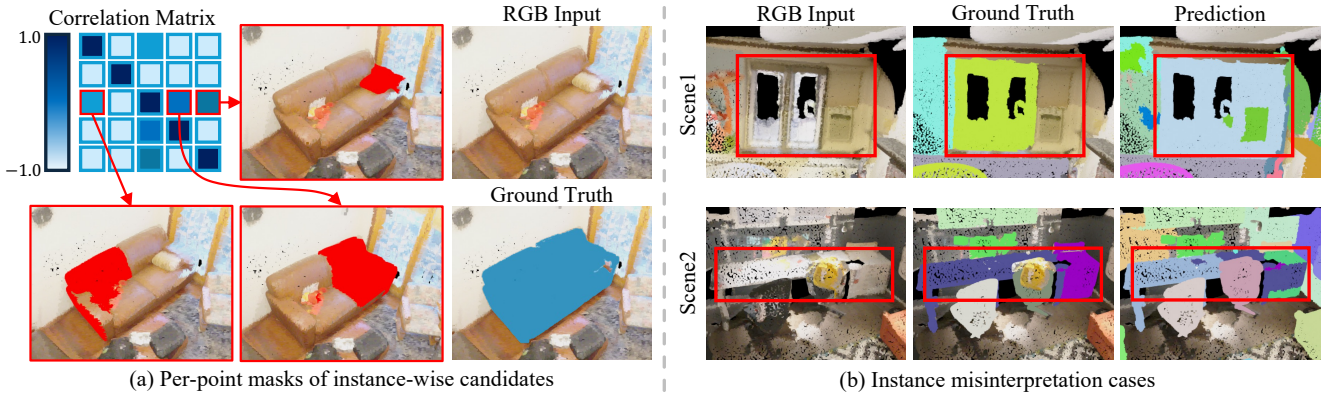
1. Introduction

3D Instance Segmentation (3DIS) is a fundamental 3D computer vision task that significantly contributes to real-world applications. Given 3D point cloud scenes, the 3DIS task identifies respective instances and assigns semantic class labels with the goal of comprehensively understanding entire spatial environments. In real-life 3D scenarios, substantial occlusion and truncation commonly arise, especially when objects are overlapped or obscured by others. To tackle these challenges, early approaches mainly focused on accurately generating region proposals (top-down) [21, 50, 52] or effectively grouping points with clustering algorithms (bottom-up) [6, 24, 45]. However, they presuppose that the intermediate processes, such as bounding box detection [18] or heuristic voting mechanism [34], generate near perfect results, which is often not the case.

Recently, kernel-based [19, 20, 32, 48] and transformer-based [29, 31, 39, 43] 3DIS approaches have been proposed, aiming to overcome the limitations of traditional frameworks. Kernel-based methods leverage instance-aware kernels for dynamic convolution, which decodes instance masks. They represent instances as kernels, replacing clustering algorithms with point sampling processes. On the other hand, transformer-based methods train instance queries to identify features about individual objects using attention mechanisms. These queries enable the direct prediction of per-point categories and instance labels. In both architectures, models utilize hundreds of instance-wise candidate features, to estimate the final instance mask. For example, ISBNet [32] (kernel-based) produces 256 instance kernels, and MAFT [29] (transformer-based) leverages 400 instance queries. However, despite the significant improvements based on their architectural advantages, they have yet to fully optimize the handling of these numerous candidates and still face several challenges, as illustrated in Fig. 1. To tackle these challenges, we introduce two main modules, IKA and ISG, each thoughtfully designed to enhance the informativeness of instance candidate features.

First, we empirically observe that multiple candidates from previous works often encode cues from the same single instance, as shown in Fig. 1 (a). These candidates are then typically decoded into considerable fragment masks representing different components of each instance, such as the *backrest* or *armrest* of the *sofa*. In this work, to estimate the complete coverage of each instance, we introduce an *Instance-wise Knowledge Aggregation* (IKA) network, which integrates the scattered intelligence from multiple fragments beyond relying solely on a few fragments based on confidence scores. Specifically, we optimize the correlations among instance-wise candidate features, encouraging the network to be aware of the distributed instance-specific knowledge. We initially identify candidates representing the same instance based on the similarity of their features, as

*Corresponding author (spk7@korea.ac.kr)



(a) Per-point masks of instance-wise candidates (b) Instance misinterpretation cases

Figure 1. Examples of two challenging cases of 3DIS. (a) Per-point masks of instance-wise candidates: highly correlated instance-wise candidates usually represent incomplete fragments of the same single instance, such as the *backrest* or *armrest* of a *sofa*. And (b) Instance misinterpretation cases: existing studies often confuse instances from backgrounds (Scene1) or misunderstand the spatial range and appearance of instances (Scene2). Note that we also provide additional examples for both cases in the supplementary material (Fig. 1).

highly correlated candidates usually capture the information from the same instance. Then, inspired by correlation regularization mechanisms [2, 23, 53], we softly guide identified candidates to be closer to each other in the latent space, aiming to enhance solidarity among them. As a result, IKA facilitates the network to learn relationships between candidates and aggregate their comprehensive details, maximizing the advantages of using hundreds of candidates.

In addition to IKA, we explore additional avenues to enrich the knowledge of instance-wise candidates. Unlike the human visual cognitive system, which perceives instances with sufficient prior knowledge, the network understands instances depending on their external properties from point features. However, point features are frequently unstable, primarily due to ambiguous inherent noises [13, 36, 47, 49] caused by sparse and incomplete point clouds. These noises are likely to confuse the models, making it hard to precisely interpret the specific spatial range and unique appearance of instances, as shown in Fig. 1 (b). To mitigate the negative impact of such noises, we propose an *Instance-wise Structural Guidance* (ISG) network specifically designed to improve candidates' structural understanding of 3D instances from point features. Here, we utilize the truncated singular value decomposition (SVD) [14, 42] algorithm to strengthen essential cues (*e.g.*, shape) of instance features while explicitly reducing fuzzy noises. Then, we effectively transfer fundamental clues for object cognition from ambiguity-minimized features to corresponding original features based on the cross-correlations between them. Ultimately, our instance features are now highly informative across complex 3D instances with aggregated and structural knowledge from IKA and ISG. Our novel modules are applicable to frameworks employing considerable instance candidates, regardless of their underlying model structures, including kernel-based and transformer-based architectures.

Given landmark datasets for 3DIS, ScanNetV2 [7], ScanNet200 [38], S3DIS [1], and STPLS3D [5], we validate the effectiveness of our novel framework. Our method outperforms the existing state-of-the-art methods. To summarize, our main contributions are listed as follows:

- To extend the benefits of producing multiple instance-wise candidates, we introduce the Instance-wise Knowledge Aggregation (IKA) network, which associates scattered instance-specific information of the same single instance by optimizing correlations among them.
- We design the Instance-wise Structural Guidance (ISG) network to improve the structural knowledge of candidates. Specifically, we use a simple yet strong truncated singular value decomposition process to emphasize essential cues from point features while filtering out noises.
- We analyze the effectiveness of our framework, **IKNE**, on four challenging datasets. Comprehensive experiments in various 3D scenarios demonstrate that our novel method achieves new State-of-the-Art performance on 3DIS.

2. Related Work

3D Instance Segmentation (3DIS) aims to distinguish individual instances within 3D scenes and assign corresponding semantic categories. Typically, 3DIS approaches can be categorized into four types: proposal-based [21, 50, 52], grouping-based [6, 11, 12, 24, 45], kernel-based [19, 20, 32, 48], and transformer-based [29, 31, 37, 39, 43]. Based on the impressive achievements of detection methods [18, 46], proposal-based approaches first detect instances and employ proposed regions as hard references to predict masks. However, these strategies, such as 3D-SIS [21] or 3D-BoNet [50], heavily rely on the quality of the outputs from the detection process. On the other hand, grouping-based methods catch 3D instances through a bottom-up pipeline, aggregating closely related points into instances using predicted semantic categories and center offsets. Yet, they still depend on intermediate manually tuned processing, such as point grouping [24] or heuristic voting mechanisms [34].

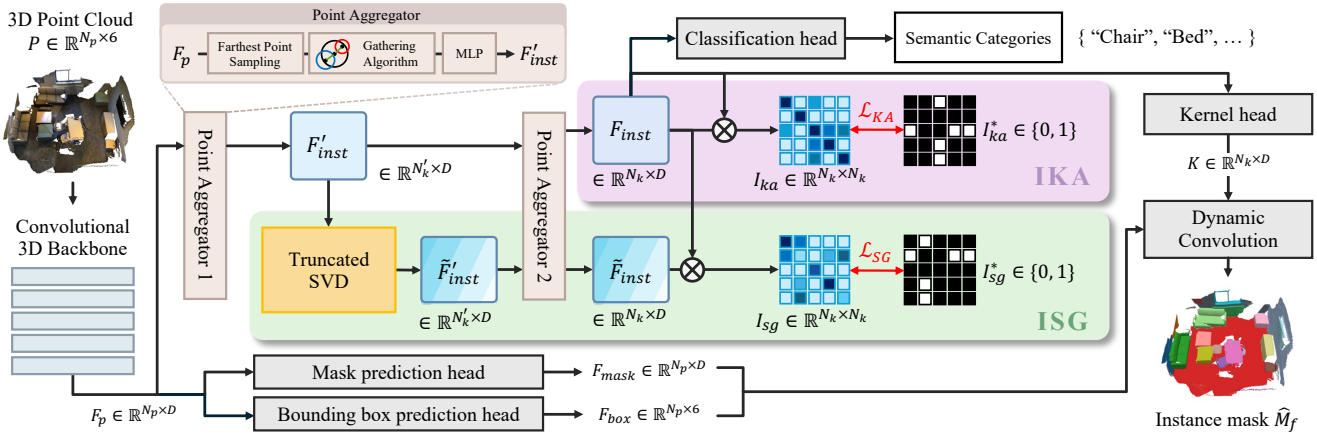


Figure 2. An overview of our framework IKNE. Built upon the modern kernel-based structure, our model consists of four main modules: (1) Sparse Convolutional 3D Backbone; (2) Instance-wise Knowledge Aggregation (IKA, Sec. 3.2), which softly associates scattered cues from highly correlated instance features of F_{inst}; (3) Instance-wise Structural Guidance (ISG, Sec. 3.3), which smartly instructs F_{inst} with essential clues from $\tilde{F}_{inst}$, using an effective SVD algorithm; and (4) Dynamic Convolution, which produces $\hat{M}_f$ from kernels K.

To address these limitations, kernel-based approaches have been introduced. For example, DyCo3D [19] employs a clustering algorithm to generate kernels for dynamic convolutions to predict instance masks, while ISBNet [32] presents a cluster-free method using instance-wise kernels. Recently, Spherical Mask [41], based on ISBNet, has overcome the low-quality results of coarse-to-fine strategies by using spherical representation to mitigate the propagation of false negatives and overestimation of instance size. Furthermore, transformer-based methods, Mask3D [39], MAFT [29], and EASE [37] utilize instance-wise queries to encode individual instance information through attention mechanisms. In this work, we focus on improving the quality of numerous instance-wise candidates introduced from recent studies with two novel approaches, IKA and ISG.

Singular Value Decomposition (SVD) is a fundamental algorithm in linear algebra, widely used for matrix factorization [14, 27, 42], dimensionality reduction [22, 51], and 2D image processing [9, 16, 26, 35, 40]. SVD decomposes a matrix $A \in \mathbb{R}^{m \times n}$ into the dot product of three matrices, $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$, and $V \in \mathbb{R}^{n \times n}$, each containing meaningful vectors and scalar values, as follows:

$$A = U \cdot \Sigma \cdot V^T \tag{1}$$

where U and V are the matrices with orthonormal columns, satisfying $U^T U = V^T V = I$, and Σ is a diagonal matrix with nonnegative singular values σ in descending order as $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r$. Based on SVD, truncated SVD [17] is an insightful dimension-reduction technique that minimizes nonessential noises while highlighting essential cues from the original matrix A. It produces a low-rank approximation $\tilde{A} \in \mathbb{R}^{m \times n}$ by retaining the most important top k singular values and their corresponding vectors from U, Σ, and V:

$$\tilde{A} = U_k \cdot \Sigma_k \cdot V_k^T \tag{2}$$

where $U_k \in \mathbb{R}^{m \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, and $V_k \in \mathbb{R}^{n \times k}$ are the truncated matrices with the top k singular value vectors.

In the field of 2D image processing, SVD has been extensively utilized for low-rank approximation. [16, 35] employ SVD to factorize and estimate low-rank patches, effectively reducing noise and enhancing image quality. Similarly, [4, 25] operate SVD to detect tampering by extracting and analyzing low-dimensional representations of images for image forensics. Also, SVD showcases robust performance in image compression [3, 33] and recovery [40]. Further, [9] leverages SVD within CNNs to learn subspace information for hyperspectral image reconstruction. However, despite its potential benefits, the introduction of SVD has received limited attention in 3D vision, especially for point clouds. Based on these observations, we apply SVD to address point features that contain inherent noises due to the unordered properties of point clouds, facilitating the model to become more robust against complex ambiguities.

3. Method

In this section, we introduce a novel 3D Instance Segmentation framework IKNE. We provide an overview of the whole pipeline (Sec. 3.1 and Fig. 2), followed by detailed technical discussions: Instance-wise Knowledge Aggregation (Sec. 3.2) and Instance-wise Structural Guidance (Sec. 3.3).

3.1. Overview

Recent 3DIS methods [19, 20, 32, 37, 39] commonly follow a similar paradigm; they spread hundreds of instance candidates across 3D scenes to capture instance-specific information in various ways and refine them into final instance masks. In this work, we encourage these numerous instance candidates to be highly informative, improving their understanding of diverse real-world instances. We describe our instance-wise knowledge enhancement approaches based on the modern kernel-based architecture [32]. As shown in

Fig. 2, our model consists of four main modules: (1) Sparse Convolutional 3D Backbone; (2) Instance-wise Knowledge Aggregation, which relates the scattered information of individual instances; (3) Instance-wise Structural Guidance, which effectively instructs candidates with spatial cues from augmented instance features; and (4) Dynamic Convolution network. Note that we also describe our **transformer-based** architecture in the supplementary material (Sec.3).

Kernel-based 3DIS Architecture. First, sparse convolutional U-Net backbone [15] takes a colored point cloud $P \in \mathbb{R}^{N_p \times 6}$ as input and voxelizes P into voxels to extract point-wise feature maps $F_p \in \mathbb{R}^{N_p \times D}$. Given the point features, the point aggregator samples a set of instance candidate features, referring to point positions based on the Farthest Point Sampling (FPS) [10]. Specifically, we employ two-stage point aggregator blocks to produce optimal instance features across 3D scenes. In the first stage, N'_k number of candidates $F'_{inst} \in \mathbb{R}^{N'_k \times D}$ are sampled from full-resolution feature F_p, while in the second stage, N_k ($< N'_k$) instance features $F_{inst} \in \mathbb{R}^{N_k \times D}$ are sampled based on the F'_{inst}. We then classify the category C_i for $i = \{1, 2, \ldots, N_c\}$ of each instance from F_{inst} via linear classification head f_{cls} using cross-entropy loss as follows:

$$\mathcal{L}_{cls} = -\mathbb{E}_{c,w_c \sim \mathbb{D}} \left[\sum_{r \in N_c} w_c[r] \log f_{\text{cls}}(F_{inst})[r] \right] \quad (3)$$

where w_c indicates one-hot encoded category labels and $\mathbb{D}$ represents (input) data distribution. Also, we predict instance kernels $K \in \mathbb{R}^{N_k \times D}$ based on F_{inst} through MLPs. Furthermore, each point-wise prediction head processes F_p to generate mask features $F_{mask} \in \mathbb{R}^{N_p \times D}$ and axis-aligned bounding boxes $F_{box} \in \mathbb{R}^{N_p \times 6}$ as auxiliary information for mask prediction. Finally, the dynamic convolution network produces the final instance mask $\hat{M}_f$ using instance kernels K and F_{mask}. To optimize $\hat{M}_f$, we formulate the following loss $\mathcal{L}_{mask}$ as a sum of the two losses:

$$\mathcal{L}_{mask} = \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}(M_f, \hat{M}_f) + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}(M_f, \hat{M}_f) \quad (4)$$

where M_f is the ground-truth mask, and $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{dice}}$ denote binary cross-entropy and Dice losses [8], respectively.

3.2. Instance-wise Knowledge Aggregation (IKA)

Recent studies [19, 20, 29, 31, 32, 39, 48] in 3DIS deal with instance-wise features F_{inst} as essential components, aiming to encode rich semantic and geometric information about instances. They spread hundreds of candidates over expansive 3D scenes to cover as many instances as possible; this number N_k far exceeds the total number N_i of instances present in any single scene (see Tab.4 in Supp.). Although these candidates provide practical information for mask prediction, they often represent separate parts of each instance,

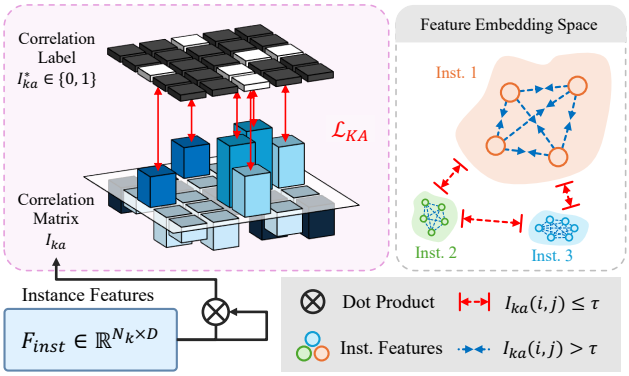


Figure 3. In IKA, we specify highly correlated instance candidates ($I_{ka} > \tau$) representing identical instances based on the correlation matrix I_{ka}. Then, we softly aggregate scattered clues of related candidates, optimizing I_{ka} with pseudo-binary label I^*_{ka}.

resulting in fragmented instance masks. With these insights, we focus on associating scattered knowledge to fully leverage the advantages of multiple instance candidates.

To handle multiple candidates, we first identify those representing the same instance based on their pairwise affinity. Specifically, we compute a correlation matrix $I_{ka} \in \mathbb{R}^{N_k \times N_k}$ using the dot product of F_{inst} and its transpose F^T_{inst}, then normalize I_{ka} to a range between 0 and 1. Here, we note that candidates representing the same instance yield relatively high correlation values. Thus, based on I_{ka}, we consider pairwise candidates with correlation values surpassing a predefined threshold τ as fragments from the same instance. Then, as shown in Fig, 3, we softly guide these candidates by optimizing their correlations in I_{ka}, motivated by the success of recent self-supervised learning strategies [2, 53]. We extend this approach to further enrich coherence among highly correlated candidate features while also encouraging the network to be aware of scattered cues for each instance. In particular, we dynamically construct a pseudo-binary correlation label matrix $I^*_{ka} \in \{0, 1\}^{N_k \times N_k}$, where elements of highly correlated pairs ($> \tau$) in I_{ka} are 1, while the rest ($\leqslant \tau$) are 0 for all candidate pairs. Given I_{ka} and I^*_{ka}, we formulate the element-wise correlation regularization loss $\mathcal{L}_{KA}$ as follows:

$$I_{ka} = F_{inst} \cdot F^T_{inst}, \;\; I^*_{ka}(i,j) = \begin{cases} 1 & \text{if } I_{ka}(i,j) > \tau \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

$$\mathcal{L}_{KA} = \sum_{I_{ka} > \tau} \| I^*_{ka}(i,j) - I_{ka}(i,j) \|_2 + w \cdot \sum_{I_{ka} \leqslant \tau} \| I^*_{ka}(i,j) - I_{ka}(i,j) \|_2 \quad (6)$$

where $\|\cdot\|_2$ denotes the L_2 norm, and w is the weight value used to balance between highly correlated pairs and others. In Eq. 6, the first term facilitates candidates representing the identical instance closer together in the latent space,

whereas the second term decreases correlations between unrelated candidates, thereby reducing confusion from irrelevant knowledge. In conclusion, IKA establishes practical connections that strengthen solidarity across multiple candidates encoding scattered cues by regularizing loss $\mathcal{L}_{KA}$.

3.3. Instance-wise Structural Guidance (ISG)

In real-world scenarios, 3D instances are usually placed near each other or overlapping, arranged in inconsistent patterns without following conventional standards. Also, these instances are represented as sparse point clouds, which often induce fuzzy inherent noises in point-wise features. Hence, the instance-wise features F_{inst} from these settings are unclear, causing models to misinterpret the spatial extent of each instance. In this work, we explicitly reduce such ambiguity of F_{inst} by leveraging intelligent structural guidance from clarity-enhanced instance-wise feature $\tilde{F}_{inst}$.

For accurate segmentation, it is crucial to capture the clear morphological shape of complex instances. From this intuition, we aim to encourage the model to better learn the structural appearance of instances. Thus, we implement effective structural guidance based on a truncated singular value decomposition (SVD) [14, 42] algorithm primarily used for noise reduction. As illustrated in Fig. 4, given $F'_{inst} \in \mathbb{R}^{N'_k \times D}$ from the first point aggregator, we initially decompose F'_{inst} into three matrices based on SVD: (1) $U \in \mathbb{R}^{N'_k \times N'_k}$, (2) $\Sigma \in \mathbb{R}^{N'_k \times D}$, and (3) $V \in \mathbb{R}^{D \times D}$, as:

$$F'_{inst} = U \cdot \Sigma \cdot V^T = \sum_{i}^{N'_k} u_i \cdot \sigma_i \cdot v_i^T \tag{7}$$

where $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_n)$ are the matrices with orthonormal columns, and $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix containing singular values, arranged in the descending order. Typically, vectors with higher singular values in Σ include meaningful information for object cognition, while the lower-rank vectors are regarded as less important, potentially causing ambiguity. Then, we preserve only the top t singular values from Σ, along with their corresponding singular vectors from U and V, facilitating an ambiguity-reduced yet information-rich representation of the original features F'_{inst} as follows:

$$\tilde{F}'_{inst} \in \mathbb{R}^{N'_k \times D} = \tilde{U} \cdot \tilde{\Sigma} \cdot \tilde{V}^T = \sum_{i}^{N'_k} \tilde{u}_i \cdot \tilde{\sigma}_i \cdot \tilde{v}_i^T \tag{8}$$

where $\tilde{U} \in \mathbb{R}^{N'_k \times t}$, $\tilde{\Sigma} \in \mathbb{R}^{t \times t}$, and $\tilde{V} \in \mathbb{R}^{D \times t}$ are truncated matrices using only t columns from U, Σ, and V, then $\tilde{F}'_{inst}$ is reconstructed from them. Based on both F'_{inst} and $\tilde{F}'_{inst}$, we sample instance-wise features F_{inst} and $\tilde{F}_{inst}$, respectively, employing the shared second stage point aggregator. Given F_{inst} and $\tilde{F}_{inst}$, we calculate the cross-correlation matrix $I_{sg} \in \mathbb{R}^{N_k \times N_k}$ and also generate a pseudo-binary correlation label $I^*_{sg} \in \{0,1\}^{N_k \times N_k}$, as in the IKA network.

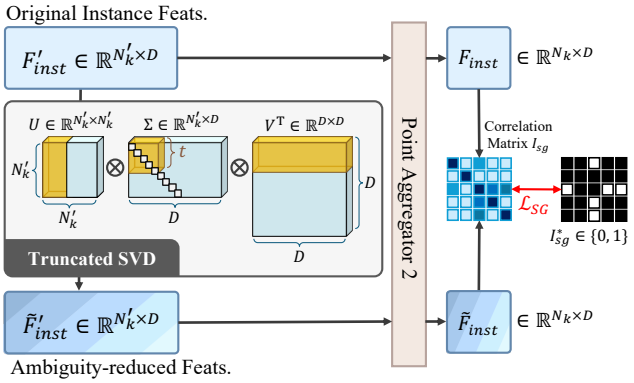


Figure 4. In ISG, we decompose F'_{inst} into three matrices (U, Σ, V) using the SVD algorithm and keep top t singular vectors to filter out less important noises. Then, we effectively enhance F_{inst} with highlighted essential cues from $\tilde{F}_{inst}$ by optimizing $\mathcal{L}_{SG}$.

We then effectively transfer structural clues from $\tilde{F}_{inst}$, utilizing the correlation regularization loss $\mathcal{L}_{SG}$ as follows:

$$I_{sg} = F_{inst} \cdot \tilde{F}^T_{inst}, \;\; I^*_{sg}(i,j) = \begin{cases} 1 & \text{if } I_{sg}(i,j) > \tau \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

$$\mathcal{L}_{SG} = \sum_{I_{sg} > \tau} \left\| I^*_{sg}(i,j) - I_{sg}(i,j) \right\|_2 + \tilde{w} \cdot \sum_{I_{sg} \leqslant \tau} \left\| I^*_{sg}(i,j) - I_{sg}(i,j) \right\|_2 \tag{10}$$

where $\tilde{w}$ represents the balancing hyper-parameter, and the role of each term in Eq. 10 follows Eq. 6. By optimizing $\mathcal{L}_{SG}$, ISG conditionally instructs instance features F_{inst} to incorporate spatial configuration details from clarity-enhanced features $\tilde{F}_{inst}$, improving discriminative power.

Loss Function. Finally, our full network minimizes $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{KA}\mathcal{L}_{KA} + \lambda_{SG}\mathcal{L}_{SG} \tag{11}$$

where each λ is a hyper-parameter decided via extensive grid searches to balance the impact of respective loss terms.

4. Experiments

4.1. Experimental Setup

Datasets. We train and evaluate the overall performance using four landmark benchmarks for 3D Instance Segmentation (3DIS): ScanNetV2 [7], ScanNet200 [38], S3DIS [1], and STPLS3D [5]. These four datasets provide 3D point cloud data collected in diverse real-world environments. Note that detailed descriptions of each dataset and all implementation details are provided in Supplementary Material.

Evaluation Metrics. We evaluate the 3DIS performance using the Average Precision (AP), a conventional metric in computer vision tasks. We report the mean average precision (mAP) across IoU (Intersection of Union) thresholds

Method	ScanNetV2 Val			ScanNetV2 Test		
	mAP	mAP_{50}	mAP_{25}	mAP	mAP_{50}	mAP_{25}
3D-SIS [21]	–	18.7	35.7	16.1	38.2	55.8
MTML [28]	20.3	40.2	55.4	28.2	54.9	73.1
PointGroup [24]	34.8	56.7	71.3	40.7	63.6	77.8
DyCo3D [19]	35.4	57.6	72.9	39.5	64.1	76.1
3D-MPA [12]	35.5	59.1	72.4	35.5	61.1	73.7
MaskGroup [54]	42.0	63.3	63.3	43.4	66.4	79.2
SoftGroup [44]	46.0	67.6	78.9	50.4	76.1	86.5
SSTNet [30]	49.4	64.3	74.0	50.6	69.8	78.9
Mask3D [39]	55.2	73.7	82.9	56.6	78.0	82.9
SPFormer [43]	56.3	73.9	82.9	54.9	77.0	85.1
Q-Former [31]	56.5	74.2	83.3	58.3	78.7	87.4
ISBNet [32]	56.8	73.3	81.3	55.9	76.3	87.0
MAFT [29]	58.4	75.9	84.5	59.6	78.6	86.0
EASE [37]	60.2	77.2	84.2	59.8	78.9	86.7
SphMask [41]	62.3	79.9	88.2	61.6	**81.2**	87.5
IKNE (Ours)	**62.9**	**81.8**	**88.7**	**61.7**	**81.2**	**88.0**

Table 1. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the ScanNetV2 [7] dataset.

Method	S3DIS Area 5				S3DIS 6-fold CV			
	AP	AP_{50}	$Prec_{50}$	Rec_{50}	AP	AP_{50}	$Prec_{50}$	Rec_{50}
3D-Bonet [50]	–	–	57.5	40.2	–	–	65.6	47.6
PointGroup [24]	–	57.8	61.9	62.1	–	64.0	69.6	69.2
3D-MPA [12]	–	–	63.1	58.0	–	–	66.7	64.1
DyCo3D [19]	–	–	64.3	64.2	–	–	–	–
MaskGroup [54]	–	65.0	62.9	64.7	–	69.9	66.6	69.6
SSTNet [30]	42.7	59.3	65.5	64.2	54.1	67.8	73.5	73.4
SoftGroup [44]	51.6	66.1	**73.6**	66.6	54.4	68.9	75.3	69.8
SPFormer [43]	–	66.8	72.8	67.1	–	69.2	74.0	71.1
ISBNet [32]	56.3	67.5	70.5	72.0	60.8	70.5	77.5	77.1
DKNet [48]	–	–	70.8	65.3	–	–	75.3	71.1
Mask3D [39]	56.6	68.4	68.7	66.3	64.5	75.5	72.8	74.5
Q-Former [31]	57.7	69.9	70.5	72.2	62.0	73.3	72.7	73.4
MAFT [29]	–	69.1	–	–	–	–	–	–
EASE [37]	59.0	71.6	69.4	68.7	65.3	76.4	73.6	74.6
SphMask [41]	60.5	72.3	71.3	76.3	64.0	72.3	78.1	77.7
IKNE (Ours)	**61.1**	**73.0**	72.0	**77.0**	**65.8**	**76.9**	**79.4**	**78.9**

Table 2. Comparison of 3DIS performance with state-of-the-art approaches on the S3DIS [1] Area 5 and 6-fold cross validation.

incremented by 5%, ranging from 50% to 95%. Also, we assess mAP_{50} and mAP_{25}, which measure accuracy based on IoU thresholds of 50% and 25%. For the S3DIS [1], we further provide mean precision (mPrec) and mean recall (mRec) at a 50% IoU threshold, following previous works.

4.2. Performance Comparison with SOTA Methods

In this section, we quantitatively compare our proposed framework IKNE with previous methods. Despite significant advances with high scores, these methods, regardless of architecture, still waste considerable instance candidates that may contain valuable information. To maximize the benefits of using numerous candidates, we present novel instance-wise knowledge aggregation (IKA) and guidance (ISG) networks. In Tab. 1, we first evaluate mean average precision (mAP) with different IoU thresholds across 18 classes on the ScanNetV2 [7] dataset. IKNE generally outperforms prior works, achieving new state-of-the-art scores in terms of mAP, mAP_{50}, and mAP_{25} on both validation and hidden test sets. While our method slightly exceeds the latest coarse-to-fine approach, Spherical Mask [41], ours transcends our primary baselines, which rely on hundreds of instance candidates, such as ISBNet [32] (kernel-based) and MAFT [29] (transformer), by large margins across all metrics. In Tab. 2, we assess 3DIS performance on Area 5 and the 6-fold cross-validation set of the S3DIS [1] dataset. For Area 5 evaluation, we train with Areas 1 to 6, excluding 5, and validate with Area 5; for 6-fold cross-validation, we average the validation scores across all 6 areas. IKNE demonstrates impressive performance in both evaluations, reaching 61.1 / 73.0 (Area 5) and 65.8 / 76.9 (6-fold) in mAP / mAP_{50}. Also, in Tab. 3, ours achieves the highest scores on the ScanNet200 [38] dataset containing fine-grained 200 categories. Notably, as shown in Tab. 4, IKNE surpasses existing methods on the STPLS3D [5] dataset, including outdoor 3D scenes, with improvements of up to +1.5 / 2.0 / 2.0 in mAP / mAP_{50} / mAP_{25}. These results highlight that we successfully handle instance candidates, improving their informativeness for understanding real-world instances from both indoor and outdoor environments. Note that "-" in each table indicates unreported scores from published papers.

Method	mAP	mAP_{50}	mAP_{25}
SPFormer [43]	25.2	33.8	39.6
Mask3D [39]	27.4	37.0	42.3
Q-Former [31]	28.1	37.1	43.4
MAFT [29]	29.2	38.2	43.3
IKNE (Ours)	**29.9**	**38.9**	**44.9**

Table 3. Comparison of 3DIS performance with state-of-the-art approaches on the ScanNet200 [38] validation (with 200 classes).

Method	mAP	mAP_{50}	mAP_{25}
PointGroup [24]	23.3	38.5	48.6
HAIS [6]	35.1	46.7	52.8
SoftGroup [44]	47.3	63.1	71.4
Mask3D [39]	63.4	79.2	85.6
IKNE (Ours)	**64.9**	**81.2**	**87.6**

Table 4. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the STPLS3D [5] test dataset.

Method (**Kernel-based**)			ScanNet Val	S3DIS Area 5
Baseline (K)	IKA	ISG	mAP / mAP$_{50}$	mAP / mAP$_{50}$
✓	-	-	56.8 / 73.3	56.3 / 67.5
✓	✓	-	62.7 / 80.8	60.4 / 72.6
✓	-	✓	62.8 / 81.1	60.6 / 71.9
✓	✓	✓	**62.9 / 81.8**	**61.1 / 73.0**

Table 5. Ablation study to see the effect of two main modules (IKA and ISG) based on the kernel-based pipeline (K).

Method (**Transformer**)			ScanNet Val	S3DIS Area 5
Baseline (T)	IKA	ISG	mAP / mAP$_{50}$	mAP / mAP$_{50}$
✓	-	-	58.4 / 75.9	56.6 / 68.4
✓	✓	-	59.1 / 76.2	58.8 / 70.7
✓	-	✓	59.4 / 77.2	58.4 / 70.1
✓	✓	✓	**60.8 / 77.9**	**59.0 / 71.6**

Table 6. Ablation study to see the effect of two main modules (IKA and ISG) based on the transformer-based pipeline (T).

Method			**Kernel-based**	**Transformer**
Baseline	IKA	ISG	Average IoU	Average IoU
✓	-	-	0.8627	0.8486
✓	✓	-	0.8828	0.8728
✓	-	✓	0.8825	0.8730
✓	✓	✓	**0.8904**	**0.8803**

Table 7. Ablation study to see the effect of two main modules across both pipelines, using IoU scores on the ScanNetV2 [7] val.

4.3. Ablation Studies

Effect of IKA and ISG Networks. We evaluate variants of our method w/ and w/o IKA and ISG modules based on kernel (K) and transformer-based (T) architectures. As shown in Tab. 5 and 6, the addition of each module surpasses the score of baseline models [29, 32, 39] across all experiments, regardless of model structure. Specifically, IKA boosts performance, with gains of up to +5.9 / 7.5 for ScanNetV2 [7] and +4.1 / 5.1 for S3DIS [1] in mAP / mAP$_{50}$, underscoring the importance of integrating scattered clues from multiple candidates. Also, ISG contributes to the spatial understanding of instance candidates via intelligent structural guidance from clarity-enhanced features, improving up to +6.0 / 7.8 and +4.3 / 4.4. Eventually, utilizing all modules together gains further advances, resulting in the best performance with notable progress on both structures. These results demonstrate each module plays a meaningful role in improving the discriminative capability of instance features.

Furthermore, to assess how each module influences the accurate determination of instance mask coverage, we use the IoU metric, which is closely associated with segmentation performance. IoU calculates the ratio of the intersection to the union of predicted and ground truth masks. In Tab. 7, we compare the average IoU of our predicted instance masks and those of the baselines for both architectures. As shown in Fig. 1 (a) and Fig.1 (a) (in Supp.), for previous works, a single instance is often separated into several fragments. These fragments potentially cause low IoU scores. The IoU improvements with IKA across both architectures confirm its effectiveness in reducing the negative impact of multiple fragments and enhancing the accuracy of mask coverage. Also, the IoU tends to be low, particularly when instance masks include the surrounding backgrounds or when adjacent instances are not clearly distinguished, as in Fig. 1 (b) and Fig.1 (b) (in Supp.), leading to overestimated or underestimated instance masks. In addition to IKA, ISG enhances the IoU scores, verifying our guidance is sufficiently practical. We further validate the significance of ours with additional experiments in Supp. (see Sec.4).

Architecture	Threshold τ	ScanNet Val	S3DIS Area 5
		mAP / mAP$_{50}$	mAP / mAP$_{50}$
Baseline (K)	-	56.8 / 73.3	56.3 / 67.5
Kernel-based	0.6	53.6 / 73.4	55.3 / 65.5
Kernel-based	0.7	57.2 / 76.4	59.1 / 69.2
Kernel-based	0.8	61.1 / 79.9	60.7 / 72.3
Kernel-based	0.9	**62.9 / 81.8**	**61.1 / 73.0**
Baseline (T)	-	58.4 / 75.9	56.6 / 68.4
Transformer	0.6	57.9 / 75.8	56.3 / 67.5
Transformer	0.7	59.2 / 76.2	57.4 / 68.6
Transformer	0.8	59.9 / 77.1	58.6 / 70.4
Transformer	0.9	**60.8 / 77.9**	**59.0 / 71.6**

Table 8. Ablation study to investigate the threshold τ for determining instance-wise candidates representing the same instance.

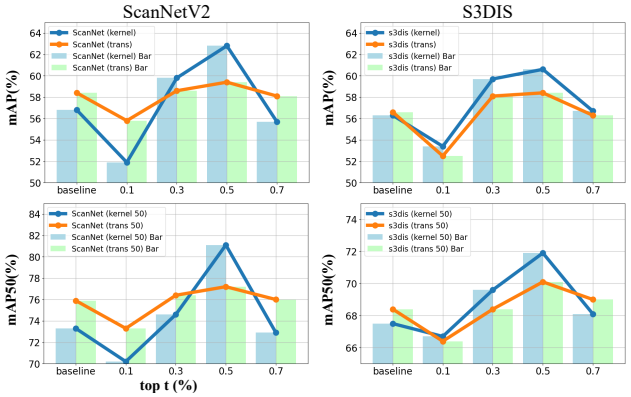


Figure 5. Comparison to explore the optimal top t for reducing ambiguity while keeping important cues of instance features.

Thresholds τ for Instance-wise Identification. We specify candidates representing the same instance using predefined threshold τ from the correlation matrices I_{ka} and I_{sg}. In Tab. 8, based on kernel (K) and transformer (T) architectures [29, 32, 39], we explore the impact of τ for useful knowledge interaction. When τ is lower than 0.7, the scores degrade due to the inclusion of inaccurate information from different instance representative features. Yet, with more precise identification using the higher τ ($\geqslant 0.8$), our knowledge interaction yields considerable improvements. This result emphasizes that clues from accurately identified candidates are precious for enhancing instance interpretation. We

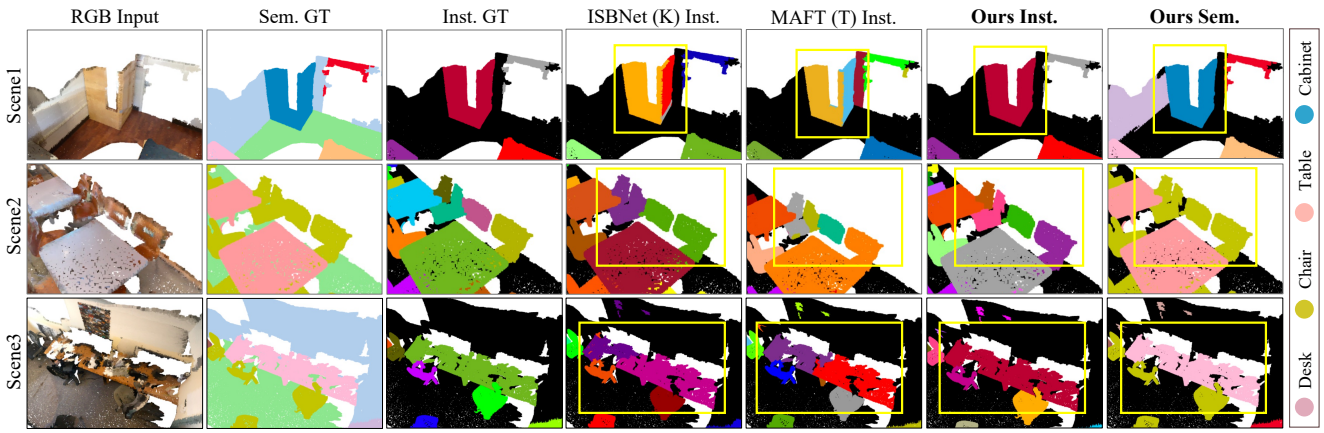


Figure 6. Qualitative comparison of instance (Inst.) and semantic (Sem.) masks between the baselines (ISBNet (K) [32], MAFT (T) [29]) and ours with Ground Truth masks on the ScanNetV2. The critical differences are highlighted using yellow boxes. Best viewed in color.

also examine τ for the STPLS3D [5] in Supp. (Sec.4.3).

Analysis of the Top t values. In this work, we truncate the top t vectors of (U, Σ, V) from F'_{inst} and reconstruct $\tilde{F}'_{inst}$ using $(\tilde{U}, \tilde{\Sigma}, \tilde{V})$ based on a truncated SVD algorithm [17]. Here, the t value impacts the balance between data compression and information preservation. Therefore, it is crucial to find the optimal value of t for proper balance. To this end, we conduct experiments using various t for both kernel (blue) and transformer (orange) architectures in Fig. 5. We represent t as a percentage, so the x-axis indicates the proportion of vectors retained during truncation across three matrices. We observe that the trends of all experiments are similar. In settings with a lower t percentage (0.1), the scores consistently decrease due to large information loss. However, with a higher t (0.7), the effect of noise reduction is insufficient; thereby, the performance is comparable to baselines [29, 32]. We ultimately set t as **0.5**, resulting in the highest accuracy while allowing us to preserve meaningful information and effectively filter out inherent noises.

4.4. Qualitative Results

Visual Comparison. In Fig. 6, we qualitatively validate the effectiveness of our novel framework IKNE. We visualize the predicted semantic (Sem.) and instance (Inst.) masks of the kernel (K) [32] and transformer (T) [29] baselines with ours on the ScanNetV2 [7]. We also provide the corresponding ground truth for solid comparisons. In Scene 1, ours accurately segments the *cabinet* as a single instance, whereas the baselines separate it into multiple fragments. In Scene 2, ours correctly distinguishes neighboring instances (four *chairs* and a *table*), while baselines erroneously recognize two *chairs* as one (K) or a *table* and a *chair* as one (T). Also, in Scene 3, where objects are disorderly adjacent, ours clearly captures the spatial range of the *desk* and *chairs* compared to the baselines. These qualitative results verify that our instance features contain highly informative cues.

Structural Guidance from $\tilde{\mathbf{F}}_{\mathbf{inst}}$. In Fig. 7, we qualitatively confirm the significance of structural cues from the noise-reduced features $\tilde{F}_{inst}$ in smartly guiding the original features F_{inst}. As shown in Scene 1 and 2, instance features (red) from $\tilde{F}_{inst}$ accurately capture the spatial range of instances (*chair* and *desk*), whereas instance features (blue) from F_{inst} struggle to cover the complete coverage. Also, in Scene 3, where the distinction is unclear (blue) from surroundings (*curtain* and *wall*), the candidate (red) from $\tilde{F}_{inst}$ clearly segments such instances. These findings highlight guidance from clarity-enhanced features is worthwhile.

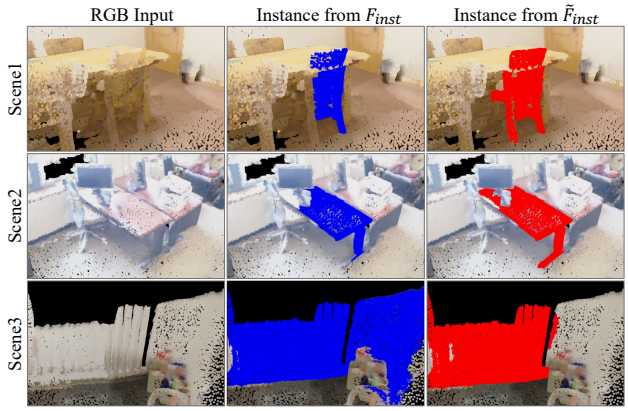


Figure 7. Instance feature visualizations of the original features F_{inst} (blue) and ambiguity-reduced features $\tilde{F}_{inst}$ (red).

5. Conclusion

In this paper, we introduce a novel instance-wise knowledge enhancement approach, IKNE, for 3D instance segmentation. We focus on optimizing the efficiency of hundreds of instance candidates by effectively handling those representing the same instance. To address these candidates, we propose an IKA to integrate spread clues based on correlation regularization. Also, we present an ISG, which enhances instance features with fundamental cues for object understanding from ambiguity-reduced instance features using the truncated SVD. Our extensive experiments demonstrate the effectiveness of our methods, achieving new state-of-the-art results on both kernel and transformer architectures.

Acknowledgement

This work was supported by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2023-00227409, Development of sketch-based semantic 3D modeling technology for creating user-centric Metaverse content spaces for indoor spaces, 90%; RS-2024-00345025, International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, 5%), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), 1%; IITP-2025-RS-2024-00436857, 2%), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00521602, 2%), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2, 5, 6, 7

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 4

[3] Ori Bryt and Michael Elad. Compression of facial images using the k-svd algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282, 2008. 3

[4] Chin-Chen Chang, Piyu Tsai, and Chia-Chen Lin. Svd-based digital image watermarking scheme. *Pattern Recognition Letters*, 26(10):1577–1586, 2005. 3

[5] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022. 2, 5, 6, 8

[6] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 1, 2, 6

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5, 6, 7, 8

[8] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *Proceedings of the European conference on computer vision (ECCV)*, pages 562–578, 2018. 4

[9] Renwei Dian, Shutao Li, and Xudong Kang. Regularizing hyperspectral and multispectral image fusion by cnn denoiser. *IEEE transactions on neural networks and learning systems*, 32(3):1124–1135, 2020. 3

[10] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997. 4

[11] Cathrin Elich, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. 3d bird's-eye-view instance segmentation. In *German Conference on Pattern Recognition*, pages 48–61. Springer, 2019. 2

[12] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 2, 6

[13] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3722–3731, 2021. 2

[14] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*, pages 134–151. Springer, 1971. 2, 3, 5

[15] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 4

[16] Qiang Guo, Caiming Zhang, Yunfeng Zhang, and Hui Liu. An efficient svd-based method for image denoising. *IEEE transactions on Circuits and Systems for Video Technology*, 26(5):868–880, 2015. 3

[17] Per Christian Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27:534–553, 1987. 3, 8

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[19] Tong He, Chunhua Shen, and Anton Van Den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 354–363, 2021. 1, 2, 3, 4, 6

[20] Tong He, Wei Yin, Chunhua Shen, and Anton Van den Hengel. Pointinst3d: Segmenting 3d instances by points. In *European Conference on Computer Vision*, pages 286–302. Springer, 2022. 1, 2, 3, 4

[21] 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*, pages 4421–4430, 2019. 1, 2, 6

[22] Peg Howland, Moongu Jeon, and Haesun Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003. 3

[23] Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: Structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[24] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 1, 2, 6

[25] XiaoBing Kang and ShengMin Wei. Identifying tampered regions using singular value decomposition in digital image forensics. In *2008 International conference on computer science and software engineering*, pages 926–930. IEEE, 2008. 3

[26] Sungjune Kim, Hyung-gun Chi, Hyerin Lim, Karthik Ramani, Jinkyu Kim, and Sangpil Kim. Higher-order relational reasoning for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15251–15260, 2024. 3

[27] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980. 3

[28] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019. 6

[29] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3693–3703, 2023. 1, 2, 3, 4, 6, 7, 8

[30] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2783–2792, 2021. 6

[31] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18516–18526, 2023. 1, 2, 4, 6

[32] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 2, 3, 4, 6, 7, 8

[33] HS Prasantha, HL Shashidhara, and KN Balasubramanya Murthy. Image compression using svd. In *International conference on computational intelligence and multimedia applications (ICCIMA 2007)*, pages 143–145. IEEE, 2007. 3

[34] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 2

[35] Ajit Rajwade, Anand Rangarajan, and Arunava Banerjee. Image denoising using the higher order singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):849–862, 2012. 3

[36] Zhongzheng Ren, Ishan Misra, Alexander G Schwing, and Rohit Girdhar. 3d spatial recognition without spatially labeled 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13213, 2021. 2

[37] Wonseok Roh, Hwanhee Jung, Giljoo Nam, Jinseop Yeom, Hyunje Park, Sang Ho Yoon, and Sangpil Kim. Edge-aware 3d instance segmentation network with intelligent semantic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20644–20653, 2024. 2, 3, 6

[38] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 2, 5, 6

[39] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 1, 2, 3, 4, 6, 7

[40] Qiquan Shi, Yiu-Ming Cheung, and Jian Lou. Robust tensor svd and recovery with rank estimation. *IEEE Transactions on Cybernetics*, 52(10):10667–10682, 2021. 3

[41] Sangyun Shin, Kaichen Zhou, Madhu Vankadari, Andrew Markham, and Niki Trigoni. Spherical mask: Coarse-to-fine 3d point cloud instance segmentation with spherical representation. *arXiv preprint arXiv:2312.11269*, 2023. 3, 6

[42] Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993. 2, 3, 5

[43] 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*, pages 2393–2401, 2023. 1, 2, 6

[44] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 6

[45] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 1, 2

[46] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[47] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2

[48] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *European Conference on Computer Vision*, pages 235–252. Springer, 2022. 1, 2, 4, 6

[49] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Dening Lu, Mingqiang Wei, and Jun Wang. Venet: Voting enhancement network for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3712–3721, 2021. 2

[50] 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. *Advances in neural information processing systems*, 32, 2019. 1, 2, 6

[51] Dan Yang, Zongming Ma, and Andreas Buja. A sparse singular value decomposition method for high-dimensional data. *Journal of Computational and Graphical Statistics*, 23 (4):923–942, 2014. 3

[52] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 1, 2

[53] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 2, 4

[54] Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 6