UNIGEO: A UNIFIED 3D INDOOR OBJECT DETECTION FRAMEWORK INTEGRATING GEOMETRY-AWARE LEARNING AND DYNAMIC CHANNEL GATING

Xing Yi[1], *Jinyang Huang*[⋆1], *Feng-Qi Cui*[2], *Anyang Tong*[1], *Ruimin Wang*[1], *Liu Liu*[1], *Dan Guo*[1]

[1.] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[2.] Institute of Advanced Technology, University of Science and Technology of China, Hefei, China

ABSTRACT

The growing adoption of robotics and augmented reality in real-world applications has driven considerable research interest in 3D object detection based on point clouds. While previous methods address unified training across multiple datasets, they fail to model geometric relationships in sparse point cloud scenes and ignore the feature distribution in significant areas, which ultimately restricts their performance. To deal with this issue, a unified 3D indoor detection framework, called UniGeo, is proposed. To model geometric relations in scenes, we first propose a geometry-aware learning module that establishes a learnable mapping from spatial relationships to feature weights, which enabes explicit geometric feature enhancement. Then, to further enhance point cloud feature representation, we propose a dynamic channel gating mechanism that leverages learnable channel-wise weighting. This mechanism adaptively optimizes features generated by the sparse 3D U-Net network, significantly enhancing key geometric information. Extensive experiments on six different indoor scene datasets clearly validate the superior performance of our method.

Index Terms— 3D Indoor object detection, 3D scenes understanding, computer vision, scene perception

1. INTRODUCTION

Indoor 3D object detection is a core task in scene understanding with applications in robotics, pose estimation [1] and AR/VR systems. Traditional methods [2–14] struggle to generalize, relying heavily on specific datasets and failing to build a unified learning framework. This limitation stems from their inability to decouple geometric features from semantic information, leading models to overfit dataset-specific surface characteristics rather than capturing essential object attributes.

To overcome these limitations, the State-of-the-art (SOTA) method proposed UniDet3D [15], a unified multi-dataset 3D detection framework that allows collaborative training across multiple data sources, enhancing generalization and building a basis for generic 3D detection. However, this method adopted a sparse 3D U-Net backbone with significant shortcomings, i.e., it processed spatial coordinates and voxel characteristics independently, leading geometric information to a decline in deeper layers and lacking explicit spatial connection modeling. Additionally, this method also failed to adapt to non-uniform point cloud distributions, challenging to detect geometrically different areas like edges and corners. These limitations jointly hinder the model's performance in complicated 3D scene interpretation tasks.

⋆ Corresponding author: hjy@hfut.edu.cn

In this paper, we propose UniGeo, a unified 3D indoor detection framework that incorporates geometric-aware learning and a dynamic channel gating mechanism. Specifically, we introduce a geometric-aware learning module to model scene geometric relationships, which establishes a mapping from spatial scene geometric relationships to feature weights, allowing for explicit feature improvement based on 3D geometric structures. Firstly, we employ Euclidean distance to model the geometric topology of point clouds in the scene, and then assign differentiated weights to point cloud features using an exponential decay function based on the topological information, thereby enhancing focus on significant regions. Then, to further address the issue of the backbone network modeling point cloud's limited spatial distribution capability, we suggest a dynamic channel gating mechanism that adaptively learns the channel weight to help modulate the characteristic response strength on various channels, efficiently enhance the local characteristic information, and suppress the background noise information.

To evaluate the effectiveness and generalization capability of UniGeo, we conduct detailed experiments on six indoor scene datasets: ScanNet [16], S3DIS [17], MultiScan [18], 3RScan [19], ScanNet++ [20], and ARKitScenes [21]. The results demonstrate that our method achieves state-of-the-art performance, outperforming existing mainstream approaches in terms of both best and average performance on most datasets. Furthermore, through a series of ablation studies on model components, algorithm selection comparisons, and hyperparameter analysis, we thoroughly validate the superiority of UniGeo. The main contributions are as follows:

- To model geometric relationships in sparse point cloud scenes and achieve high detection accuracy, a unified 3D indoor detection framework, UniGeo, is proposed. By introducing a geometric-aware learning and a dynamic channel gating mechanism to establish a learnable mapping from spatial relationships to feature weights, we effectively handle non-uniformly distributed scene point clouds and achieve explicit feature enhancement based on 3D scene geometric structures.
- Extensive evaluation experiments across various metrics on six indoor scene detection datasets consistently demonstrate the superior performance of our proposed method.

2. METHOD

2.1. Approach Overview

The overall architecture of our method is depicted in Fig. 1. The input point cloud typically consists of N points, which can be represented as $\mathcal{P} = (p_i | i = 1, ..., N) \in \mathbb{R}^{N \times 6}$. Each point has coordinates x, y, z and colors r, g, b. Following previous methods, we first take the point cloud $\mathcal{P}$ as the input of the geometry-aware

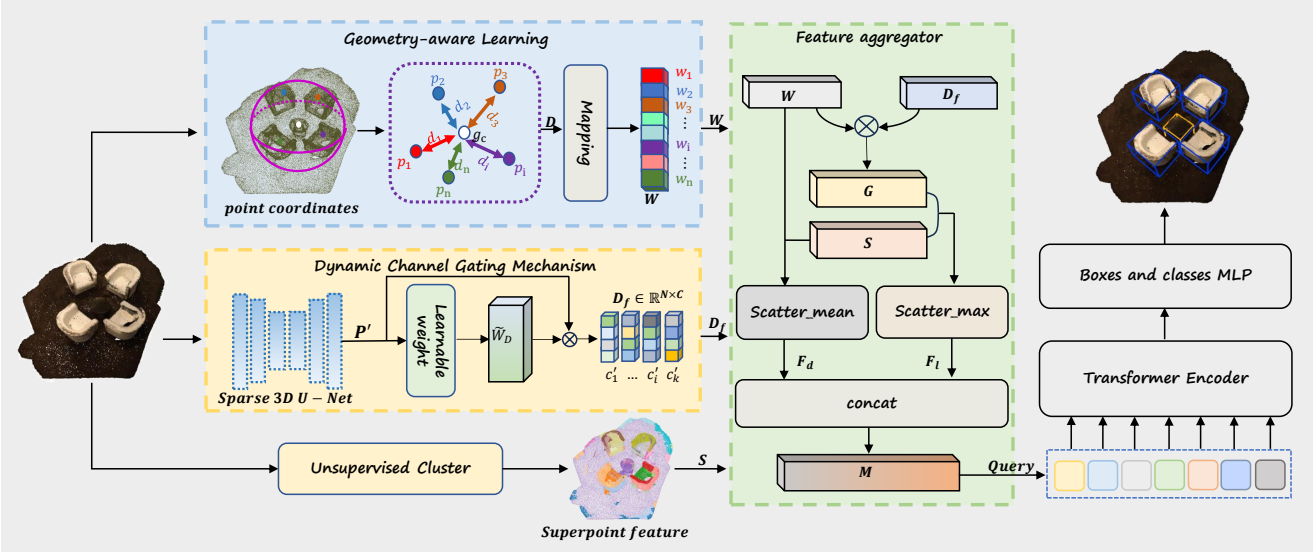

Fig. 1. Overview of our method. UniGeo takes a point cloud as input and pass through a geometry-aware learning module and a dynamic channel gating mechanism to generate geometry weights and channel features. A feature aggregator combines these with superpoint features into a hybrid representation, which then serves as input queries to a transformer encoder. Finally, a box MLP and a class MLP predict the 3D bounding boxes from the transformer encoder outputs.

learning module and the dynamic channel gating mechanism and obtain their spatial weighting coefficients $\mathcal{W} = \{w_i | i = 1, ..., N\} \in \mathbb{R}^{N\times 1}$ and channel feature $\mathcal{D}_f = \{d_f^i | i = 1, ..., N\} \in \mathbb{R}^{N\times C}$, respectively. Meanwhile, we obtain the superpoint feature $S = (s_i | i = 1, ..., N) \in \mathbb{R}^{M\times C}$ from the point cloud by unsupervised clustering. Then, we aggregate these features through a feature aggregator and obtain a hybrid representation $\mathcal{M}$. Finally, the hybrid representation $\mathcal{M}$ serve as queries for the transformer encoder. After processing through the transformer encoder, the output is fed into boxes and classes MLP to predict the detection boxes and classes results.

2.2. Geometry-aware Learning

A Geometry-Aware Learning(GAL) module that models spatial geometric structures is first proposed and to establish a weight mapping between topological information and point cloud data. By adaptively weighting scene features, the module enhances the model's focus on significant regions. Specifically, we take the 3D coordinates $\mathcal{P} = \{p_i = (x_i, y_i, z_i) | i = 1, ..., N\} \in \mathbb{R}^{N\times 3}$ as input of the geometry-aware learning module, and we further compute the geometric centroid g_c as the reference point to model geometric structures. In particular, g_c can be calculated as:

$$g_c = (\bar{x}, \bar{y}, \bar{z}) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i, \frac{1}{N}\sum_{i=1}^{N} y_i, \frac{1}{N}\sum_{i=1}^{N} z_i \right). \quad (1)$$

To model the importance and geometric relationships of different regions in scenes, we compute the Euclidean distance from each sampled point $\mathcal{P}' = \{p_i' | i = 1, ..., N\} \in \mathbb{R}^{N\times 3}$ to the geometric centroid g_c, thereby learning geometric feature representations. The Euclidean distance $\mathcal{D} = \{d_i | i = 1, ..., N\} \in \mathbb{R}^{N\times 1}$ can be represented as:

$$d_i = \|p_i - g_c\|_2 = \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (z_i - \bar{z})^2}. \quad (2)$$

To preserve the consistency of the point cloud topological structure, we employ min-max scaling normalize the Euclidean distance representation $\mathcal{D}$, thus obtaining the normalized spatial distance feature $\tilde{\mathcal{D}} = \{\tilde{d}_i | i = 1, ..., N\} \in \mathbb{R}^{N\times 1}$.

To enhance the global geometric structure of the scene, we employ an exponential decay function to process the normalized feature $\tilde{\mathcal{D}}$. This constructs a mapping between the spatial topological structure and the scene point cloud data, and ultimately obtains the spatial weighting coefficients $\mathcal{W} = w_i \mid i = 1, \ldots, N \in \mathbb{R}^{N\times 1}$ for each sample point. The weighting coefficients $\mathcal{W}$ are calculated as:

$$w_i = \exp(-\alpha \cdot \tilde{d}_i), \quad (3)$$

where the hyperparameter α denotes the exponential decay coefficient, and we set this hyperparameter to $\alpha = 2$.

2.3. Dynamic Channel Gating Mechanism

We propose a dynamic channel gating mechanism that employs learnable channel weights to adaptively optimize point cloud features generated by the sparse 3D U-Net network. This approach enhances key geometric information while effectively suppressing background noise and irrelevant features. Specifically, we first define a trainable weight vector $\tilde{\mathcal{W}}_{\mathcal{D}} = \left[w_1^d, w_2^d, ..., w_c^d\right] \in \mathbb{R}^C$ and initialize as $w_c^d = 0.1$. Then, the learned weight parameters are transformed into channel gating coefficients through a sigmoid function, where these coefficients represent the importance of each channel. To enhance geometric feature channels (e.g., edges) and suppress irrelevant background information, we combines the learned gating coefficients with the point features $\mathcal{P}' = \{p_i' | i = 1, ..., N\} \in \mathbb{R}^{N\times C}$ generated for each point by the sparse 3D U-Net. The channel feature $\mathcal{D}_f = \{d_f^i | i = 1, ..., N\} \in \mathbb{R}^{N\times C}$ can be calculated as:

$$\mathcal{D}_f = \left[w_i^d \odot p_i' | i = 1, 2, ..., N \right] \in \mathbb{R}^{N\times C}, \quad (4)$$

where C is the number of channel.

Method	Venue	ScanNet		ARKitScenes		S3DIS		MultiScan		3RScan		ScanNet++	
		mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
Best Results													
FCAF3D [14]	ECCV'2022	71.5	57.3			66.7	45.9	53.8	40.7	60.1	42.6	22.3	11.4
TR3D [13]	ICIP'2023	72.9	59.3			74.5	51.7	56.7	42.3	62.3	45.4	26.2	14.5
SPGroup3D [9]	AAAI'2024	74.3	59.6			69.2	47.2						
V-DETR [12]	ICLR'2024	77.4	65.0										
SOFW [4]	TMM'2025	70.9	52.3										
UniDet3D [15]	AAAI'2025	77.5	**66.1**	**61.3**	47.1	75.2	60.8	64.2	51.6	64.7	48.6	**26.4**	17.2
UniGeo (Ours)	-	**77.7**	65.6	60.0	**47.7**	**80.5**	**71.8**	**69.6**	**56.3**	**68.1**	**56.1**	25.6	**19.1**
Average across 25 trials													
FCAF3D [14]	ECCV'2022	70.7	56.0			64.9	43.8	52.5	39.2	59.6	40.4	21.4	11.0
TR3D [13]	ICIP'2023	72.0	57.4			72.1	47.6	55.0	41.2	61.5	44.2	24.3	13.9
SPGroup3D [9]	AAAI'2024	73.5	58.3			67.7	43.6						
V-DETR [12]	ICLR'2024	76.8	64.5										
UniDet3D [15]	AAAI'2025	**77.1**	**65.2**	**60.2**	46.0	73.3	57.9	62.4	50.8	62.1	45.6	24.4	16.3
UniGeo (Ours)	-	76.8	64.5	59.3	**46.8**	**77.9**	**66.3**	**65.9**	**53.5**	**64.2**	**49.7**	**24.6**	**17.8**

Table 1. Comparison with state-of-the-arts methods on 6 indoor scenes datasets, UniGeo demonstrates superior performance. Best results indicate the model's best evaluation results, Average across 25 trials represents the average evaluation score obtained by training 5 times and test each trained model 5 times independently.

2.4. Feature Aggregation

We use unsupervised clustering to convert the point cloud $\mathcal{P} = [p_i|i = 1, ..., N] \in \mathbb{R}^{N\times 6}$ into superpoint [22] features $S \in \mathbb{R}^{M\times C}$ according to the principles of UniDet3D [15]. Then, a hybrid feature fusion approach is proposed for the feature aggregation step that combines the benefits of geometry-aware feature $\mathcal{G}$ and dynamic channel gating features $\mathcal{D}_f$ to efficiently incorporate multi-dimensional data. Particularly, we obtain the geometry-aware features $\mathcal{G}$ through element-wise multiplication between the weight parameters $\mathcal{W} = \{w_i|i = 1, ..., N\} \in \mathbb{R}^{N\times 1}$ and the dynamic channel gating features $\mathcal{D}_f = \{d_f^i|i = 1, ..., N\} \in \mathbb{R}^{N\times C}$, achieving spatially adaptive feature recalibration. Geometry-aware features $\mathcal{G} = [g_i|i = 1, ..., N] \in \mathbb{R}^{N\times C}$ can be calculated as:

$$g_i = (w_i \odot d_f^i|i = 1, 2, ...N). \tag{5}$$

Then, we aggregate the Geometry-aware features $\mathcal{G}$ with the superpoints feature $S = [s_i|i = 1, ..., N] \in \mathbb{R}^{M\times C}$ by using a scattered mean to obtain the global geometric feature representation F_d, which can be expressed as:

$$\mathcal{F}_d = scatter\_mean\left[s_i, g_i|i = 1, 2, ..., M\right]) \in \mathbb{R}^{M\times C}. \tag{6}$$

Meanwhile, we aggregate the dynamic channel gating features $\mathcal{D}_f = \{d_f^i|i = 1, ..., N\} \in \mathbb{R}^{N\times C}$ with the superpoints feature $S = [s_i|i = 1, ..., N] \in \mathbb{R}^{M\times C}$ using a scattered max to obtain most discriminative local feature F_l, which is denoted by:

$$\mathcal{F}_l = scatter\_max\left[s_i, d_i|i = 1, 2, ..., M\right]) \in \mathbb{R}^{M\times C}. \tag{7}$$

Subsequently, The hybrid feature representation$\mathcal{M}$ is obtained by concatenating the global geometric feature $\mathcal{F}_d$ with the local feature $\mathcal{F}_l$. $\mathcal{M}$ remains sensitive to the overall geometric layout while preserving critical local details. $\mathcal{M}$ can be calculated as:

$$\mathcal{M} = concat\left[F_d, F_l\right]) \in \mathbb{R}^{M\times C}. \tag{8}$$

2.5. Transformer Encoder and Loss

Following the stage of feature aggregation, the backbone features are processed by a transformer encoder network. This network takes $\mathcal{M}$ queries as input and outputs $\mathcal{M}$ object proposal features. Besides, the transformer architecture relies solely on self-attention mechanisms among the input queries. For the matching strategy and training objective, we adopt the same approach as used in UniDet3D [15]. The total loss $\mathcal{L}_{total}$ can be defined as:

$$\mathcal{L}_{total} = \beta L_{cls} + L_{reg}, \tag{9}$$

where the classification loss $\mathcal{L}_{cls}$ is computed with a cross-entropy loss. For each matched proposal-ground truth pair, the regression loss $\mathcal{L}_{reg}$ is computed as a DIoU loss between the predicted and ground truth bounding boxes. β is the weight hyperparameter for $\mathcal{L}_{cls}$, and we set it to 0.5 based on experience.

3. EXPERIMENTAL

3.1. Experimental Setup

Datasets and metrics: We conduct our experiments on six real-world indoor detection datasets, including ScanNet [16], S3DIS [17], ARKitScenes [21], ScanNet++ [20], 3RScan [19], and MultiScan [18]. The specific data categories and detailed information are provided in Table 2. We use mean average precision (mAP) with thresholds of 0.25 and 0.5 as the evaluation metrics. To obtain statistically significant results, we run training 5 times and test each trained model 5 times independently. Furthermore, we report the best and average metrics across 5 × 5 trials to ensure consistency with the experimental standards of other methods.

Dataset Name	#Train Scenes	#Val Scenes	#Classes
ScanNet	1201	312	18
ARKitScenes	4493	549	17
S3DIS	204	68	5
MultiScan	174	42	17
3RScan	385	47	18
ScanNet++	230	50	84
Overall	6687	1068	99

Table 2. Quantitative statistics of indoor datasets in our mixture.

Parameter settings: We implement our method in the MMDetection3D [23] framework. The AdamW optimizer is used with an initial learning rate of 0.0002, a weight decay of 0.05, a batch size of 1, and a polynomial scheduler with a base of 0.9 for 1024 epochs. All experiments are conducted on four RTX 2028 Ti GPUs.

3.2. Analysis of Results

We compare UniGeo with the latest methods on six datasets: ScanNet [16], S3DIS [17], MultiScan [18], 3RScan [19], and ScanNet++ [20]. The quantitative results are presented in Table 1, and qualitative results are shown in Fig. 2. In the best results, our method achieves state-of-the-art performance on most datasets. Particularly, our approach attains gains of **5.3%** and **11%** in mAP_{25} and mAP_{50} over UniDet3D [15] on the S3DIS [17] dataset. In the average across

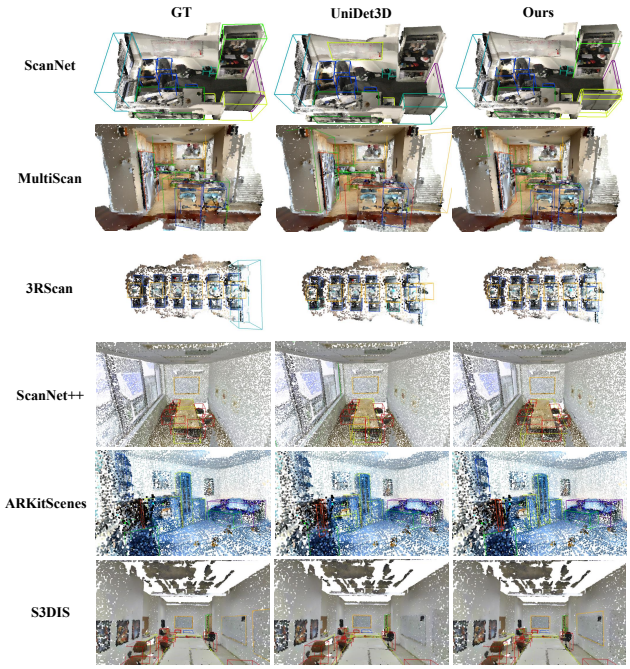

Fig. 2. Qualitative comparison on six indoor scenes datasets.

25 trials setting, our method maintains strong performance and leads by **4.6%** and **8.4%** on the S3DIS [17] dataset. Our method's superior performance stems from its enhanced focus on perceiving and learning critical scene features, coupled with its advanced modeling of scene geometry. Although our approach performs better on most datasets and metrics, significant domain differences among datasets limit information capture in ambiguous scenarios, leading to slightly lower accuracy on certain metrics.

3.3. Ablation Study

To evaluate the combined efficiency of the dynamic channel gating (DCG) mechanism and geometric-aware learning (GAL), we design an ablation experiment for the module. The experiment evaluates the following three configurations: (1) using only DCG, (2) using only GAL, (3) combining both DCG and GAL. The results are shown in Table 3, The DCG mechanism enhances feature discriminability through channel-wise feature calibration, while the GAL module strengthens geometric feature representation via explicit spatial relationship modeling. The two components exhibit significant complementary advantages and jointly optimize the representation learning of sparse 3D data.

Best results	S3DIS		MultiScan	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
Only GAL	77.8	67.3	67.3	55.3
Only DCG	78.1	67.2	68.0	54.1
GAL+DCG (Ours)	**80.5**	**71.8**	**69.6**	**56.3**

Table 3. Ablation study of dynamic channel gating mechanism and geometry-aware learning.

Best results	S3DIS		MultiScan	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
Manhattan	76.8	67.0	69.1	56.1
Mahalanobis	76.3	65.6	68.5	56.3
Euclidean(Ours)	**80.5**	**71.8**	**69.6**	**56.3**

Table 4. Ablation study of different distance algorithms.

We conduct an ablation study on the feature centroid distance measurement method in the GAL module to evaluate the effectiveness of different distance algorithms. The results are shown in Table 4, we systematically compare three typical metrics: Euclidean distance, Mahalanobis distance, and Manhattan distance. Extensive experimental results show that Euclidean distance consistently outperforms both Mahalanobis and Manhattan distances, which can be attributed to its superior adaptability to spatial relationships. Although Mahalanobis distance can capture correlations within feature distributions, its performance is limited by the stability of covariance matrix estimation. Meanwhile, Manhattan distance suffers from excessive sparsity, which inevitably leads to incomplete preservation of geometric features.

Best results	S3DIS		MultiScan	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
$\alpha = 1.0$	80.4	70.8	68.0	55.8
$\alpha = 1.5$	80.0	70.4	68.4	56.5
$\alpha = 2.5$	79.1	69.1	66.0	54.6
$\alpha = 3.0$	76.0	65.8	69.4	**57.2**
$\alpha = 2.0$ (Ours)	**80.5**	**71.8**	**69.6**	56.3

Table 5. Ablation study of the hyperparameter α.

To identify the optimal value of the decay hyperparameter α in the GAL module, we evaluate five different configurations: $\alpha \in [1.0, 1.5, 2.0, 2.5, 3.0]$. As shown in Table 5, the model obtains the optimum performance when $\alpha = 2.0$, demonstrating that this value strikes an optimal balance between keeping local feature sensitivity and decreasing long-range noise interference. A smaller α leads to insufficient decay, hence causing interference from distant places, while a larger α results in over-decay and loss of essential spatial contextual information.

We conduct an ablation study on the classification loss weight hyperparameter β, evaluating five configurations with β values ranging from 0.3 to 0.7. As shown in Table 6, the model achieves optimal performance when $\beta = 0.5$. Both higher and lower β values degrade detection accuracy. Specifically, an excessively large β causes the model to over emphasize the classification task, resulting in insufficient optimization of localization tasks such as bounding box regression. Conversely, an excessively small β weakens the model's ability to distinguish critical semantic features, leading to increased incorrect Predictions. Both scenarios ultimately impair detection performance.

Best results	S3DIS		MultiScan	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
$\beta = 0.3$	**80.6**	70.6	67.5	56.1
$\beta = 0.4$	80.1	68.4	68.1	56.2
$\beta = 0.6$	75.1	66.3	68.1	55.6
$\beta = 0.7$	78.1	66.9	65.3	52.9
$\beta = 0.5$ (Ours)	80.5	**71.8**	**69.6**	**56.3**

Table 6. Ablation study of the hyperparameter β.

4. CONCLUSION

In this paper, we propose UniGeo, a unified 3D indoor detection framework that integrates geometry-aware learning with a dynamic channel gating mechanism. To address the limitations of sparse 3D U-Net networks, we employ a multi-scale feature extraction strategy to capture both local details and global structural information. Moreover, we model geometric relationships through a geometry-aware learning module and enhance local feature representation via dynamic channel gating, effectively integrating global geometric information with discriminative local features. Extensive experiments on multiple indoor scene datasets demonstrate that our approach achieves state-of-the-art performance and outperforms existing mainstream methods. Ablation studies further validate the effectiveness of UniGeo.

5. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (Grant No. 62302145, 62272144), Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGTB0225, JZ2024HGTG0309, JZ2024AHST0337), Major Scientific and Technological Project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012, 202423k09020001), National Key R&D Program of China (NO.2024YFB3311602), and the Anhui Provincial Natural Science Foundation (2408085J040).

6. REFERENCES

[1] Songming Jia, Yan Lu, Bin Liu, Xiang Zhang, Peng Zhao, Xinmeng Tang, Yelin Wei, Jinyang Huang, Huan Yan, and Zhi Liu, "Breaking coordinate overfitting: Geometry-aware wifi sensing for cross-domain 3d pose estimation," in *Proceedings of the 32nd Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2026)*, 2026.

[2] Haiyang Wang, Lihe Ding, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, and Liwei Wang, "Ca-group3d: Class-aware grouping for 3d object detection on point clouds," *Advances in neural information processing systems*, vol. 35, pp. 29975–29988, 2022.

[3] 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*, 2019, pp. 9277–9286.

[4] Kun Dai, Zhiqiang Jiang, Tao Xie, Ke Wang, Dedong Liu, Zhendong Fan, Ruifeng Li, Lijun Zhao, and Mohamed Omar, "Sofw: A synergistic optimization framework for indoor 3d object detection," *IEEE Transactions on Multimedia*, 2025.

[5] Yun Zhu, Le Hui, Hang Yang, Jianjun Qian, Jin Xie, and Jian Yang, "Learning class prototypes for unified sparse-supervised 3d object detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9911–9920.

[6] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang, "Mlcvnet: Multi-level context votenet for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10447–10456.

[7] 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*. PMLR, 2022, pp. 180–191.

[8] Zhenyu Wang, Ya-Li Li, Xi Chen, Hengshuang Zhao, and Shengjin Wang, "Uni3detr: Unified 3d detection transformer," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39876–39896, 2023.

[9] Yun Zhu, Le Hui, Yaqi Shen, and Jin Xie, "Spgroup3d: Superpoint grouping network for indoor 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 7811–7819.

[10] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong, "Group-free 3d object detection via transformers," *arXiv preprint arXiv:2104.00678*, 2021.

[11] Wencheng Han, Junbo Yin, Xiaogang Jin, Xiangdong Dai, and Jianbing Shen, "Brnet: Exploring comprehensive features for monocular depth estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 586–602.

[12] Yichao Shen, Zigang Geng, Yuhui Yuan, Yutong Lin, Ze Liu, Chunyu Wang, Han Hu, Nanning Zheng, and Baining Guo, "V-detr: Detr with vertex relative position encoding for 3d object detection," 2023.

[13] Danila Rukhovich, Anna Vorontsova, and Anton Konushin, "Tr3d: Towards real-time indoor 3d object detection," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 281–285.

[14] Danila Rukhovich, Anna Vorontsova, and Anton Konushin, "Fcaf3d: Fully convolutional anchor-free 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 477–493.

[15] Maksim Kolodiazhnyi, Anna Vorontsova, Matvey Skripkin, Danila Rukhovich, and Anton Konushin, "Unidet3d: Multi-dataset indoor 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4365–4373.

[16] 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*, 2017, pp. 5828–5839.

[17] 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*, 2016, pp. 1534–1543.

[18] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva, "Multiscan: Scalable rgbd scanning for 3d environments with articulated objects," *Advances in neural information processing systems*, vol. 35, pp. 9058–9071, 2022.

[19] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner, "Rio: 3d object instance re-localization in changing indoor environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7658–7667.

[20] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.

[21] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al., "Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data," *arXiv preprint arXiv:2111.08897*, 2021.

[22] Loic Landrieu and Martin Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4558–4567.

[23] MMDetection3D Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," `https://github.com/open-mmlab/mmdetection3d`, 2020.