

Learning Robust 3D Representations via Multi-Granularity Gaussian Mixture Guided Cross-Modal Fusion

Li Han

lhan98@lnnu.edu.cn

Liaoning Normal University

Yuping Zhang

Liaoning Normal University

Chenyang Fu

Liaoning Normal University

Yongxin Song

Liaoning Normal University

Xiaoran Qi

Liaoning Normal University

Research Article

Keywords: Cross-modal learning, Point cloud, Gaussian Mixtures, Unsupervised learning, Semantic guidance

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9962358/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Learning Robust 3D Representations via Multi-Granularity Gaussian Mixture Guided Cross-Modal Fusion

Li Han^{1*} • Yuping Zhang • Chenyang Fu¹ • Yongxin Song¹ • Xiaoran Qi¹

¹Liaoning Normal University, School of Computer Science & Artificial Intelligence, Dalian, 116021, China

Abstract— This paper presents a novel cross-modal framework for 3D representation learning, which is driven by a multi-granularity Gaussian mixture model (MGM) to achieve geometrically interpretable and robust feature extraction. Our key innovation reformulates representation learning as a process of hierarchical probabilistic modeling, decoding latent features into Gaussian distributions through a tree-structured recursive mechanism with dynamic covariance adaptation. This MGM module explicitly captures the intrinsic geometric and statistical properties of 3D shapes. To enhance fine-grained cross-modal alignment, we introduce a Multi-scale Visual Enhancement (MVE) module that provides rich semantic guidance by distilling multi-scale visual features. Beyond simple feature fusion, our framework pioneers a new learning paradigm that establishes fine-grained semantic associations across modalities under explicit geometric constraints. Extensive experiments validate the framework’s superiority and robustness. It achieves state-of-the-art performance, with average gains of 0.2%-0.5% in classification, and 0.5%-3.0% in part segmentation and few-shot learning, demonstrating strong generalization capability against noise and data variations.

Keywords—Cross-modal learning • Point cloud • Gaussian Mixtures • Unsupervised learning • Semantic guidance

I. INTRODUCTION

Point cloud understanding is a cornerstone of 3D computer vision and underpins critical applications ranging from autonomous driving and robotics to VR/AR. Pioneering works such as PointNet and PointNet++ [1, 2] first demonstrated the feasibility of directly processing unordered point sets using shared multiplayer perceptrons (MLPs), inspiring a rich line of subsequent architectures including KPConv, PointConv, and PASNL [3-5]. The field has since been further advanced by Transformer-based models (e.g., Point Transformer [6], PCT [7], and the PTV series [8, 9]), which leverage self-attention to capture long-range dependencies and have delivered substantial performance gains. Despite these advances, existing approaches share a common reliance on dense annotations, fixed sampling strategies, and considerable computational resources, inherently hindering their robustness and adaptability to diverse real-world scenarios.

To alleviate the scarcity of large-scale 3D annotations, cross-modal learning has emerged as a promising direction. By aligning features across modalities, these methods exploit abundant 2D data and semantic labels to enrich 3D representations. Current methods generally follow two technical pathways. The first converts 3D point clouds into image-based representations to reduce the modality gap, thereby enabling the use of pre-trained vision models such as CLIP for cross-modal feature enhancement, as exemplified by PointCLIP and CLIP2Point [10,11]. The second employs techniques including knowledge distillation, masked autoencoding, contrastive learning, and adversarial learning to bridge modalities, with representative works including CrossPoint [12], PointM2AE [13], PointContrast [14], and PointCMAE [15]. More recently, CrossOver [16] has achieved instance-segmentation-free alignment of point clouds, images, text, and CAD models in scenarios under data-missing scenarios. Despite this progress, a critical limitation persists: existing methods predominantly operate at the global or coarse-grained level, establishing only approximate correspondences across modalities. This lack of fine-grained, geometrically aware semantic guidance ultimately undermines both the discriminability and transferability of the learned 3D embeddings.

To fundamentally address this granularity gap, we introduce a multi-granularity, Gaussian-driven cross-modal learning paradigm. Departing from prevalent discriminative or hybrid frameworks, our core innovation is a reformulation of 3D representation learning as a structured, hierarchical generative process. This is realized through dual-branch variational autoencoder (VAE) architecture. Its encoder jointly captures structural and probabilistic properties of point clouds, while its decoder—a multi-granularity Gaussian mixture module—explicitly models the spatial distribution of points, endowing the latent representation with a hierarchical and geometrically grounded structure. This generative formulation provides a natural scaffold for establishing fine-grained cross-modal associations, shifting the paradigm from coarse alignment toward precise, geometry-informed semantic fusion.

Building upon this foundation, we propose a self-supervised alignment mechanism. Multi-scale image embeddings are aligned with hierarchical point features via a novel intra- and inter-modal consistency loss, while a negative

log-likelihood reconstruction loss ensures generative fidelity. Through joint optimization, the model converges to a unified, geometrically constrained embedding space where 2D and 3D semantics are precisely matched. Extensive experiments on shape classification, part segmentation, and few-shot learning demonstrate the superior discriminability, robustness, and transferability of our framework, which consistently outperforms state-of-the-art methods.

II. RELATED WORKS

As the pioneering deep-learning architecture for 3D data, PointNet [1] handles unordered point sets via permutation-invariant symmetric functions. PointNet++ [2] extends it with a hierarchical sampling-and-grouping strategy, yet its rigid neighborhood partitioning limits the modeling of fine-grained local structures. To address this, Wang et al. propose DGCNN [17], where dynamic EdgeConv layers iteratively recast local geometry as a k-NN graph and perform graph convolutions, allowing receptive fields to evolve adaptively with the features. Inspired by the success of attention and Transformer architectures in 2D vision, researchers have increasingly explored long-range dependencies in point clouds. Seminal works like Point Transformer [6] integrate learnable position encoding with self-attention to fuse multi-scale features. PCT [7] further incorporates an offset-attention module to enhance local geometric sensitivity. Concurrently, methods like S2ANet [18] and SparseVoxNet [19] demonstrate the efficiency of local spectral grouping and volumetric representation for 3D feature learning. Despite their high accuracy, these approaches are often computationally intensive, memory-heavy, and most critically dependent on large annotated datasets, which impedes their adaptability to new domains.

Self-supervised learning has emerged as a pivotal strategy to mitigate data scarcity in 3D vision. One significant research thread leverages probabilistic representations, particularly Gaussian Mixture Models (GMMs). Early work by Ben-Shabat et al. [20] encode voxelized point clouds into GMMs, a concept later extended by Eckart et al. [21] into a hierarchical GMM (hGMM) for multi-scale descriptors. State-of-the-art in this direction, PointGMM [22], couples hGMM with a variational autoencoder, using an attention-based encoder to predict parameters and achieving top-tier results in generation and registration. Beyond probabilistic modeling, other innovative paradigms have emerged, including the cross-reconstruction framework by Zhang [23] and Gaussian-splatting techniques such as GSFaceMorpher [24], underscoring the versatility of these representations for complex tasks.

In parallel, proxy-task-based self-supervision has demonstrated remarkable generalization capabilities. This category includes OcCo [25], which learns local geometric context through occlusion completion, as well as methods that predict geometric transformations to foster invariance [26]. The success of contrastive learning has also been adapted to 3D. DepthContrast [27] creates positive pairs from different viewpoints and incorporates geometric augmentations to boost discrimination, while ClusterNet [28] jointly optimizes contrastive and clustering objectives to improve feature separability. Nonetheless, the scale of publicly available 3D data is still orders of magnitude smaller than its 2D counterpart;

consequently, sophisticated point-based self-supervision often saturates prematurely and struggles to learn truly transferable representations.

Cross-modal learning transcends the representational limitations of single-sensor data by integrating complementary information from multiple sources. Pioneering work in this domain, such as CrossPoint [12] established a global alignment of image-to-point-cloud via contrastive learning. This was followed by methods like PointM2AE [13] and CSL [29], which pre-train a masked auto-encoder on large-scale, unlabeled image – point pairs to learn bidirectional latent mappings. Subsequent research has pursued finer-grained alignment through diverse strategies. PointContrast [14] enforces dense point-level correspondences. While VIP-GAN [30] introduces an adversarial learning framework to align visual and point features at a fine scale. STRL [31] further strengthens modality-invariant representations by exploiting temporal coherence in synchronized video–LiDAR sequences. Building on these foundations, recent works have tackled increasingly granular challenges, advancing cross-modal 3D retrieval [32] and image-to-point cloud registration [33]. The pursuit of fine-grained alignment is also advanced by techniques designed for non-rigid point clouds, such as anisotropic hybrid field harmonization [34]. Despite this significant progress, two principal limitations persist: coarse global feature matching and the absence of a structured mechanism for cross-modal information exchange. To overcome these challenges, we propose a cross-modal framework built upon Multi-granularity Gaussian Mixtures (MGM), which unifies hierarchical probabilistic 3D modeling with fine-grained cross-modal fusion.

III. METHODOLOGY

Our framework jointly models local geometry and global spatial context for enhanced 3D understanding. At its core lies an MGM decoder that transforms graph-based features into a hierarchical set of Gaussian components. By explicitly parameterizing probability density, it produces structure-augmented representations that seamlessly integrate fine-grained geometric details with robust statistical information, significantly boosting resilience against noise, point sparsity, and shape deformations. Complementing this, a Multi-scale Visual Enhancement (MVE) module refines image features to deliver rich visual-semantic guidance, further strengthening cross-modal alignment for fine-grained scene interpretation. As illustrated in Fig. 1, the entire pipeline is trained end-to-end via a self-supervised cross-modal pre-training strategy. The MVE module, point encoder (PE), and MGM decoder are jointly optimized under a cross-modal semantic consistency objective, driving the framework toward a unified and transferable multi-modal embedding space.

3.1 Multi-Granularity Gaussian Mixture Learning

We formulate our multi-granularity Gaussian mixture learning module within a VAE framework that maps latent-space features to multi-granularity GMM parameters. This formulation jointly enforces both distribution consistency and geometric reconstruction consistency, substantially enhancing the geometric representation capacity of point cloud features.

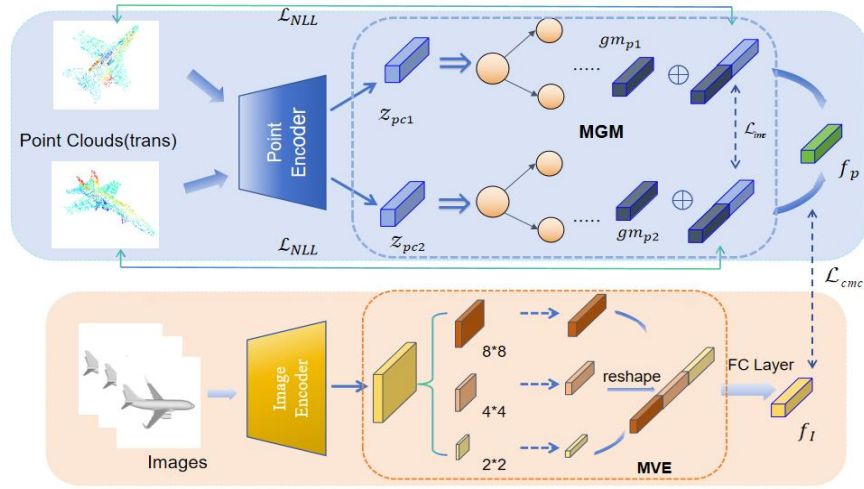


Fig.1. Overview of the proposed cross-modal 3D representation learning framework. The framework comprises two synergistic components: a Multi-granularity Gaussian Modeling (MGM) module for robust point cloud encoding, and a Multi-scale Visual Enhancement (MVE) module for fine-grained visual guidance. Both are jointly optimized under a unified cross-modal objective.

Consider a point-cloud X modeled as a mixture of K weighted, overlapping Gaussian distributions. The probability density for a point $x_i \in X$ generated by the GMM is given by:

$$P(X | \Theta^{l=1}) = \prod_{i=1}^m P(x_i | \Theta^{l=1}) = \prod_{i=1}^m \sum_{k=1}^K \pi_k P(x_i | \Theta_k^{l=1}) \quad (1)$$

where m denotes the number of points, $l=1$ indicates the root of the multi-granularity Gaussian hierarchy, and $\Theta_k = \{\pi_k, \mu_k, \Sigma_k\}$ represents the parameters of the k -th Gaussian component. Here, π_k is the mixture weight, while μ_k and Σ_k denote the mean vector and covariance matrix, respectively.

To capture fine-grained structures, we recursively refine coarse-grained Gaussian coefficients. Specifically, given the posterior probability $E^{l=d-1}$ and coarse-grained parameters $\Theta^{l=d-1}$ from a parent layer, the fine-grained Gaussian mixture at the child layer $l=d$ is expressed as:

$$P(X | E^{l=d-1}, \Theta^{l=d}) = \prod_{i=1}^m \sum_{k=1}^{K^*} \pi_k^{l=d-1} P(x_i | \Theta_k^{l=d-1}) \quad (2)$$

where K^* denotes the number of Gaussian components in the child layer. Following a hard assignment strategy, each point is assigned to the Gaussian component whose parent node yields the highest posterior probability. This process iteratively constructs a tree-structured, coarse-to-fine, multi-granularity Gaussian mixture model, wherein each child node represents a refined, finer-grained counterpart of its parent. In our implementation, a weight-shared MLP recursively splits parent nodes from top to bottom to generate these child nodes, producing multi-granularity distributional features that are subsequently decoded into the final GMM parameters.

To facilitate effective information flow across the hierarchy, we introduce an attention mechanism that dynamically modulates the anisotropy of each Gaussian component. This mechanism enlarges covariance in smooth regions, thereby reducing the number of required components, and shrinks it in high-curvature regions to increase component

density. As a result, the model adaptively captures complex geometries while maintaining computational efficiency. The attention modulation is formalized as:

$$h_{l,j} = \text{Atten}(W_p, W_e) \cdot h_{l-1,j} \quad (3)$$

where W_p, W_e are the learnable weights associated with the parent distribution $\Theta_k^{l=d-1}$ and the current distribution $\Theta_k^{l=d}$, respectively.

To improve robustness against noise and sparsity, we employ a data augmentation strategy that applies random transformations to the input point cloud X . The augmented data $X_{p1}, X_{p2} \in \mathbb{R}^{m \times 3}$ are encoded into feature vectors $z_{p1}, z_{p2} \in \mathbb{R}^D$ by the point encoder, which are then processed by the MGM decoder to explicitly model the underlying spatial distribution.

The parameter-generation function $h_{l,j}$ in Eq.(3) is implemented as a three-layer MLP:

$$\begin{pmatrix} \mu_{(l,j)} \\ \Sigma_{(l,j)} \\ \pi_{(l,j)} \end{pmatrix}^T = h_{l,j}(W_p W_e \{h_{l-1,j}\}_{j=1}^K) \quad (4)$$

Each hidden layer is followed by Batch Normalization and a ReLU activation. The output layer applies three separate linear projections to produce the Gaussian parameters with appropriate constraints: a softmax activation over the K components for the mixture weights $\pi_{(l,k)}$ (ensuring they sum to unity); a tanh activation for the mean vectors $\mu_{(l,k)}$; and a softplus activation for the diagonal covariance elements $\Sigma_{(l,k)}$ to guarantee positive-definiteness.

Through iterative feature propagation and node splitting, an optimized MGM-tree representation is constructed. Each node is parameterized by a Gaussian mixture model $\Theta^{(l)} = \{\pi_k^{(l)}, \mu_k^{(l)}, \Sigma_k^{(l)}\}$. The Gaussian parameters from the top

layer $l=L$ constitute the multi-granularity distributional feature representation:

$$G = \mu^{(l)} \oplus \Sigma^{(l)} \oplus \pi^{(l)} \quad (5)$$

where $\oplus$ denotes feature concatenation. The final synthesized representation is obtained by concatenating this multi-granularity feature with geometric features z :

$$f_p = z \oplus G \quad (6)$$

Figure 2 presents a comparison of shape reconstruction between the proposed multi-granularity Gaussian mixtures and a traditional GMM. Both models represent the input point cloud using 64 Gaussian distributions. The traditional GMM generates all mixtures in a single, monolithic step at the final layer. In contrast, our MGM employs a four-level tree structure, progressively refining the representation with 8, 16, 32, and finally 128 Gaussians. The visual results demonstrate that MGM captures finer geometric details and more accurately reconstructs the underlying semantic structures.

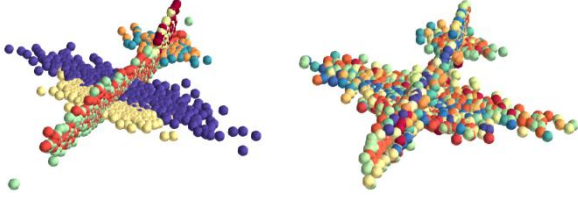


Fig.2. Comparison of shape reconstruction between traditional GMM (left) and our MGM (right)

In our experimental analysis, we further evaluate the impact of two key hyperparameters -tree depth L and number of Gaussian components K on model performance through a comprehensive ablation study.

3.2 Multi-Scale Visual Enhancement

To overcome the coarse-grained feature correspondence that limits existing cross-modal approaches, we propose a multi-scale visual-semantic enhancement scheme. This module employs a spatial pyramid strategy to extract fine-grained image semantics and strengthens their cross-modal guidance through local semantic association.

Given a point cloud dataset $D = \{(X_i, I_i)\}_{i=1}^N$, where $I_i \in \mathbb{R}^{H \times W}$ is the RGB image rendered from point cloud X_i via a randomly positioned camera. Each image is processed by a ResNet-based visual encoder, which maps it to a high-dimensional feature space $f_i \in \mathbb{R}^{H \times W \times C}$, with C denoting the feature dimension and $H \times W$ the spatial resolution.

Leveraging a spatial pyramid strategy, we construct a multi-scale feature-pooling block. At each spatial scale $n \times n$ (with $n \in \{2, 4, 8\}$), the module concurrently applies max-pooling and average-pooling operations to capture discriminative fine-grained semantics:

$$\hat{f}_i^n = \text{AvgPool}(f_i^n) + \text{MaxPool}(f_i^n) \quad (7)$$

where AvgPool, MaxPool represent average pooling and max pooling, respectively. Features pooled across all scales are subsequently integrated via concatenation:

$$f_i' = \hat{f}_i^2 \oplus \hat{f}_i^4 \oplus \hat{f}_i^8 \quad (8)$$

The combined multi-scale feature is then transformed by a fully-connected layer to produce the locally-enhanced semantic visual feature f_l .

3.3 Joint Loss Function

We propose a dual-branch collaborative learning framework comprising the visual-semantic enhancement module and the multi-granularity Gaussian mixture learning module. The framework is trained in a self-supervised manner, achieving effective cross-modal fusion through a joint optimization objective.

A ResNet [35] backbone serves as the image encoder; after processing by the MVE module, it yields multi-scale visual semantic features. Concurrently, a DGCNN-based point encoder is employed alongside the MGM module to form an encoder-decoder architecture, generating a multi-granularity point feature representation that integrates both geometric and spatial distributional information.

To enable fine-grained cross-modal alignment, we introduce a joint optimization strategy incorporating three complementary objectives: intra-modal structural consistency L_{imc} , cross-modal semantic consistency L_{cmc} , and probabilistic distribution consistency L_{NLL} .

$$\mathcal{L} \equiv \lambda_1 L_{imc} + \lambda_2 L_{cmc} + \lambda_3 L_{NLL} \quad (9)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contribution of each term.

Intra-modal Semantic Structure Consistency

We adopt a contrastive learning paradigm to improve the robustness of point cloud feature extraction.

For a given point cloud sample X_i and its randomly augmented counterparts X_{i1}, X_{i2} , the point encoder and MGM decoder jointly generate their respective fused multi-granularity features f_{i1}, f_{i2} .

The intra-modal structural consistency loss for a positive pair is defined as:

$$l(i, f_{i1}, f_{i2}) = -\log \frac{\exp(\cos(f_{i1}, f_{i2})/\tau)}{\sum_{j=1}^m \exp(\cos(f_{i1}, f_{j2})/\tau)} \quad (10)$$

where $\cos(\cdot)$ denotes the cosine similarity and τ is the temperature coefficient. This formulation strengthens the cohesion of intra-class samples while repelling negative pairs. The overall intra-modal consistency loss is computed symmetrically over the entire training set:

$$L_{imc} = \frac{1}{2m} \sum_{i=1}^m (l(i, f_{i1}, f_{i2}) + l(i, f_{i2}, f_{i1})) \quad (11)$$

where m is the number of samples.

Cross-Modal Fine-Grained Matching Consistency

Unlike prior works that rely on global alignment, our formulation operates at instance level, enabling precise correspondence:

$$L_{cmc} = \frac{1}{2m} \sum_{i=1}^m (l(i, f_I, f_p) + l(i, f_p, f_I)) \quad (12)$$

where f_I denotes the multi-scale enhanced visual feature from Eq.(8) and f_p represents the fine-grained point cloud features.

Probabilistic Distribution Consistency

Our method models local shape and spatial distribution of a point cloud via multi-granularity Gaussian mixtures. This approach improves the semantic discriminability of the learned representation and enhances robustness to geometric transformations and shape noise. To optimize the MGM decoder, we employ a negative log-likelihood loss.

For a given point cloud X , the maximum-likelihood loss at the l -th layer is defined as:

$$\ell_{NLL}(X | \Theta^{l=d}) = \begin{cases} \log P(X | \Theta^{l=1}) & d = 1 \\ \log P(X | E^{l=d-1}, \Theta^{l=d}) & d \neq 1 \end{cases} \quad (13)$$

The overall distribution consistency loss aggregates the losses across all layers:

$$L_{NLL}(X, \Theta) = -\frac{1}{|X|} \sum_{d=1}^D \ell_{NLL}(X | \Theta^{l=d}) \quad (14)$$

Where $|X|$ denotes the number of points. This hierarchical negative log-likelihood objective ensures that each layer of the Gaussian mixture tree faithfully reconstructs the underlying point distribution at its corresponding granularity.

IV. EXPERIMENTS AND ANALYSIS

4.1 Implement Details & Datasets

All experiments were conducted on a workstation equipped with an NVIDIA RTX A4000 GPU, using Python 3.10 and PyTorch 1.11.0. The proposed network was trained in a self-supervised manner on the large-scale ShapeNet-Core dataset and evaluated on the ModelNet40 and ScanObjectNN benchmarks. For each point cloud, a 2D image was rendered from an arbitrary viewpoint. To enhance model robustness and generalization, these rendered images underwent a series of augmentations, including random cropping, color jittering, and horizontal flipping. Each point cloud was uniformly sub-sampled to 1,024 points via farthest-point sampling (FPS). Finally, all images were resized to 224×224 pixels to conform to the input requirements of the visual encoder.

MVE Encoder. The image branch employs ResNet-50 architecture, initialized with ImageNet pre-trained weights to serve as a visual semantic extractor. The output features are processed by a multi-scale pooling operation that applies both average and max pooling operations with kernel sizes of 7×7, 5×5, and 3×3. The resulting multi-scale features are fused and

projected through a fully connected layer to produce a 1,024-dimensional visual feature vector f_I .

Point Cloud Encoder: The geometric branch utilizes a DGCNN as its backbone. The network comprises four EdgeConv layers, each of which dynamically constructs a local graph for every point using k -NN algorithm ($k=20$). The resulting features are then processed by a multi-layer perceptron (MLP) projection to produce a 512-dimensional global point cloud descriptor f_p .

MGM decoder: The 512-dimensional point cloud feature is fed into the MGM decoder to generate a hierarchical, distribution-aware representation. The decoder initializes a Gaussian mixture model with $K=8$ components. This representation is then refined over $L=3$ hierarchical levels, generating child nodes containing 32, 64, and 128 Gaussian distributions, respectively. The parameters of the final set of $K=128$ Gaussians are concatenated with the original DGCNN features to form a fused multi-granularity representation.

Cross-Modal alignment objective: To achieve cross-modal alignment, the visual and fused point cloud features are projected into a unified latent space using a parameter-shared two-layer MLP. The model is optimized through a cross-modal consistency contrastive learning objective, which leverages the joint geometry-distribution representation to enhance feature discriminability.

Training configuration: The network is pre-trained using the Adam optimizer with momentum 0.9 and weight decay 1e-4, and an initial learning rate of 0.001. We use a batch size of 16 for pre-training and 32 for validation on the ModelNet40 dataset. The model is trained for 250 epochs. The temperature coefficient in the contrastive loss function is set to 0.1, and the coefficients for the joint regularization terms are tuned to stabilize training and enhance performance; a detailed ablation study of these hyperparameters is presented in Section 4.5.

4.2 Point Cloud Classification

We first evaluate our model's feature learning capability on point cloud classification. A linear SVM classifier is trained on features extracted from the training splits of ModelNet40 and ScanObjectNN. The resulting classification accuracy is compared against representative self-supervised deep learning methods.

As summarized in Table 1, our method consistently outperforms existing point-based approaches on ModelNet40. Specifically, 3D-GAN [36] exploits adversarial training to enrich 3D data, yet its limited learning efficiency leads to a modest accuracy of 65.3%. Latent-GAN [37] improves upon this by applying an adversarial loss in the latent space, albeit with significant computational overhead. Other methods, including SO-Net [38] with its self-organizing map for hierarchical structure learning, FoldingNet [39] with its surface-reconstruction folding decoder, and 3D-PointCapsNet [40] with its part-level capsule networks, have progressively advanced performance, with the latter reaching 88.9% accuracy.

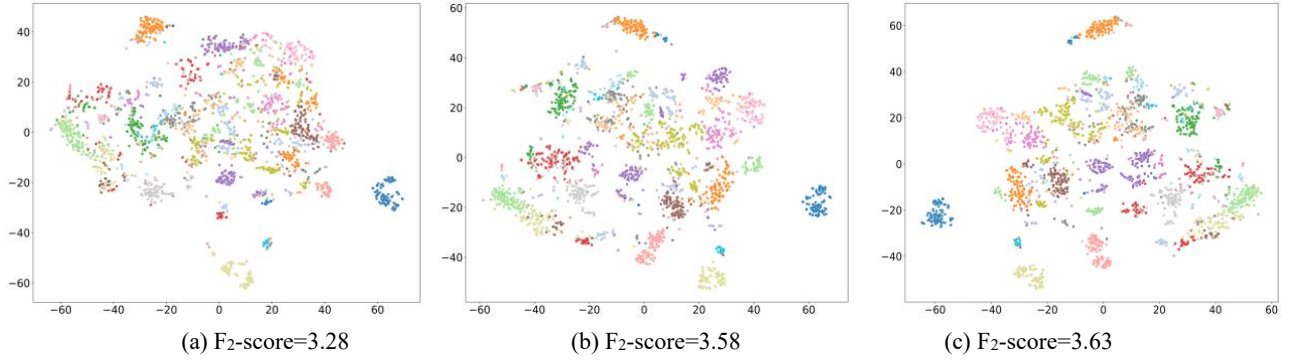


Fig.3. t-SNE visualization of feature embeddings. (a) DGCNN alone, (b) CrossPoint [12], (c) Ours.

Our method surpasses these approaches, achieving 91.4%, demonstrating the effectiveness of the proposed multi-granularity Gaussian-driven cross-modal learning paradigm.

Table 1. Linear Classification Results on ModelNet40

Methods	Backbone	Params (M)	Linear Acc (%)
Generative / GAN-based Methods			
3D-GAN [36]	3D-GAN	-	88.3
Latent-GAN [37]	Latent-GAN	-	85.7
Unsupervised Methods			
SO-Net [38]	SO-Net	~17.5	87.3
FoldingNet [39]	FoldingNet	~2.4	88.4
MRTNet [41]	MRTNet	-	86.4
3D-PointCapsNet [40]	CapsNet	-	88.9
Self-supervised Methods			
DepthContrast [27]	PointNet++	~1.5	85.4
ClusterNet [28]	DGCNN	~1.8	86.8
VIP-GAN [30]	DGCNN	~1.8	90.2
MAP-VAE [33]	PointNet	~3.5	88.4
OcCo [25]	DGCNN	~1.8	89.2
STRL [31]	DGCNN	~1.8	90.9
Rotation [26]	DGCNN	~1.8	90.8
Jigsaw [42]	DGCNN	~1.8	90.6
Cross-modal Methods			
CrossPoint [12]	DGCNN + ResNet	~1.8 + ~25	91.2
PointMAE [43]	Transformer	22.1	91.0
DHGCN [44]	DGCNN	~1.8	90.9
PartCon [45]	DGCNN	~1.8	90.9
Ours	DGCNN + ResNet-50	~1.8 + ~26	91.4

Further gains have been achieved through cross-modal learning. VIP-GAN [30] adopts adversarial alignment between visual and point cloud representations, reaching 90.2% accuracy. MAP-VAE [33] employs a variational framework for probabilistic cross-modal mapping, while STRL [31] exploits spatio-temporal consistency to learn modality-invariant features. Jigsaw [42] enhances structural reasoning via a block-shuffling strategy, improving accuracy to 90.6%. CrossPoint[12] utilizes cross-modal contrastive learning between images and point clouds, achieving a linear accuracy of 91.2%. More recently, Transformer-based approaches such as PointMAE [43] have incorporated masking strategies to strengthen representation learning, reporting 91.0% accuracy. DHGCN [44] introduces a dual-hierarchical graph architecture that jointly models local geometry and global semantics within the DGCNN backbone, achieving 90.9% accuracy. PartCon [45]

enhances cross-modal alignment through part-aware contrastive learning, where semantic parts are explicitly mined and aligned across modalities, also attaining 90.9% linear classification performance.

Our method achieves a state-of-the-art linear classification accuracy of 91.4%, outperforming the above point-based approaches by an average margin of 4% and surpassing cross-modal methods by approximately 1.3%. While the improvement over PointMAE is 0.4%, our DGCNN-based framework offers a superior trade-off between performance and computational efficiency (-1.8M vs. 22.1M parameters).

To evaluate the quality of the learned feature representations, we visualize the feature embeddings via t-SNE and quantitatively assess inter-class separability using the F_2 -score, where a higher value indicates more discriminative features. As shown in Fig. 3a, features extracted using DGCNN alone on the ModelNet40 validation set yield an F_2 -score of 3.28. The resulting clusters exhibit substantial overlap, indicating weak inter-class separation and low intra-class cohesion. In comparison, features learned by cross-modal contrastive learning (CrossPoint [12]), shown in Fig. 3b, improve the F_2 -score to 3.58 and show better separation, although several outliers and ambiguous regions persist. In contrast, our approach, which integrates multi-granularity geometric and distributional features with cross-modal semantic enhancement, produces well-separated clusters (Fig. 3c) and achieves the highest F_2 -score of 3.63. This demonstrates a marked improvement in discriminating visually similar categories, such as desks and tables.

The quality of the learned representations is further validated through shape reconstruction. As shown in Figure 4, qualitative comparisons reveal that our model generates shapes with more accurate semantic structures and finer geometric details compared to those from the CrossPoint baseline.

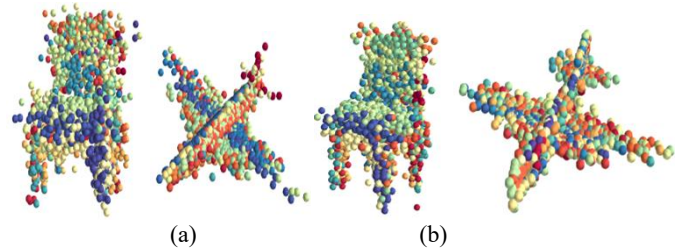


Fig.4. Reconstructed shapes using deep features from (a) CrossPoint [12] and (b) our MGM.

4.3 Point Cloud Segmentation

To verify the fine-grained representation capability of our model, we conduct part segmentation experiments on the ShapeNetPart dataset. The segmentation network is trained using cross-entropy loss, and evaluated via the mean Intersection-over-Union (mIoU) metric.

As reported in Table 2, our approach achieves a state-of-the-art mIoU of 86.2%, outperforming all existing approaches by an average margin of 0.9%. Specifically, it surpasses the strongest existing baselines—MaskPoint [47], PointCMAE [15] and Point-PQAE [23] (86.0-86.1%)—by 0.1-0.2 percentage points.

Quantitative per-category results are summarized in Table 3. Our model demonstrates consistent superiority, achieving the highest segmentation accuracy in 10 out of 16 fine-grained categories. This indicates a particular strength in handling complex geometries. Specifically, we attain 92.7% on the *chair* category (outperforming PointBERT [46] by 1.7% and DGCNN [17] by 2.1%), 85.4% on *pistol* (a 1.1% improvement over PointBERT), and 86.0% on *lamp* (exceeding the Transformer-based baseline [6] by 0.7%).

Table 2. Part Segmentation mIoU on ShapeNetPart

Part Segmentation	mIoU (%)
PointNet [1]	83.7
PointNet++ [2]	85.1
DGCNN [17]	85.1
Transformer [6]	85.1
PointBERT [46]	85.6
Jigsaw [42]	85.3
OcCo [25]	85.0
PointContrast [14]	85.1
MaskPoint [47]	86.0
CrossPoint [12]	85.5
PointCMAE [15]	86.0
Point-PQAE [23]	86.1
Ours	86.2

Table 3. Per-Category Segmentation Results (mIoU, %) on ShapeNetPart

	mIoU	airplane	bag	cap	car	chair	earphone	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate	table
PointNet [1]	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
PointNet++ [2]	85.1	82.4	79	87.7	77.3	90.8	71.8	91.0	85.9	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
DGCNN [17]	85.1	84	83.4	86.7	77.8	90.6	74.7	91.2	87.5	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
Transformer [6]	85.1	82.9	85.4	87.7	78.8	90.5	80.8	91.1	87.7	85.3	95.6	73.9	94.9	83.5	61.2	74.9	80.6
PointBERT [46]	85.6	84.3	84.8	88.0	79.8	91.0	81.7	91.6	87.9	85.2	95.6	75.6	94.7	84.3	63.4	76.3	81.5
OcCo [25]	85.0	83.3	85.2	88.3	79.9	90.7	74.1	91.9	87.6	84.7	95.4	75.5	94.4	84.1	63.1	75.7	80.8
MaskPoint [47]	86.0	84.2	85.6	88.1	80.3	91.2	79.5	91.9	87.8	86.2	95.3	76.9	95.0	85.3	64.4	76.9	81.5
PointCMAE [15]	86.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ours	86.2	84.9	85.7	88.4	79.9	92.7	77.0	91.7	88.0	86.0	95.8	77.0	94.1	85.4	64.6	76.8	82.6

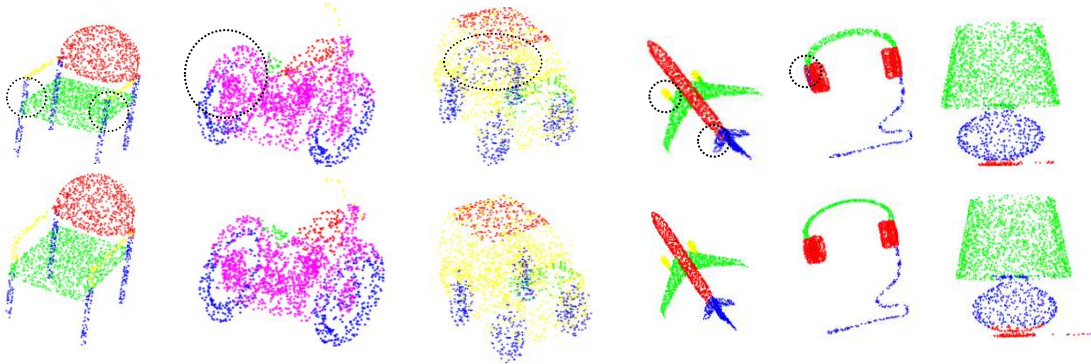


Fig. 5. Segmentation comparison: ground-truth labels (top) vs. our results (bottom).

Figure 5 presents a comparative analysis of segmentation results against ground truth annotations. Our method exhibits superior performance in segmenting fine-grained parts for complex structural models. In the *chair* category, for instance, the baseline method misclassifies the armrest as part of the leg component, whereas our model accurately distinguishes these semantically distinct parts. Similarly, for categories such as

motorbike, *cars*, and *aircraft*, our method achieves more precise localization of key components including wings, wheels, and rear sections, highlighting its capability to capture subtle geometric distinctions while maintaining semantic consistency.

4.4 Few-shot Classification

We further evaluate the few-shot classification performance of our model under the standard N -way K -shot protocol.

As reported in Table 4, our model achieves top-tier performance under both 5-way 10-shot ($93.7 \pm 3.4\%$) and 5-way 20-shot ($96.3 \pm 2.9\%$) configurations. Increasing the sample count from 10 to 20 per class yields a 2.6 percentage point improvement in accuracy. In the 5-way 20-shot setting, our approach outperforms PointCMAE [15] by 2.2%. Notably, while traditional methods such as PointNet and DGCNN exhibit significant performance degradation in the few-shot scenario, our cross-modal pre-training strategy effectively mitigates overfitting under limited supervision.

However, on the more challenging 10-way tasks, our result is slightly lower than that of PointCMAE, which benefits from masked token modeling and a Transformer architecture to attain stronger generalization across a large number of classes, albeit at the cost of increased computational overhead (22.1M vs. ~ 1.8 M parameters).

Table 4. Few-shot Classification Results (%) on ModelNet40

Methods	5-way 10-shot	5-way 20-shot	10-way 10-shot	10-way 20-shot
PointNet [1]	52.0 $\pm$ 3.8	57.8 $\pm$ 4.9	46.6 $\pm$ 4.3	35.2 $\pm$ 4.8
DGCNN [17]	31.6 $\pm$ 2.8	40.8 $\pm$ 4.6	19.9 $\pm$ 2.1	16.9 $\pm$ 1.5
3D-GAN [36]	55.8 $\pm$ 3.4	65.8 $\pm$ 3.1	40.3 $\pm$ 2.1	48.4 $\pm$ 1.8
Latent-GAN [37]	41.6 $\pm$ 5.3	46.2 $\pm$ 6.2	32.9 $\pm$ 2.9	25.5 $\pm$ 3.2
3D-PointCapsNet [40]	42.3 $\pm$ 5.5	53.0 $\pm$ 5.9	38.0 $\pm$ 4.5	27.2 $\pm$ 4.7
PointNet++ [2]	38.5 $\pm$ 4.4	42.4 $\pm$ 4.5	23.1 $\pm$ 2.2	18.8 $\pm$ 1.7
PointCNN [48]	65.4 $\pm$ 2.8	68.6 $\pm$ 2.2	46.6 $\pm$ 1.5	50.0 $\pm$ 2.3
RSCNN [49]	65.4 $\pm$ 8.9	68.6 $\pm$ 7.0	46.6 $\pm$ 4.8	50.0 $\pm$ 7.2
Jigsaw [42]	34.3 $\pm$ 1.3	42.2 $\pm$ 3.5	26.0 $\pm$ 2.4	29.9 $\pm$ 2.6
cTree [51]	60.0 $\pm$ 2.8	65.7 $\pm$ 2.6	48.5 $\pm$ 1.8	53.0 $\pm$ 1.3
CLIP2Point [11]	86.4 $\pm$ 1.6	90.7 $\pm$ 2.0	80.7 $\pm$ 2.1	—
OcCo [25]	90.6 $\pm$ 2.8	92.5 $\pm$ 1.9	82.9 $\pm$ 1.3	86.5 $\pm$ 2.2
CrossPoint [12]	92.5 $\pm$ 3.0	94.9 $\pm$ 2.1	83.6 $\pm$ 5.3	87.9 $\pm$ 4.2
PointCMAE [15]	90.4 $\pm$ 4.2	94.1 $\pm$ 3.9	89.2$\pm$5.5	92.3$\pm$4.5
Ours	93.7$\pm$3.4	96.3$\pm$2.9	86.0 $\pm$ 4.3	88.5 $\pm$ 3.4

To further verify the generalization capability of our method, we conduct experiments on the ScanObjectNN dataset (Table 5). Under the 5-way 20-shot configuration, our method achieves $82.7 \pm 4.9\%$, exceeding CrossPoint [12] by 3.7 percentage points. As the number of shots is increased from 10 to 20, our method exhibits a performance gain of 6.1%—substantially higher than competing approaches—highlighting its superior ability for learning discriminative features from limited data. In 10-way 20-shot scenario, our model outperforms OcCo [25] by a substantial margin of 13.6%, owing to the use of multi-granularity Gaussian features that effectively mitigate the impact of occlusions. Despite the increased standard deviations across all methods due to dataset noise, our method maintains consistently high accuracy, underscoring its robustness under real-world conditions.

Furthermore, we compare our approach with CLIP-based methods [10, 11, 50]. These methods transform point cloud data into 2D images that CLIP can process, thereby "borrowing" the generic visual-semantic knowledge learned

from massive image-text pairs to address data-scarce scenarios such as zero-shot or few-shot learning. However, due to the inherent discrepancies in modality conversion and inevitable structural information loss, their accuracy remains nearly 17.6 percentage points lower than that of our proposed method.

Table 5. Few-shot Classification Results (%) on ScanObjectNN

Method	5-way 10-shot	5-way 20-shot	10-way 10-shot	10-way 20-shot
Jigsaw [42]	62.0 $\pm$ 5.6	67.8 $\pm$ 5.1	37.8 $\pm$ 4.3	41.8 $\pm$ 2.4
cTree [51]	65.2 $\pm$ 3.8	72.2 $\pm$ 2.7	45.6 $\pm$ 3.1	48.2 $\pm$ 2.8
OcCo [25]	68.4 $\pm$ 3.4	71.6 $\pm$ 2.9	42.4 $\pm$ 2.7	43.0 $\pm$ 3.0
CLIP2Point [11]	59.2 $\pm$ 1.1	67.8 $\pm$ 0.9	47.5 $\pm$ 0.8	56.3 $\pm$ 0.7
PointCLIP [10]	63.5 $\pm$ 1.3	68.4 $\pm$ 1.1	—	—
PPCITNet [50]	62.5 $\pm$ 1.3	70.1 $\pm$ 1.0	51.8 $\pm$ 0.9	59.4 $\pm$ 0.8
CrossPoint [12]	72.4 $\pm$ 1.4	77.2 $\pm$ 1.4	57.0 $\pm$ 1.3	61.6 $\pm$ 1.2
PointCMAE [15]	74.8 $\pm$ 1.5	79.0 $\pm$ 1.2	62.9 $\pm$ 1.7	73.9 $\pm$ 2.2
Ours	76.6$\pm$2.6	82.7$\pm$4.9	64.5$\pm$2.2	75.2$\pm$2.8

4.5 Ablation Study

We conduct a series of ablation studies to systematically evaluate the effectiveness and robustness of the proposed framework under varying conditions, including different loss functions, the number of Gaussian mixtures, and noise levels.

4.5.1 Collaborative Optimization and Joint Loss Functions

Pre-training was performed for 50 epochs on ShapeNet, with ModelNet used for validation, to assess the contributions of individual components. Four configurations were designed:

Baseline: Single-branch architecture optimized solely with intra-modal contrastive loss (L_{imc}).

Cross-modal Extension: Dual-branch architecture incorporating cross-modal semantic consistency loss (L_{cmc}).

Cross-modal Enhancement: Addition of the MVE module for intra-modal feature augmentation, optimized with joint loss ($L_{imc} + L_{cmc}$).

Full Framework: Integration of the MGM module with all three constraints ($L_{imc} + L_{cmc} + L_{NULL}$).

All experiments used identical hyperparameters. As shown in Figure 6, the full framework with the joint loss achieves a linear classification accuracy of 89.5% within only 25 epochs, demonstrating smoother convergence compared to the significant oscillations observed with individual losses. Crucially, the incorporation of the negative log-likelihood loss for probability density calibration considerably improves model robustness, promotes effective feature disentanglement, and enhances generalization.

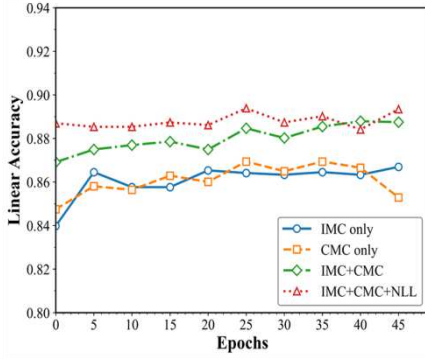
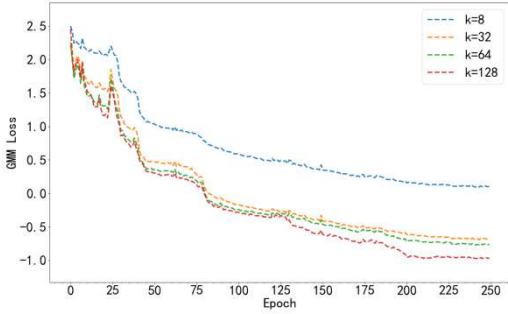


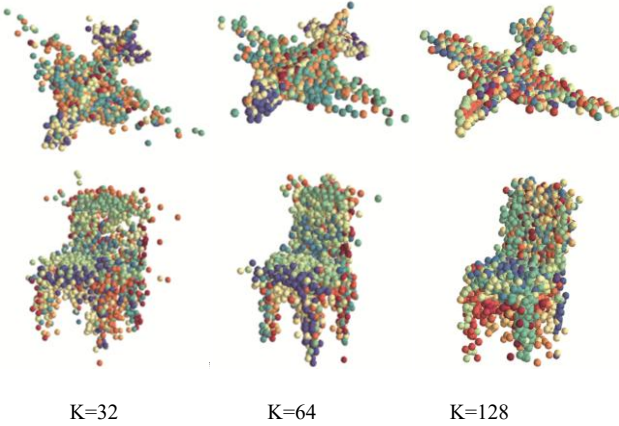
Fig. 6. Training accuracy curves under different loss configuration.

4.5.2 Number of Gaussian Mixtures and Noise Robustness

We investigate the impact of the number of Gaussian mixture components $K \in \{8, 32, 64, 128\}$ on learning performance. Results on ShapeNet (Fig.7a) demonstrate that as the multi-granularity Gaussian-mixture tree is refined, the loss curve converges to a more stable model after only 200 iterations when $K=128$.



(a)



(b)

Fig. 7. Influence of the number of Gaussian mixture components. (a) Loss curves, (b) Generated shapes.

We also employ these components for shape generation (Fig.7b). With increasing K , the enhanced fine-grained feature representation preserves more accurate semantic structures and geometric details in the generated shapes.

To evaluate noise robustness against input perturbations, we introduce Gaussian noise with standard deviations $\sigma = 0.03$

and 0.05 to the test data. As shown in Fig.8a, the linear accuracy of the CrossPoint baseline decreases by 5.1% and 7.3%, respectively, indicating high sensitivity to noise. In contrast, our method, which explicitly models point distributions via the MGM module, exhibits significant smaller decreases of only 1.8% and 4.1% under the same conditions.

We further evaluate robustness with respect to input point quantity (Fig. 8b). Performance improves markedly from 128 to 512 and continues to enhance before stabilizing between 1024 and 2048 points, demonstrating that our method benefits from denser point clouds while remaining effective even with sparse inputs.

4.5.3 Number of Rendered Views

We analyze the influence of the number of rendered views on cross-modal learning. As shown in Table 6, employing a single view yields the highest linear classification accuracy, while the incorporation of additional views leads to slight performance degradation. This suggests that the multi-scale spatial pooling within the MVE module effectively extracts comprehensive semantic information from individual views. In contrast, images from divergent viewpoints introduce heterogeneous feature representations that complicate cross-modal alignment. Consequently, pre-training the model with a single view achieves optimal performance while significantly reducing computational cost.

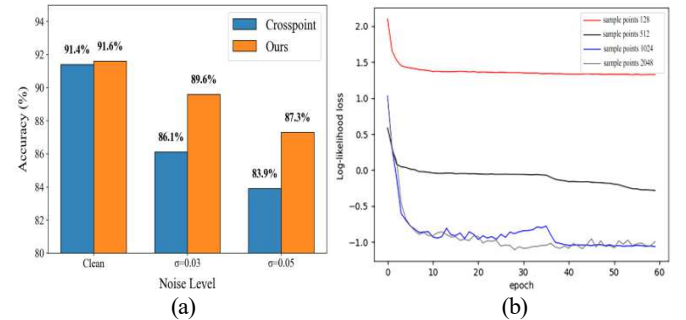


Fig. 8. Ablation studies on (a) Gaussian noise robustness, (b) input point sparsity.

Table 6. Impact of the Number of Rendered Images on Classification Accuracy

Images(n)	1	2	3	4
Linear Acc (%)	91.43	91.16	91.27	90.06

V. CONCLUSION

In this paper, we have presented a novel cross-modal learning framework for point cloud understanding, built upon multi-granularity Gaussian mixtures. The core of our approach is a recursive MGM decoder that explicitly models fine-grained probability distributions in point clouds, thereby establishing geometry-distribution collaborative feature-fusion mechanism.

To complement this geometric modeling, we introduce a spatial pyramid pooling module that provides fine-grained visual-semantic guidance, along with a triple joint optimization strategy that enhances cross-modal semantic alignment. Together, these contributions substantially advance self-supervised point-cloud representation learning. Extensive experiments demonstrate that the proposed method achieves high linear classification accuracy on shape recognition tasks and delivers notable gains in few-shot and part segmentation across multiple datasets. These results validate the effectiveness of integrating multi-granularity distribution modeling with cross-modal semantic guidance, offering a promising direction for 3D shape understanding under low-annotation regimes. For future work, we plan to extend Gaussian mixture modeling to dynamic point cloud sequences and to develop lightweight decoders suitable for large-scale semantic segmentation and real-time applications.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This research was supported by grants from *** and ***.

STATEMENTS AND DECLARATIONS

Competing Interests: All authors declare that they have no conflict of interests.

Data Availability and Access: The datasets generated and analyzed in this study, along with the code for replication, are available in the following GitHub repository: <https://github.com/Zxy0910/mgm.git>

REFERENCES

- [1] Qi C R, Su H, Mo K, et al. "PointNet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017: 652-660.
- [2] Qi C R, Yi L, Su H, et al. "PointNet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, 30.
- [3] Thomas H, Qi C R, Deschaud J E, et al. "KPConv: Flexible and deformable convolution for point clouds." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019: 6411-6420.
- [4] Wu W, Qi Z, Fuxin L. "PointConv: Deep convolutional networks on 3d point clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019: 9621-9630.
- [5] Yan X, et al. "PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
- [6] Zhao H, et al. "Point Transformer." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.
- [7] Guo M, et al. "PCT: Point cloud transformer." *International Journal of Computer Vision (IJCV)*. 2021: 187-199.
- [8] Wu X, Tian Z, Wen X, et al. "Point Transformer V2: Grouped vector attention and partition-based pooling." *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [9] Wu X, Jiang L, Wang P, et al. "Point Transformer V3: Simpler, faster, stronger." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, 2024:4840-4851.
- [10] Zhang R, Guo Z, Zhang W, et al. "PointCLIP: Point cloud understanding by CLIP." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022: 8552-8562.
- [11] Huang T, Dong B, Yang Y, et al. "Clip2Point: Transfer clip to point cloud classification with image-depth pre-training" *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023: 22157-22167.
- [12] Afham M, et al. "CrossPoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.
- [13] Chen Z, et al. "Point cloud self-supervised learning via 3D to multi-view masked autoencoder." *arxiv preprint arxiv:2311.10887*, 2023.
- [14] Xie S, et al. "PointContrast: Unsupervised pre-training for 3D point cloud understanding." *European Conference on Computer Vision (ECCV)*. Springer, Cham, 2020.
- [15] Ren B, et al. "Bringing masked autoencoders explicit contrastive properties for point cloud self-supervised learning." *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2024.
- [16] Sarkar S D, Miksik O, Pollefeys M, et al. "CrossOver: 3D scene cross-modal alignment." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025: 8985-8994.
- [17] Phan A V, et al. "DGCNN: A convolutional neural network over large-scale labeled graphs." *Neural Networks* 108 (2018): 533-543.
- [18] Liu Y, Sun, X, et al. "S2ANet: combining local spectral and spatial point grouping for point cloud processing." *Virtual Reality & Intelligent Hardware*. 6 (2024): 267-279.
- [19] Karambakhsh A, Sheng B, Li P, et al. "SparseVoxNet: 3D object recognition with sparsely aggregation of 3-D dense blocks." *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 35(1): 532-546.
- [20] Ben-Shabat Y, Lindenbaum M, Fischer A. "3DMFV: Three-dimensional point cloud classification in real-time using convolutional neural networks, *IEEE Robot and Automation. Letter*. 2018, 3(4): 3145-3152.
- [21] Eckart B, Kim K, Kautz J. "HGM-R: Hierarchical Gaussian mixtures for adaptive 3D registration." *Proceedings of the European Conference on Computer Vision. (ECCV)*. Springer, Munich, Germany, 2018: 705-721.
- [22] Hertz A, et al. "PointGMM: A neural GMM network for point clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
- [23] Zhang X, Zhang S, Yan J. "Towards more diverse and challenging pre-training for point cloud learning: self-supervised

- cross reconstruction with decoupled views. " *arxiv preprint arxiv:2509.01250*, 2025.
- [24] Shi X, Zhao H, Jiang Y, et al. "GSFaceMorpher: High-fidelity 3D face morphing via Gaussian splatting." *Computer Animation and Virtual Worlds*, 2025, 36(3): e70036.
- [25] Wang H, et al. "Unsupervised point cloud pre-training via occlusion completion." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.
- [26] Yu H, Hao J, et al. "Rotation-invariant transformer for point cloud matching." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.
- [27] Chhipa P C, et al. "Depth contrast: Self-supervised pre-training on 3D maps for mining material classification." *European Conference on Computer Vision (ECCV)*. Springer, Cham, 2022.
- [28] Chen C, et al. "ClusterNet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [29] Cai M, Luo C, Lee Y J, et al. "Cross-Modal Self-Supervised Learning with Effective Contrastive Units for LiDAR Point Clouds." *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024: 9468-9475.
- [30] Han Z, et al. "View-inter-prediction-GAN: Unsupervised representation learning for 3D shapes by learning global shape memories to support local view predictions." *Proceedings of the AAAI conference on artificial intelligence*. 2019, 33(01).
- [31] Huang S, et al. "Spatio-temporal self-supervised representation learning for 3d point clouds." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.
- [32] Wang X, Hu X, Huang C, et al. "Self-Supervised Cross-Modal Learning for Image-to-Point Cloud Registration." *arxiv preprint arxiv:2509.15882*, 2025.
- [33] Han Z, et al. "Multi-angle point cloud-VAE: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction." *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.
- [34] Wang J, Lu X, Bennamoun M, et al. "Non-rigid point cloud registration via anisotropic hybrid field harmonization." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TRAM)*, 2025.
- [35] Koonce B. "ResNet 50." *Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization*. Aprss, Berkeley, CA, 2021. 63-72.
- [36] Ko J, et al. "3D GAN inversion with pose optimization." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023.
- [37] Afzali J, César Q C, Rossella A. "Latent GAN: using a latent space-based GAN for rapid forecasting of CFD models." *International Conference on Computational Science (ICCS)*. Springer, Cham, 2021.
- [38] Li J, Chen B M, Lee G H. "So-Net: Self-organizing network for point cloud analysis." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [39] Yang Y, et al. "FoldingNet: Point cloud auto-encoder via deep grid deformation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [40] Zhao Y, et al. "3D point capsule networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [41] Ji W, et al. "MRTNet: Multi-resolution temporal network for video sentence grounding." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.
- [42] Sauder J, Bjarne S. "Self-supervised deep learning on point clouds by reconstructing space." *Advances in Neural Information Processing Systems (NeurIPS)*, 2019,32.
- [43] Pang Y, et al. "Masked autoencoders for 3d point cloud self-supervised learning." *World Scientific Annual Review of Artificial Intelligence*. 2023,1: 2440001.
- [44] Jiang J, Zhao L, Lu X, et al. "DHGCN: Dynamic hop graph convolution network for self-supervised point cloud learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024, 38(11): 12883-12891.
- [45] Cheng Z, Wang Y. "Partial contrastive point cloud self-supervised representation learning." *Scientific Reports*, 2025, 15(1): 16182.
- [46] Yu X, Tang L, Rao Y, et al. "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022: 19313-19322.
- [47] Li H, Lin B, Zhang C, et al. "Mask-Point: automatic 3D surface defects detection network for fiber-reinforced resin matrix composites." *Polymers*, 2022, 14(16): 3390.
- [48] Li Y, Bu R, Sun M, et al. "PointCNN: Convolution on x-transformed points." *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, 31.
- [49] Liu Y, Fan B, Xiang S, et al. "Relation-shape convolutional neural network for point cloud analysis" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019: 8895-8904.
- [50] Ghose S, Li M, Qian Y, et al. "CLIP-based point cloud classification via point cloud to image translation." *International Conference on Pattern Recognition (ICPR)*. Springer, Cham, 2025.DOI:10.1007/978-3-031-78447-7_12.