View-GFN: A Novel View-Based Graph Convolution and Sampling Fusion Network for 3D Shape Recognition

Min Pang *, Jichao Jiao, Yingjian Zhang

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0848.v1

Keywords: 3D shape recognition; view-based methods; graph neural networks; hierarchical graph coarsening; multi-scale fusion

 doi:10.20944/preprints202604.0848.v1

Article

View-GFN: A Novel View-Based Graph Convolution and Sampling Fusion Network for 3D Shape Recognition

Min Pang [1,2,*], Jichao Jiao [2] and Yingjian Zhang [2]

[1] China Research Institute of Radiowave Propagation, No. 33 Xianshan East Road, Chengyang District, Qingdao City, China
[2] School of Electronic Engineering, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing, 100876, China
* Correspondence: pangmin2021@bupt.edu.cn; Tel.: +86-1830-021-7266

Abstract

Three-dimensional (3D) shape recognition is a fundamental task in computer vision, where view-based methods have recently achieved state-of-the-art performance. However, effectively capturing and exploiting the rich geometric correspondences between different views remains a key challenge, as such information is crucial for accurate shape representation. Existing methods often fall short in explicitly modeling these structured correlations, which limits their ability to fully leverage discriminative shape information. To address this limitation, we propose a novel View-based Graph Convolution and Sampling Fusion Network (View-GFN). View-GFN employs a hierarchical architecture that progressively coarsens the view-graph to learn multi-scale features. In this structure, views are treated as graph nodes, and a predefined-value strategy is introduced to initialize the adjacency matrix (AM) for constructing initial node correlations. For effective graph coarsening, we develop a novel view down-sampling method based on a cluster assignment matrix. Furthermore, a Graph Convolution and Sampling Fusion (CSF) module is designed to seamlessly integrate deep feature embeddings with the topological information derived from view down-sampling. Extensive experiments on benchmark datasets, including ModelNet40 and RGB-D, demonstrate that View-GFN achieves a superior recognition accuracy of 97.8%, surpassing previous methods while reducing the number of model parameters by nearly 50%. These results validate the superiority of our hierarchical fusion strategy in capturing multi-view geometric information both effectively and efficiently.

Keywords: 3D shape recognition; view-based methods; graph neural networks; hierarchical graph coarsening; multi-scale fusion

1. Introduction

Humans perceive the world in three dimensions; therefore, understanding and parsing 3D objects is a fundamental and crucial task in computer vision. In recent years, 3D shape recognition has emerged as a highly active research direction in this field, playing a pivotal role in practical applications such as autonomous driving, robotic perception, and virtual reality. With the rapid advancement of deep learning technologies, researchers have proposed a plethora of innovative architectures for 3D shape recognition and analysis. Depending on the underlying 3D data representations, these approaches can be broadly categorized into three main paradigms: voxel-based [1,2], point cloud-based [3–9], and multi-view-based methods [10–15].

Specifically, voxel-based methods extract features and recognize targets by discretizing the 3D space into regular 3D grids [1,2]. Point cloud-based methods represent targets as unordered sets of spatial points and directly process point cloud features using advanced geometric aggregation or semantic modeling networks [3,4]. For instance, recent studies have further introduced techniques such as skeleton-aware sampling [5], zero-shot geometry-driven aggregation [6], as well as the latest

 doi:10.20944/preprints202604.0848.v1

sample-adaptive auto-augmentation [7], multi-scale topological networks [8], and dynamic acoustic field fitting [9], significantly enhancing the accuracy and robustness of point cloud recognition. On the other hand, multi-view-based methods project 3D objects into a sequence of two-dimensional (2D) images and integrate the features of these multi-view images into a global 3D shape descriptor through deep recognition networks [10,11]. Compared to the former two paradigms, view-based methods generally exhibit superior performance in 3D recognition tasks. This is not only because they can acquire comprehensive and dense geometric and textural information from different perspectives via 2D images [12]; more importantly, this paradigm can seamlessly leverage pre-trained image network models that have proven exceptionally powerful and mature in the 2D vision domain [13]. Recent research has further extended this paradigm to heterogeneous dynamic graph representations [14] and cross-modal fusion of multi-view images and point clouds [15], demonstrating immense potential.

Despite the remarkable progress achieved by multi-view-based methods, efficiently and losslessly modeling the complex geometric correspondences across views remains a core bottleneck hindering further breakthroughs in recognition performance. Early multi-view fusion strategies (e.g., classic view pooling operations) typically treat all input views equally, ignoring the inherent spatial correlations and structured semantic information among them. To introduce relational modeling, Graph Convolutional Network (GCN)-based methods (e.g., view-GCN) have been proposed in recent years, which treat views as nodes for message passing. However, these graph topology-based methods still exhibit evident theoretical and practical deficiencies in their network evolution mechanisms. First, during the graph coarsening (i.e., view down-sampling) stage, existing methods predominantly rely on selective "hard sampling" based on feature scores (e.g., directly discarding lower-ranked view nodes). Such a crude dropping strategy not only leads to the irreversible loss of long-tail critical features but also severely destroys the global manifold topological structure of 3D objects. Second, the existing graph construction process often over-relies on predefined rigid viewpoint coordinates for the initialization of the adjacency matrix (AM). This restricted initialization mechanism lacks a global connectivity prior, making the model less robust to fluctuations in the number of input views and rendering it difficult to adaptively capture long-range dependencies across views. Furthermore, the feature updating of graph nodes and the evolution of the topological structure are usually decoupled. This fragmented architecture further restricts the network from fully exploiting deep discriminative shape information.

To overcome the limitations of existing methods in view relationship modeling and geometric information preservation, this paper proposes a novel View-based Graph Convolution and Sampling Fusion Network (View-GFN). Specifically, we introduce theoretical and structural innovations from three dimensions: graph initialization, graph coarsening, and feature fusion. Firstly, unlike traditional graph construction methods that rely on fixed spatial coordinates, we propose an adjacency matrix (AM) initialization strategy with a global connectivity prior. This strategy endows the initial graph with a global receptive field, which not only significantly enhances the information interaction capability of graph convolutions in the early stages of feature aggregation but also provides the model with strong adaptability to variations in the number of input views. Secondly, to address the issue that traditional selective sampling (e.g., Top-K dropping) easily disrupts the intrinsic geometric topology of 3D objects, we introduce a hierarchical graph coarsening method based on a clustering assignment matrix. By softly mapping semantically similar view features into super-nodes, this method effectively preserves the key geometric topological properties across views while eliminating redundant information. Finally, we design a Graph Convolution and Sampling Fusion (CSF) module, which breaks the barrier of independent graph updating and feature extraction, seamlessly integrating deep local feature embeddings with the coarsened macroscopic topological structure.

Our main contributions can be summarized as follows: (1) We propose a novel multi-view 3D shape recognition framework (View-GFN). As the core of this framework, we introduce a graph coarsening mechanism based on a clustering assignment matrix. By substituting traditional hard dropping strategies with a learnable soft-assignment mechanism, it achieves feature fusion rather than mere screening during view down-sampling, effectively mitigating the loss of discriminative

 doi:10.20944/preprints202604.0848.v1

information during dimensionality reduction from an information-theoretic perspective. (2) We design a Graph Convolution and Sampling Fusion (CSF) module that unifies feature embedding and graph topology evolution within the same framework. By jointly optimizing node feature updating and graph structure coarsening, it eliminates the error accumulation inherent in traditional two-stage methods, significantly improving the representation capability of multi-scale geometric features from an optimization theory perspective. (3) We propose an adjacency matrix (AM) initialization strategy based on a global connectivity prior. By constructing an initial dense graph structure with predefined values, this strategy equips the shallow graph convolutions with a global receptive field. From a graph signal processing perspective, this method enhances information aggregation efficiency and exhibits natural robustness to changes in the number of input views.

Extensive experiments on two benchmark datasets, ModelNet40 and RGB-D, demonstrate that View-GFN achieves a 97.8% recognition accuracy while reducing the total number of model parameters by nearly 50% compared to existing mainstream methods. This compellingly validates the excellent performance and practical value of the proposed hierarchical fusion strategy in efficiently capturing multi-view geometric information.

2. Related Work

In this section, we provide a systematic review of the literature closely related to our work from three perspectives: multi-view 3D shape recognition, graph construction and relational modeling, and hierarchical graph coarsening and pooling.

2.1. Multi-View 3D Shape Recognition

Converting 3D objects into 2D projections to leverage mature 2D convolutional neural networks (CNNs) for discriminative feature extraction has become a core paradigm in the field of 3D shape analysis. As a pioneering work, MVCNN [16] introduced a view pooling strategy that aggregates multi-view features via element-wise maximum operation to generate global shape descriptors. This work laid the foundation for multi-view methods, enabling 3D recognition tasks to fully benefit from 2D network models pre-trained on large-scale image datasets such as ImageNet.

Subsequent studies have pursued improvements in fusion strategies and feature extraction. GVCNN [17] introduced a group-view convolutional approach that partitions views into different groups based on feature similarity. MHBN [18] proposed harmonized bilinear pooling to capture second-order statistics across cross-view image patches. Furthermore, several works have focused on viewpoint optimization and sequence modeling. For instance, RotationNet [19] treats viewpoints as latent variables for joint optimization, achieving simultaneous improvement in both classification and pose estimation. Methods based on RNNs or LSTMs [20] attempt to capture spatial evolution patterns across view sequences using temporal models.

Despite significant progress, most of these methods rely on simple pooling operations or sequential aggregation, treating each view as an isolated image sample. This paradigm fails to explicitly establish structured topological relationships between views, thereby overlooking the rich geometric correspondence information embedded across different perspectives. This limitation motivated the introduction of graph neural networks (GNNs) into the multi-view domain. View-GCN [21] represents the first attempt to explicitly treat views as graph nodes and perform message passing through graph convolution, opening new directions for graph-driven multi-view fusion research.

2.2. Graph Construction and Relational Modeling

The performance of GNNs heavily depends on the quality of the initial graph topology. In multi-view 3D recognition, defining appropriate node adjacency relationships for a set of views constitutes a fundamental challenge. Existing graph-based methods primarily adopt two initialization strategies. The first is geometry-driven static graph construction, exemplified by View-GCN [21], which initializes the adjacency matrix (AM) using the physical 3D coordinates of camera viewpoints via the K-nearest neighbors (KNN) algorithm. While this approach introduces spatial priors, its fixed graph structure

 doi:10.20944/preprints202604.0848.v1

fails to reflect the dynamic semantic evolution of view relationships and exhibits high sensitivity to variations in the number of input views. The second strategy is semantic-driven dynamic graph construction. For example, Xu et al. [22] proposed a path aggregation graph network that dynamically constructs a view-relation graph by computing semantic correlations between view features. While this approach captures deep semantic relationships, it typically involves expensive pairwise similarity computation, incurring significant computational overhead.

Different from these methods, this paper proposes an AM initialization strategy based on a global connectivity prior. In contrast to methods relying on local geometric constraints [21] or high-overhead dynamic feature dependencies [22], our approach constructs a densely connected initial topology using pre-defined values. This design endows graph convolution with a global receptive field at shallow layers and eliminates dependence on static viewpoint coordinates, enabling the model to adaptively learn cross-view long-range dependencies while demonstrating inherent robustness to fluctuations in the number of input views.

2.3. Hierarchical Graph Coarsening and Pooling

Graph pooling is a fundamental technique for learning multi-scale graph representations. Existing methods can be broadly categorized into two families: node dropping and soft pooling. Representative node dropping methods, such as gPool [23] and SAGPool [24], learn scalar scores for nodes and deterministically retain high-scoring nodes. Despite their efficiency, this heuristic hard sampling mechanism exhibits notable limitations in 3D multi-view tasks. As seen in View-GCN [21], directly discarding view nodes may lead to irreversible loss of long-tail discriminative geometric features and disrupt the global manifold topology of 3D objects. Moreover, the selection process is often decoupled from feature extraction, limiting the effectiveness of end-to-end optimization.

On the other hand, general-purpose soft pooling methods such as DiffPool [25] and MinCutPool [26] introduce mapping mechanisms based on cluster assignment. However, these methods are primarily designed for generic graph data. When applied to densely connected multi-view graphs, their computational complexity often grows quadratically with the number of nodes, and they fail to utilize geometric priors specific to 3D vision tasks.

To tackle these challenges, we propose a hierarchical multi-view graph coarsening method based on a cluster assignment matrix. Our approach smoothly aggregates semantically similar view features into super-nodes through a learnable soft assignment mechanism, achieving dimensionality reduction while maximally preserving critical geometric topological properties. Building upon this, we design a graph convolution and sampling fusion (CSF) module that jointly optimizes feature embedding and topological evolution within a unified framework. This design effectively mitigates discriminative information loss from an information theory perspective and eliminates the error accumulation inherent in traditional two-stage methods from an optimization standpoint.

3. Methodology

3.1. Overview

In this section, we introduce View-GFN, a novel hierarchical graph fusion network for three-dimensional (3D) shape recognition. The network adopts a multi-stage abstraction architecture designed to capture multi-scale geometric features through progressive graph coarsening. Each level of the hierarchy defines a view-graph denoted as $\mathcal{G}^l = (\mathcal{V}^l, \mathcal{E}^l)$.

The initial view-graph at the first level is constructed based on M input views, where each view corresponds to a node in the graph. To define the initial correlations between nodes, we propose an initialization strategy based on a global connectivity prior. In contrast to traditional methods that rely

 doi:10.20944/preprints202604.0848.v1

on unstable viewpoint coordinates or local K-Nearest Neighbor (KNN) constraints, we initialize the initial adjacency matrix $A^1 \in \mathbb{R}^{M\times M}$ as a representation of a complete graph:

$$A^1_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases} \tag{1}$$

The core logic behind this design is that in multi-view recognition tasks, the number of input views (typically 12 or 20) is relatively limited. Adopting a complete graph topology ensures that each node attains a global receptive field during the first layer of graph convolution, thereby facilitating global information interaction at the early stages of feature learning. From the perspective of graph signal processing (GSP), a complete graph exhibits high spectral expansion, and the eigenvalue distribution of its Laplacian matrix favors the rapid smoothing and consensus building of global geometric signals in the shallow layers of the network.

The overall architecture of View-GFN is illustrated in Fig. 1. The network consists of a feature extraction module followed by three cascaded Graph Convolution and Sampling Fusion (CSF) modules. Each CSF module concurrently performs feature embedding and assignment matrix generation, enabling the hierarchical evolution of the graph structure through differentiable soft-clustering operations.

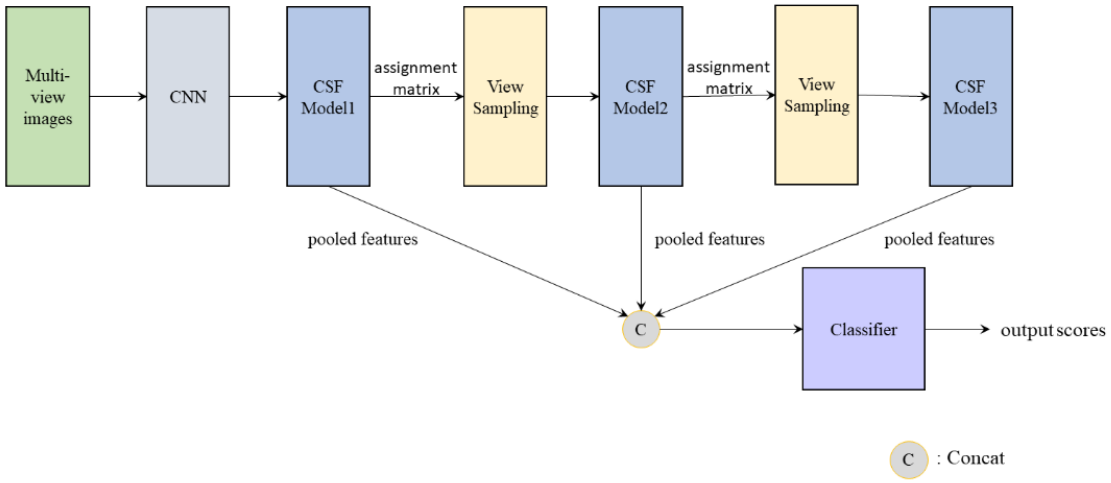

Figure 1. The overall architecture of the proposed View-GFN. The framework takes multi-view images as input, extracts initial features via a CNN backbone, and processes them through a hierarchical structure consisting of cascaded CSF models and view sampling modules. The final pooled features are concatenated to generate output scores.

3.2. Initial Feature Extraction

Given a sequence of multi-view images of a 3D object $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$, we employ ResNet-18, pre-trained on ImageNet and fine-tuned on the target dataset, as the backbone network for initial feature extraction. Each view image I_i is mapped to a c_0-dimensional discriminative feature vector. These vectors constitute the initial node feature matrix $X^1 \in \mathbb{R}^{m_1\times c_0}$ for the first-level graph, where $m_1 = M$ represents the initial number of nodes.

3.3. Cluster Assignment Based View Sampling

To achieve hierarchical compression of the graph structure, we need to aggregate m_l nodes at level l into m_{l+1} super-nodes at level $l+1$, such that $m_{l+1} < m_l$. This process is realized by learning a Cluster Assignment Matrix $S^l \in \mathbb{R}^{m_l\times m_{l+1}}$.

 doi:10.20944/preprints202604.0848.v1

Each row of S^l represents the assignment probability of a node in the current level to each super-node in the next level, satisfying the constraint $\sum_j S^l_{ij} = 1$. Utilizing S^l, the node feature matrix X^{l+1} and the adjacency matrix A^{l+1} for the next level are computed as follows:

$$X^{l+1} = (S^l)^{\mathrm{T}} Z^l \in \mathbb{R}^{m_{l+1} \times c} \tag{2}$$

$$A^{l+1} = (S^l)^{\mathrm{T}} A^l S^l \in \mathbb{R}^{m_{l+1} \times m_{l+1}} \tag{3}$$

where $Z^l \in \mathbb{R}^{m_l \times c}$ represents the enhanced node embeddings output by the CSF module. The theoretical foundation of this soft-assignment mechanism stems from spectral clustering and graph coarsening theories. From a graph-theoretic perspective, S^l acts as a differentiable low-pass filter that softly aggregates nodes that are both topologically proximal and semantically similar, thereby maximizing the preservation of the object's manifold structural features during dimensionality reduction.

3.4. Graph Convolution and Sampling Fusion Module (CSF)

The detailed internal architecture of the CSF module is illustrated in Fig. 2. It unifies the feature embedding process with the generation of the assignment matrix required for subsequent view down-sampling.

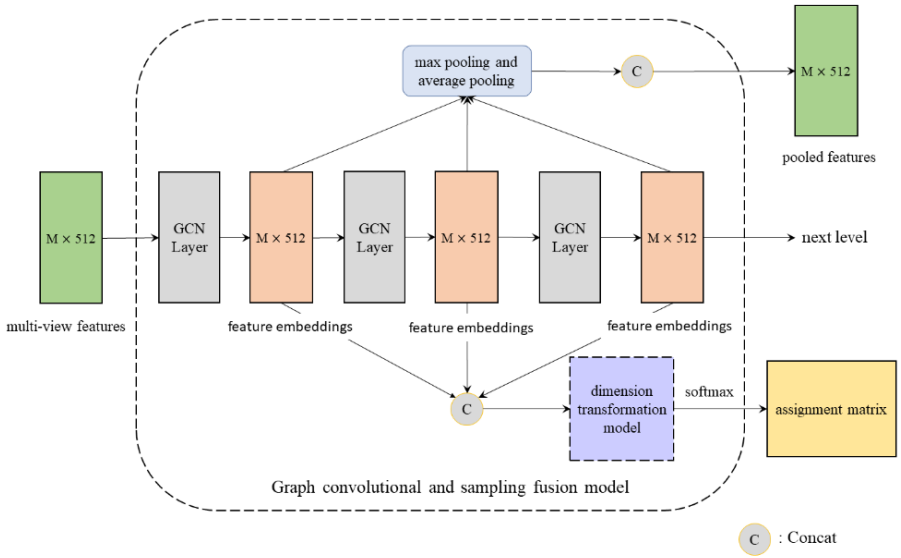

Figure 2. The detailed architecture of the Graph Convolution and Sampling Fusion (CSF) module. It demonstrates the joint optimization of multi-scale feature embeddings (via stacked GCN layers and mixed pooling) and the generation of the assignment matrix.

3.4.1. Feature Embedding

At level l, given the node features X^l and the adjacency matrix A^l, we utilize K stacked layers of graph convolutions for feature updating. To ensure numerical stability, each graph convolution layer is defined as:

$$F^l_k = \sigma\left(\tilde{D}^{l-1/2} \tilde{A}^l \tilde{D}^{l-1/2} H^l_k W^l_k\right) \tag{4}$$

where $\tilde{A}^l = A^l + I$ is the adjacency matrix with self-loops, $\tilde{D}^l$ is the degree matrix, $H^l_1 = X^l$, W^l_k is the weight matrix, and σ denotes the activation function. We define the output of the final convolution layer as the high-level semantic embedding matrix, i.e., $Z^l = F^l_K$.

 doi:10.20944/preprints202604.0848.v1

3.4.2. Assignment Matrix Generation

The generation of the assignment matrix S^l is performed synchronously with feature embedding. To fuse multi-scale information, we concatenate the intermediate outputs of each GCN layer and generate assignment weights through a non-linear mapping network:

$$S^l = \text{softmax}\Big(\text{MLP}\Big(\text{Concat}(F_1^l, \ldots, F_K^l)\Big)\Big) \tag{5}$$

By incorporating outputs from multiple neighborhood scales, the generated S^l is capable of perceiving neighborhood relationships across different levels, ensuring the structural fidelity of the clustering results in the topological space.

3.4.3. Multi-Scale Fusion and Receptive Field Analysis

The CSF module constructs a hierarchical representation of the current level through a joint optimization mechanism. We perform mixed pooling on the features F_k^l from each layer and concatenate them to obtain the output feature vector $O^l \in \mathbb{R}^{2Kc}$:

$$O^l = \text{Concat}\left(\left[\text{MaxPool}(F_k^l) \| \text{AvgPool}(F_k^l)\right]_{k=1}^{K}\right) \tag{6}$$

This fusion strategy effectively enforces the preservation of full-spectrum signals, ranging from local geometric details (shallow features) to macroscopic topological contours (deep features), thereby successfully overcoming the information bottleneck caused by limited perspectives in traditional methods.

3.5. Hierarchical Network Architecture and Loss Function

View-GFN employs a three-level cascaded structure. In our implementation, the first level typically consists of $m_1 = 20$ (or 12) views. After processing through the first CSF module, the graph is coarsened to m_2 super-nodes, and subsequently to m_3 super-nodes at the third level. The final global shape descriptor is formed by concatenating the multi-scale representations from all levels: $O = [O^1 \| O^2 \| O^3]$.

To guide S^l in learning reasonable topological clusters, we introduce a Link Prediction Loss as an auxiliary objective:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma \sum_{l=1}^{2} \| A^l - S^l (S^l)^\top \|_F^2 \tag{7}$$

This loss term constrains the super-nodes to reconstruct the adjacency relationships of the original graph as closely as possible, mathematically guaranteeing the topological consistency and structural fidelity of the soft-clustering process. The final descriptor O is then fed into a fully connected classifier for shape recognition.

4. Experiments and Results Analysis

In this section, we evaluate the performance of the proposed View-GFN through extensive experiments on benchmark datasets. We first introduce the experimental setup, followed by a comprehensive analysis of the model from multiple dimensions, including classification accuracy, robustness to view quantity, shape retrieval capabilities, and an ablation study.

4.1. Experimental Setup

4.1.1. Datasets and Evaluation Metrics

- **ModelNet40** [27]: This dataset contains 12,311 3D CAD models across 40 categories, with 9,843 models used for training and 2,468 for testing. Following the standard protocol, we render either 20 views (from the vertices of a dodecahedron) or 12 views (from a circular trajectory at an elevation of 30°) for each 3D object.

 doi:10.20944/preprints202604.0848.v1

- **RGB-D** [28]: A real-world dataset comprising 300 household objects across 51 categories. We adopt a 10-fold cross-validation strategy for evaluation on this dataset.

Evaluation Metrics: The primary metrics include *Instance Accuracy* (the ratio of correctly classified samples to the total number of samples), *Class Accuracy* (the arithmetic mean of accuracies across all classes), *mean Average Precision* (mAP, used for the retrieval task), the number of *Parameters (Params)*, and *Training Time* (forward and backward propagation time per epoch).

4.1.2. Implementation Details

We employ ResNet-18, pre-trained on ImageNet, as the initial feature extractor. View-GFN consists of three hierarchical levels with node scales set to $m_1 = 20$, $m_2 = 10$, and $m_3 = 5$, respectively. The network is trained using the SGD optimizer with a momentum of 0.9, a weight decay of 0.01, and a batch size of 20. During the fine-tuning of the feature extraction network, the initial learning rate is set to 0.01 and halved every 10 epochs. When training the entire network, the initial learning rate is set to 0.001 with a cosine annealing schedule. The balancing hyperparameter γ for the auxiliary loss is set to 0.1 by default. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2. Comparison with State-of-the-Art Methods

We compare View-GFN with various representative 3D shape recognition methods. Table 1 summarizes the classification results and model efficiency on the ModelNet40 dataset.

Table 1. Classification accuracy and model complexity comparison on ModelNet40.

Method	Type	Views	Inst Acc (%)	Class Acc (%)	Params (M)
MVCNN-new [29]	View Aggregation	12	95.0	92.4	-
GVCNN [17]	View Grouping	12	93.1	90.7	-
MHBN [18]	Bilinear Pooling	6	94.7	93.1	-
MVPNet [31]	Path Aggregation	20	97.9	96.8	-
RotationNet [19]	View Optimization	20	97.4	-	-
MLVCNN [30]	Multi-loop Views	36	94.2	-	-
View-GCN [21]	Graph Network	20	97.6	96.5	33.9
View-GFN (Ours)	**Graph Fusion**	**20**	**97.8**	**96.5**	**17.0**

Analysis: As shown in Table 1, View-GFN achieves an instance accuracy of **97.8%**, slightly higher than View-GCN (97.6%) and remarkably close to the current optimal MVPNet (97.9%). Its class accuracy is on par with View-GCN (96.5%). More importantly, the parameter size of View-GFN is only **50.1%** of that of View-GCN (17.0M vs. 33.9M), and the training time per epoch is reduced by approximately 46.7% (33.4s vs. 62.6s). This compellingly demonstrates that the proposed CSF module and soft-clustering mechanism are highly efficient in compressing redundant information while retaining discriminative geometric features.

4.3. Robustness to View Quantity

To validate the effectiveness of the "global connectivity prior AM initialization," we test the model using 12 input views on the real-world RGB-D dataset. The comparison results are presented in Table 2.

 doi:10.20944/preprints202604.0848.v1

Table 2. Classification accuracy on the RGB-D dataset (12 views).

Method	Views	Inst Acc (%)
MVCNN [16]	12	86.1
CFK [33]	120	86.8
MMDCNN [34]	120	89.5
MDSICNN [35]	120	89.9
View-GCN [21]	12	94.3
View-GFN (Ours)	**12**	**94.1**

Analysis: On the real-world RGB-D dataset, View-GFN achieves an accuracy of 94.1% with 12 views, which is comparable to View-GCN (94.3%), but with significantly fewer parameters (17.0M vs. 22.7M) and approximately 54.5% less training time (0.5s vs. 1.1s). This represents a substantial improvement over MVCNN (86.1%), which also uses 12 views. Furthermore, our 12-view View-GFN outperforms several methods that rely on 120 views (e.g., CFK, MMDCNN), highlighting its superior capability in utilizing view information efficiently.

4.4. Shape Retrieval Performance

Table 3 demonstrates the retrieval performance on the ModelNet40 dataset, evaluated using mean Average Precision (mAP).

Table 3. Retrieval task performance comparison on ModelNet40 (mAP).

Method	mAP (%)
GVCNN [17]	85.7
MVCVT [32]	95.4
MLVCNN [30]	92.8
MVPNet [31]	97.4
View-GFN (Ours)	**97.7**

Analysis: View-GFN achieves an mAP of **97.7%** in the retrieval task, outperforming MVPNet (97.4%) and MVCVT (95.4%), and significantly surpassing the earlier GVCNN (85.7%). This indicates that the multi-scale features extracted by the CSF module are not only discriminative for classification but also possess excellent semantic clustering properties, enabling the generation of high-quality global shape descriptors.

4.5. Ablation Study

Table 4 dissects the contribution of each core component to the overall performance based on the ModelNet40 dataset with 20 views.

 doi:10.20944/preprints202604.0848.v1

Table 4. Ablation study of View-GFN core components (ModelNet40, 20 views).

Configuration	Inst Acc (%)	Class Acc (%)	Description
View-GFN-FPS	96.5	95.2	Replace soft-clustering with Farthest Point Sampling (FPS)
View-GFN-SEP	97.7	96.5	Decouple feature embedding and assignment matrix generation
View-GFN-A1	97.4	96.2	AM considers only 3 nearest neighbor nodes
View-GFN-A2	97.5	96.1	AM initialized with view coordinate encoding
View-GFN (Full)	**97.8**	**96.4**	**Full model (Global AM + Soft-clustering + CSF)**

Analysis & Conclusion:

1. **Soft-clustering vs. Hard Sampling**: The instance accuracy of View-GFN-FPS drops by 1.3% (from 97.8% to 96.5%), proving that soft-clustering based on the assignment matrix retains discriminative information far better than Farthest Point Sampling.
2. **CSF Synchronous Fusion**: View-GFN-SEP performs almost on par with the full model (instance accuracy is only 0.1% lower), but incurs a significant increase in parameters. This demonstrates that the CSF module substantially reduces model complexity while maintaining high accuracy.
3. **AM Initialization Strategy**: The instance accuracies of View-GFN-A1 (local adjacency) and View-GFN-A2 (coordinate encoding) are 0.4% and 0.3% lower than the full model, respectively. This firmly validates the superiority of our global connectivity prior and predefined initial values.

5. Conclusions

This paper presents View-GFN, a novel multi-view graph convolution network for efficient 3D shape recognition. Our proposed method leverages a cluster assignment matrix for soft-clustering view down-sampling and introduces a Graph Convolution and Sampling Fusion (CSF) module to jointly optimize feature embedding and topological evolution. Furthermore, we propose a global connectivity prior initialization method for the adjacency matrix (AM), significantly enhancing the representation capability of shallow graph convolutions. Experimental results on benchmark datasets demonstrate the superiority of our approach, achieving state-of-the-art recognition accuracy while reducing the model parameters by nearly 50%. Although View-GFN demonstrates excellent performance, the dense connectivity initialization may introduce additional memory overhead when scaling to an extremely large number of views. In future work, we plan to explore sparse attention mechanisms to further optimize computational efficiency and extend this framework to more complex 3D scene understanding tasks.

References

1. Tu, T.; Chen, P.; Zhang, L. ImGeoNet: Image-induced Geometry-aware Voxel Representation for Multi-view 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023; pp. 6996–7007.
2. Xu, C.; Wu, B.; Hou, J.; et al. NeRF-Det: Learning Geometry-Aware Volumetric Representation for Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023.
3. Ding, D.; Wang, Z.; Xiong, H. Robust point cloud classification via semantic and structural modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023.

 doi:10.20944/preprints202604.0848.v1

4. Ben-Shabat, Y.; Gould, S. 3DInAction: Understanding Human Actions in 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024; pp. 19978–19987.
5. Chen, Y.; Liu, S.; Shen, X. Learnable Skeleton-Aware 3D Point Cloud Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023.
6. Li, Z.; Xu, C.; Leng, B. Geometrically-driven Aggregation for Zero-shot 3D Point Cloud Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024.
7. Wang, Y.; Chen, X.; Cao, L.; et al. Towards Robust Point Cloud Recognition With Sample-Adaptive Auto-Augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3003–3017.
8. Zhang, L.; Wang, Y.; Liu, H. Enhancing 3D Point Cloud Classification with ModelNet-R and Point-SkipNet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2025.
9. Liu, H.; Zhang, L.; Wang, Y. Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2025.
10. Wang, S.; Jiang, L.; Wu, Z.; et al. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023; pp. 3621–3631.
11. Zong, Z.; Song, G.; Liu, Y. Temporal Enhanced Training of Multi-view 3D Object Detector via Historical Object Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023; pp. 3781–3790.
12. Chen, H.; Wang, S.; Wu, Z. Pixel-Aligned Recurrent Queries for Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023.
13. Jia, X.; Zhang, Y.; Wu, B.; et al. PointCert: Point Cloud Classification with Deterministic Certified Robustness Guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023.
14. Zhao, M.; Li, H.; Tang, J. Cross-Modal 3D Shape Retrieval via Heterogeneous Dynamic Graph Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**.
15. Wu, J.; Zhang, L.; Liu, Y. Cross-Modal 3D Representation with Multi-View Images and Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2025.
16. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015; pp. 945–953.
17. Feng, Y.; Zhang, Z.; Zhao, X.; Ji, R.; Gao, Y. GVCNN: Group-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018; pp. 264–272.
18. Yu, T.; Meng, J.; Yuan, J. Multi-view harmonized bilinear network for 3D object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018; pp. 186–194.
19. Kanezaki, A.; Matsushita, Y.; Nishida, Y. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018; pp. 5010–5019.
20. Han, Z.; Shang, M.; Liu, Y.S.; et al. SeqViews2SeqLabels: Learning 3D global features via aggregating sequential views by RNN with attention. *IEEE Trans. Image Process.* **2018**, *28*, 658–672.
21. Wei, X.; Yu, R.; Sun, J. View-GCN: View-based graph convolutional network for 3D shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020; pp. 1850–1859.
22. Xu, M.; Chen, H.; Wang, Z. PAGNet: Path aggregation graph network for multi-view 3D shape recognition. *Knowl.-Based Syst.* **2021**, *229*, 107338.
23. Gao, H.; Ji, S. Graph U-Nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019; pp. 2083–2092.
24. Lee, J.; Lee, I.; Kang, J. Self-attention graph pooling. In *Proceedings of the International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019; pp. 3734–3743.

 doi:10.20944/preprints202604.0848.v1

25. Ying, Z.; You, J.; Morris, C.; et al. Hierarchical graph representation learning with differentiable pooling. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2018; pp. 4800–4810.
26. Bianchi, F.M.; Grattarola, D.; Alippi, C. Spectral clustering with graph neural networks for graph pooling. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria, 2020; pp. 874–883.
27. Wu, Z.; Song, S.; Khosla, A.; et al. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015; pp. 1912–1920.
28. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view RGB-D object dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, 2011; pp. 1817–1824.
29. Su, J.-C.; Gadelha, M.; Wang, R.; Maji, S. A deeper look at 3D shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, Munich, Germany, 2018.
30. Jiang, J.; Bao, D.; Chen, Z.; Zhao, X.; Gao, Y. MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019; pp. 8513–8520.
31. Xu, L.; Cui, Q.; Xu, W.; Chen, E.; Tong, H.; Tang, Y. Walk in views: Multi-view path aggregation graph network for 3D shape analysis. *Information Fusion* **2024**, *103*, 102131.
32. Li, J.; Liu, Z.; Li, L.; Lin, J.; Yao, J.; Tu, J. Multi-view convolutional vision transformer for 3D object recognition. *J. Vis. Commun. Image Represent.* **2023**, *95*, 103906.
33. Cheng, Y.; Cai, R.; Zhao, X.; Huang, K. Convolutional Fisher kernels for RGB-D object recognition. In *Proceedings of the 2015 International Conference on 3D Vision (3DV)*, IEEE, 2015; pp. 135–143.
34. Rahman, M.M.; Tan, Y.; Xue, J.; Lu, K. RGB-D object recognition with multimodal deep convolutional neural networks. In *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017; pp. 991–996.
35. Asif, U.; Bennamoun, M.; Sohel, F.A. A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2051–2065.