

Multi-View Aggregation Transformer with Contrastive Learning for 3D Shape Retrieval

Ze Zhang

Harbin Institute of Technology

Xiaojun Wu

wuxj@hit.edu.cn

Harbin Institute of Technology

Chenxi Wu

Harbin Institute of Technology

Guoyuan Liang

Chinese Academy of Sciences

Research Article

Keywords: 3D shape retrieval, Multi-view aggregation, Vision Transformer, Contrastive learning, Absolute cosine similaritys

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10363998/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multi-View Aggregation Transformer with Contrastive Learning for 3D Shape Retrieval

Ze Zhang¹, Xiaojun Wu^{1*}, Chenxi Wu¹, Guoyuan Liang²

^{1*}School of Intelligent Science and Engineering, Harbin Institute of Technology Shenzhen, Xili, Shenzhen, 518055, Guangdong, China.

²Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Xili, Shenzhen, 518055, Guangdong, China.

*Corresponding author(s). E-mail(s): wuxj@hit.edu.cn;

Contributing authors: 22s153129@stu.hit.edu.cn;

25s165024@stu.hit.edu.cn; gy.liang@siat.ac.cn;

Abstract

3D shape retrieval plays a critical role in mechanical design, digital heritage, and multimedia systems, where learning discriminative multi-view representations remains a core challenge. This paper presents Multi-View Aggregation Transformer (MVAT), a geometry-aligned Transformer-based multi-view aggregation framework for 3D shape retrieval. MVAT uses a vision Transformer backbone to extract single-view features and introduces hierarchical multi-view attention to model global, face-level, and vertex-level view correlations derived from a regular dodecahedron-based camera layout. To improve multi-view feature fusion and retrieval discriminability, we further design Channel Attention Aggregation (CAA) and Residual Channel Attention (RCA) modules for signed channel modulation and bounded residual feature refinement. In addition, we introduce an absolute cosine similarity metric for supervised contrastive retrieval learning. The proposed metric treats opposite feature directions as equivalent and admits an orthogonal line-prototype configuration for sign-invariant class separation. We provide a geometric feasibility analysis of this property and show its practical benefit for multi-class retrieval embeddings. Extensive experiments on ModelNet40, ModelNet10, and a mechanical component benchmark show that MVAT achieves strong retrieval performance with 93.2%, 95.1%, and 95.4% mAP, respectively. Comprehensive ablation studies validate the effectiveness of the hierarchical attention design, channel attention modules, absolute cosine similarity, and the three-stage training strategy.

Keywords: 3D shape retrieval, Multi-view aggregation, Vision Transformer, Contrastive learning, Absolute cosine similarity

1 Introduction

3D shape retrieval has become a fundamental task in computer-aided design, computer graphics, and virtual reality, driven by the rapid growth of 3D data applications. Traditional methods rely on hand-crafted descriptors such as SPH [1], LFD [2], and PANORAMA [3], which fail to capture complex geometric structures and high-level semantics. With the development of deep learning, data-driven methods have dominated this field, including view-based [4–6], point-based [7, 8], voxel-based [9], and multi-modal approaches [10, 11].

Among these, view-based methods are widely adopted for their efficiency and robustness, bridging the gap between mature 2D visual representation learning and 3D geometric understanding. Early works such as MVCNN [4] use view pooling for feature aggregation, while subsequent studies introduce group convolution [5] and graph reasoning [6]. However, most CNN-based methods suffer from two core limitations: Limited receptive fields that cannot effectively model long-range dependencies among multi-view features; Simple pooling or fusion strategies that ignore the inherent geometric symmetry of multi-view camera layouts, leading to suboptimal feature representations.

The success of Vision Transformer (ViT) [12] and contrastive learning [13] provides new insights for visual representation learning. ViT captures global dependencies via self-attention, while contrastive learning enhances discriminability by pulling similar samples close and pushing dissimilar ones apart. Despite these advances, directly applying ViT and contrastive learning to 3D shape retrieval faces three key unresolved challenges: (1) Lack of geometry-aligned hierarchical attention. Directly applying ViT to multi-view 3D retrieval ignores the inherent geometric symmetry of multi-view camera layouts, failing to align neural network inductive bias with 3D shape properties. (2) Suboptimal multi-view feature aggregation. Simple pooling or sigmoid-based channel attention strategies restrict aggregated features to non-negative space, which limits the performance of angle-based similarity metrics and leads to suboptimal discriminative representations. (3) Directional over-constraint in contrastive retrieval learning. Conventional cosine similarity distinguishes between opposite feature directions, imposing stricter angular constraints than necessary for supervised retrieval, where the goal is to retrieve semantically or geometrically similar shapes rather than to preserve feature-vector orientation.

To address these limitations, we propose Multi-View Aggregation Transformer (MVAT), a geometry-aligned Transformer-based multi-view aggregation framework for 3D shape retrieval. MVAT extracts single-view features using a pre-trained visual Transformer backbone and aggregates them through hierarchical attention modules that explicitly exploit the geometric structure of a regular dodecahedron-based camera layout. In addition, we introduce absolute cosine similarity (ACS) and a three-stage supervised contrastive training strategy to improve feature compactness and

retrieval stability. From a neural representation learning perspective, we provide a geometric feasibility analysis showing that ACS admits an orthogonal line-prototype configuration for sign-invariant class separation, which helps explain its effectiveness in high-dimensional multi-class retrieval scenarios. The main contributions are summarized as follows:

- **Geometry-aligned hierarchical attention.** We propose a Multi-View Attention (MVA) module that models global, face-level, and vertex-level view correlations according to the regular dodecahedron-based camera layout. This hierarchical design explicitly encodes geometric inductive bias into multi-view feature aggregation and improves rotational robustness for 3D shape representation.
- **Signed channel attention for discriminative fusion.** We design Channel Attention Aggregation (CAA) and Residual Channel Attention (RCA) modules for discriminative multi-view feature fusion. CAA aggregates multi-view features through full-space signed channel modulation via $\tanh$ activation, while RCA performs residual channel-wise refinement with bounded scaling in $(0, 2)$, enabling stable progressive feature enhancement across stacked refinement blocks without uncontrolled amplification.
- **Sign-invariant contrastive retrieval learning.** We introduce an absolute cosine similarity (ACS) metric for supervised contrastive 3D shape retrieval. ACS measures similarity in the real projective space $\mathbb{RP}^{C-1}$ and admits an orthogonal line-prototype configuration for sign-invariant class separation. Combined with a three-stage training pipeline, the proposed method achieves strong retrieval performance on both general object and mechanical component benchmarks.

2 Related Work

2.1 View-Based 3D Shape Retrieval

Early 3D shape retrieval methods relied heavily on hand-crafted feature descriptors. SPH [1] discussed challenges in 3D shape retrieval due to rotational differences and proposes using spherical harmonics for creating rotation-invariant descriptors. LFD [2] proposed a visual similarity-based 3D shape retrieval system that uses orthogonal projections encoded with Zernike moments and Fourier descriptors. PANORAMA [3] introduced a novel 3D shape descriptor that uses panoramic views projected onto three principal axes of an object, enhanced by 2D Fourier and Wavelet Transforms. These hand-crafted methods are limited by their inability to capture high-level semantic and geometric features of complex 3D shapes.

The advancement of deep learning in computer vision has led to the development of various techniques for visual feature extraction and aggregation, particularly benefiting 3D shape retrieval. View-based 3D shape retrieval involves setting up virtual cameras to capture rendered views of a 3D model from different angles, then aggregating features from each view to create a global feature for retrieval. One notable innovation is MVCNN [4], which pioneered the use of CNN for extracting features from single views and employed view pooling to aggregate features from multiple views, outperforming traditional hand-crafted descriptors by a large margin. Following MVCNN,

GVCNN [5] introduced a group fusion method for feature aggregation, while View-GCN [6] integrated graph neural network to model view correlations. VAM-IAM [14] further incorporated attention mechanisms into retrieval networks to enhance feature discriminability.

Despite these advances, existing view-based methods still have critical limitations: Most CNN-based methods have limited receptive fields, failing to model long-range inter-view dependencies; Existing attention designs do not leverage the inherent geometric symmetry of multi-view camera layouts, leading to mismatched inductive bias; Simple pooling or fusion strategies cannot fully preserve discriminative information across views, resulting in suboptimal global representations. Our work addresses these limitations by proposing a transformer-based multi-view aggregation framework with hierarchical attention aligned to multi-view geometric symmetry.

2.2 Attention Mechanism

The attention mechanism originated in natural language processing, and its ability to adaptively encode informative features has been widely adopted across deep learning tasks. Its core idea is to guide the network to focus on critical input parts by assigning learnable weights. Our work specifically leverages two complementary attention mechanisms: channel attention and self-attention.

Channel attention allows the network to adaptively emphasize or suppress information from different channels by learning the importance of each channel. SENet [15] pioneered the channel attention mechanism in computer vision, while CBAM [16] further enhanced it by combining max-pooling and average-pooling operations. GLAM [17] extended channel attention to image retrieval tasks, demonstrating its efficacy in enhancing feature discriminability for retrieval scenarios. However, existing channel attention designs for multi-view 3D retrieval often ignore the distribution of aggregated features, with activation functions restricting features to non-negative space, which limits the performance of angle-based similarity metrics. Our CAA and RCA modules address this issue with tailored activation and residual designs.

Self-attention enables the network to capture global dependency relationships by considering information from all positions in the input sequence simultaneously. Initially introduced in Transformer [18] for NLP tasks, ViT [12] was the first to integrate Transformer encoders into computer vision for image feature extraction. Subsequent works such as Swin Transformer [19] introduced sliding window self-attention, and MaxViT [20] proposed multi-axis self-attention for efficient global modeling. Vision Transformer often surpasses traditional CNNs in feature extraction capabilities, especially for long-range dependency modeling. However, directly applying ViT to multi-view 3D retrieval lacks hierarchical design tailored to multi-view geometric structures. Our MVA module fills this gap by designing hierarchical self-attention aligned to the symmetric camera layout.

2.3 Contrastive Learning

Contrastive learning learns discriminative data representations by comparing similarities and differences among samples. Its core idea involves bringing features of similar

samples closer together while pushing features of dissimilar samples apart, which is particularly suitable for matching and retrieval tasks.

In the domain of 3D shape retrieval, contrastive learning has been widely explored. TCL [21] proposed a contrastive learning loss that enhances feature discrimination by learning class-specific centers. ATCL [22] optimized cosine distances between shape features to enhance inter-class discriminability and intra-class compactness. AMTCL [23] integrated group-view similarity learning and contrastive loss to improve the robustness of MVCNN. However, these works mainly adopt conventional cosine similarity, which distinguishes the same and opposite feature directions. In supervised contrastive retrieval, such directional sensitivity may impose a stricter angular constraint than necessary, especially when the goal is to retrieve semantically or geometrically similar shapes rather than to preserve feature-vector orientation.

One of the most notable advances in contrastive learning is CLIP [13], which proposed language-image contrastive learning with powerful zero-shot transfer capabilities. Although CLIP has demonstrated strong performance in various retrieval tasks, its application in view-based 3D shape retrieval remains under-explored, especially regarding the adaptation of contrastive learning to structured multi-view geometry and retrieval-oriented similarity metrics. Inspired by CLIP, our work tailors supervised contrastive learning to 3D shape retrieval by introducing an absolute cosine similarity metric with a sign-invariant geometric interpretation, together with a three-stage training pipeline for stable feature learning.

A foundational work in supervised contrastive learning is SupCon [24], which generalizes contrastive learning to the fully supervised setting by leveraging label information to construct positive pairs beyond data augmentation. The SupCon loss takes the following form:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{f}_i \cdot \mathbf{f}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{f}_i \cdot \mathbf{f}_a / \tau)}, \quad (1)$$

where $P(i)$ is the set of positives for sample i , $A(i)$ is the set of all other samples in the batch, and τ is a learnable temperature. Our symmetric cross-entropy loss, introduced in Section 3.6.2, shares the same objective of intra-class compactness and inter-class separability with SupCon, but differs in three key aspects: (1) we adopt absolute cosine similarity instead of standard dot-product similarity, removing the directional constraint on feature vectors; (2) We keep the symmetric notation for consistency with CLIP-style contrastive formulations. In the intra-modal retrieval setting, since both the similarity matrix and the label matrix are symmetric, the row-wise and column-wise terms reduce to an equivalent objective; and (3) we do not use a temperature parameter, reducing hyperparameter sensitivity in the retrieval setting.

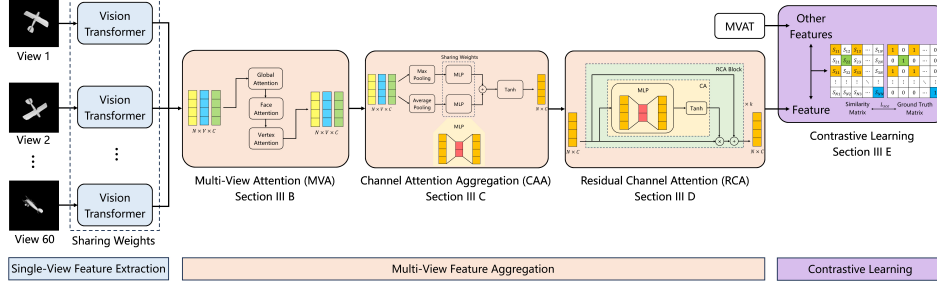


Fig. 1: Overview of the proposed MVAT framework. The network consists of two core parts: single-view feature extraction and multi-view feature aggregation. It is an end-to-end network that takes 60 views of a 3D model as input and outputs a discriminative global feature of the 3D model. During training, the global feature is compared with features of other 3D models in the same batch to compute the absolute cosine similarity matrix S , and the symmetric cross-entropy loss L_{sce} is calculated based on the ground truth matrix GT . During testing, the similarity between the query feature and all features in the test set is computed, and retrieval results are obtained by similarity ranking.

3 Method

3.1 Overall Framework

The proposed 3D shape retrieval network, termed Multi-View Aggregation Transformer (MVAT), is depicted in Fig. 1. The framework follows a two-branch design: single-view feature extraction and hierarchical multi-view feature aggregation.

Given a 3D model, we first render 60 multi-view images via a symmetric regular dodecahedron-based camera layout (detailed in Section 3.2). These 60 view images are fed into a pre-trained MaxViT backbone to extract 60 single-view features, each with dimension $C = 512$. The multi-view feature sequence is then processed by three cascaded core modules:

- Multi-View Attention (MVA) module: models hierarchical global, face-level, and vertex-level inter-view correlations, capturing multi-scale geometric dependencies across view-orientation samples.
- Channel Attention Aggregation (CAA) module: fuses the multi-view features into a compact global representation through signed channel-wise modulation, enhancing angular discriminability for retrieval.
- Residual Channel Attention (RCA) module: refines the global feature with stacked residual channel attention blocks, using bounded residual scaling for progressive feature enhancement.

The final output of MVAT is a discriminative global feature vector for the input 3D model, which is used for contrastive training and retrieval inference. In the following subsections, we detail each component of the framework in the order of data flow.

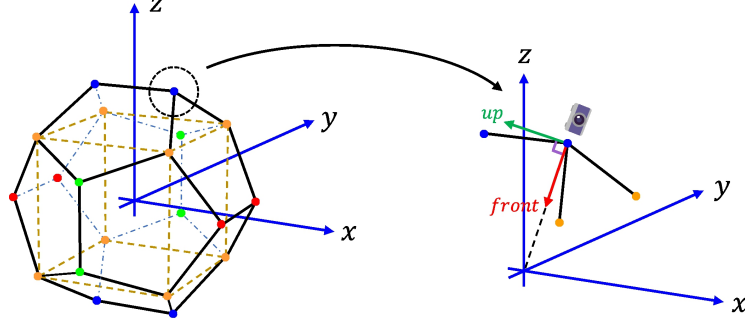


Fig. 2: Position and in-plane orientation of one view-orientation sample based on the regular dodecahedron layout.

3.2 Symmetric Multi-View Acquisition

For view-based 3D shape retrieval, the raw mesh data of a 3D model are first converted into multiple rendered images. The placement and orientation of virtual cameras are critical to retrieval robustness, especially when 3D models appear in arbitrary poses. To provide uniform spherical coverage and a structured geometric layout for attention modeling, we adopt a regular dodecahedron-based camera configuration, as shown in Fig. 2.

A regular dodecahedron consists of 12 regular pentagonal faces and 20 vertices. We first normalize each 3D model into a unit sphere and enclose it with a regular dodecahedron. A spatial viewpoint is placed at each vertex of the dodecahedron, and the viewing direction points toward the center of the model. To capture orientation-dependent appearance variations, three in-plane camera orientations are assigned to each spatial viewpoint according to the three adjacent edges connected to the corresponding vertex. In this way, the acquisition scheme produces 20 spatial viewpoints and three in-plane orientations at each viewpoint, resulting in 60 view-orientation samples.

This configuration provides two advantages. First, the 20 vertex viewpoints offer approximately uniform coverage of the object surface. Second, the three in-plane orientations at each vertex preserve local orientation information, which is useful for distinguishing geometrically similar 3D shapes. The resulting 60 rendered images are used as the input sequence of MVAT.

3.3 Multi-View Attention (MVA) Module

To fully exploit the inter-view correlations and geometric symmetry of the multi-view layout, we propose the Multi-View Attention (MVA) module, which includes three complementary hierarchical components: Global Attention (GA), Face Attention (FA), and Vertex Attention (VA). All components follow the modified Transformer encoder structure with pre-layer normalization for stable training [25]. Different attention masks are applied to model hierarchical view dependencies, as illustrated in Fig. 3.

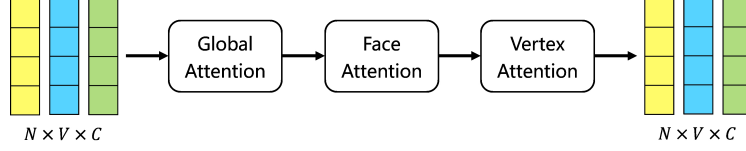


Fig. 3: Structure of the Multi-View Attention (MVA) module. N represents the batch size, V represents the number of views, and C represents the feature dimension.

The input of MVA is a multi-view feature tensor $F_{\text{in}} \in \mathbb{R}^{N \times V \times C}$, where N is the batch size, $V = 60$ is the number of views, and C is the feature dimension. The output is a refined multi-view feature tensor with the same shape, capturing hierarchical inter-view correlations.

3.3.1 Global Attention (GA)

GA computes full self-attention without any masking, enabling every view to communicate with all other views. It captures global inter-view dependencies and provides a holistic understanding of the 3D shape’s global structure. The attention mask is defined as an all-one matrix:

$$M_G = \mathbf{1} \in \mathbb{R}^{V \times V}, \quad (2)$$

where $V = 60$ denotes the number of views.

3.3.2 Face Attention (FA)

Based on the regular dodecahedron camera layout, we construct 12 face-level groups corresponding to the 12 pentagonal faces of the dodecahedron. Each face group contains the view-orientation samples associated with the five vertices of that face. Since each vertex of a regular dodecahedron is shared by three adjacent faces, the face-level groups are not disjoint; instead, they provide overlapping local neighborhoods that encode planar geometric adjacency.

FA computes masked self-attention within each face-level neighborhood, enhancing correlations among views located on the same pentagonal face and capturing local surface-consistent geometric context. The face-level attention mask M_F is defined as:

$$M_F(i, j) = \begin{cases} 1, & \text{if views } i \text{ and } j \text{ belong to at least one common face-level group,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This overlapping neighborhood design allows information to propagate across adjacent face structures while preserving the geometric locality induced by the dodecahedron layout. The structure of face self-attention is illustrated in Fig. 4.

3.3.3 Vertex Attention (VA)

For each vertex of the dodecahedron, three view-orientation samples are generated by changing the in-plane camera orientation while keeping the same spatial viewpoint and viewing direction. These three samples form a vertex-level group. Since there are

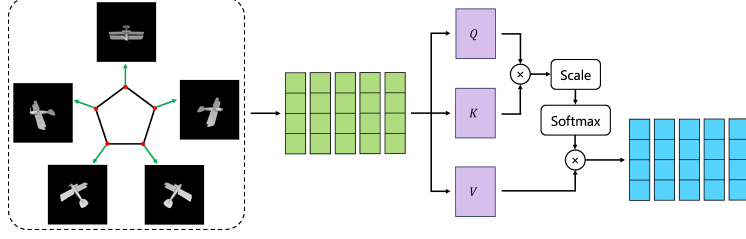


Fig. 4: Structure of Face Self-Attention on each dodecahedron face. Each red point represents a vertex, each green arrow represents the up vector of a virtual camera, and the multiplication sign represents matrix multiplication.

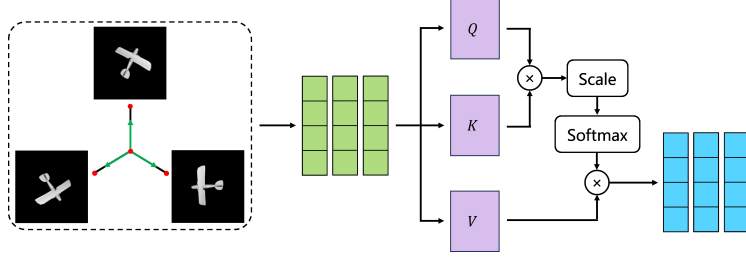


Fig. 5: Structure of Vertex Self-Attention on each dodecahedron vertex. Each red point represents a vertex, each green arrow represents the up vector of a virtual camera, and the multiplication sign represents matrix multiplication.

20 vertices, VA defines 20 disjoint vertex-level groups, each containing three view-orientation samples.

VA models fine-grained correlations among different in-plane orientations at the same spatial viewpoint, allowing the network to capture local orientation-sensitive appearance variations. The vertex-level attention mask M_V is defined as:

$$M_V(i, j) = \begin{cases} 1, & \text{if views } i \text{ and } j \text{ belong to the same vertex-level group,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The structure of vertex self-attention is illustrated in Fig. 5.

By cascading the three components in the order of $GA \rightarrow FA \rightarrow VA$, MVA progressively models inter-view dependencies from global shape-level context to face-level geometric neighborhoods and vertex-level orientation variations. This hierarchical design introduces a geometry-aware inductive bias into the attention mechanism and improves the robustness of multi-view 3D shape representation.

3.4 Channel Attention Aggregation (CAA) Module

After modeling inter-view correlations with MVA, we propose the Channel Attention Aggregation (CAA) module to aggregate the refined multi-view feature tensor $F \in$

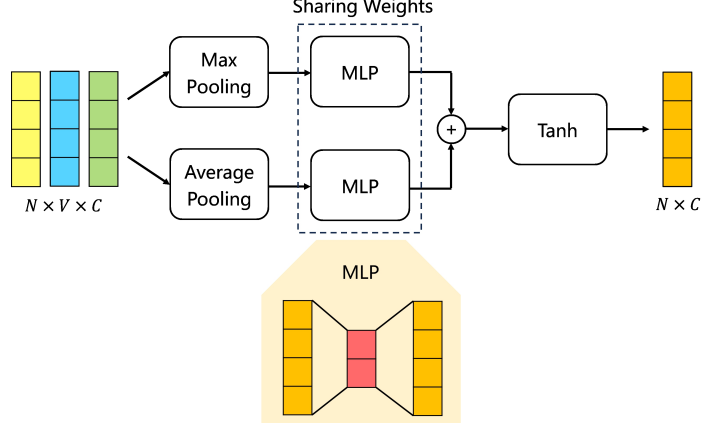


Fig. 6: Structure of the Channel Attention Aggregation (CAA) module. N represents the batch size, V represents the number of views, and C represents the feature dimension.

$\mathbb{R}^{N \times V \times C}$ into a compact global feature $F_{\text{out}} \in \mathbb{R}^{N \times C}$. CAA combines view-wise max-pooling and average-pooling to capture complementary statistics across views, and then performs channel-wise modulation in the full feature space.

The aggregation process is formulated as:

$$\begin{aligned}
 F_{\text{max}} &= \max_{v \in [1, V]} F_{:,v,:} \in \mathbb{R}^{N \times C}, \\
 F_{\text{avg}} &= \frac{1}{V} \sum_{v=1}^V F_{:,v,:} \in \mathbb{R}^{N \times C}, \\
 F_{\text{pool}} &= \frac{1}{2} (F_{\text{max}} + F_{\text{avg}}), \\
 A_{\text{ca}} &= \tanh(\text{MLP}(F_{\text{max}}) + \text{MLP}(F_{\text{avg}})), \\
 F_{\text{out}} &= F_{\text{pool}} \odot (\mathbf{1} + A_{\text{ca}}),
 \end{aligned} \tag{5}$$

where $\odot$ denotes element-wise multiplication. The MLP consists of two linear layers with a bottleneck hidden dimension of $C/4$ and a non-linear activation between them.

Different from sigmoid-based channel attention, which restricts attention responses to the non-negative range, the tanh activation produces signed channel modulation in $(-1, 1)$. This design allows the aggregated representation to preserve both positive and negative feature responses, which is beneficial for angle-based retrieval metrics such as cosine similarity and the proposed absolute cosine similarity.

3.5 Residual Channel Attention (RCA) Module

After CAA, we use the Residual Channel Attention (RCA) module to further refine the global feature. Direct channel attention with signed activation may compress the dynamic range of features when deeply stacked. To alleviate this issue, RCA adopts

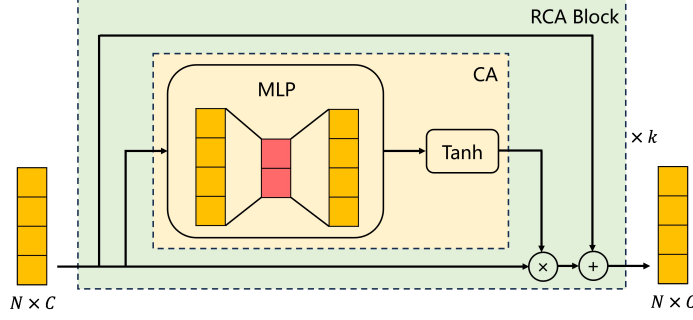


Fig. 7: Structure of the Residual Channel Attention (RCA) module. N represents the batch size, and C represents the feature dimension.

a residual channel-wise scaling structure, as shown in Fig. 7. For a single RCA block, the forward process is formulated as:

$$\begin{aligned} A_{\text{rca}} &= \tanh(\text{MLP}(F_{\text{in}})), \\ F_{\text{out}} &= F_{\text{in}} + F_{\text{in}} \odot A_{\text{rca}} = F_{\text{in}} \odot (\mathbf{1} + A_{\text{rca}}), \end{aligned} \quad (6)$$

where $F_{\text{in}} \in \mathbb{R}^{N \times C}$ is the input global feature, $A_{\text{rca}} \in (-1, 1)^{N \times C}$ is the signed channel attention response, and $\odot$ denotes element-wise multiplication. The MLP follows a bottleneck structure with input dimension C , hidden dimension $C/4$, and output dimension C .

Since A_{rca} is bounded by the tanh activation, the residual scaling factor $\mathbf{1} + A_{\text{rca}}$ lies in $(0, 2)$. Therefore, RCA can adaptively suppress or enhance each channel while avoiding unbounded feature amplification. In retrieval tasks, the final similarity is computed on L2-normalized features; thus, the main effect of RCA is to adjust the feature direction rather than the feature magnitude. This makes RCA suitable for progressive refinement of angularly discriminative embeddings.

In our framework, we stack 16 RCA blocks for global feature refinement. The effectiveness and stability of different numbers of RCA blocks are validated in the ablation study.

3.6 Contrastive Learning with Absolute Cosine Similarity

We adopt contrastive learning inspired by CLIP [13] and tailor it for 3D shape retrieval with an absolute cosine similarity (ACS) metric that provides a sign-invariant geometric target for supervised contrastive retrieval learning.

3.6.1 Absolute Cosine Similarity (ACS)

Given two L2-normalized feature vectors $f_i, f_j \in \mathbb{R}^C$, the absolute cosine similarity is defined as:

$$S_{ij} = \frac{|f_i^\top f_j|}{\|f_i\|_2 \|f_j\|_2} = |\cos \theta_{ij}| \in [0, 1], \quad (7)$$

where θ_{ij} is the angle between the two feature vectors.

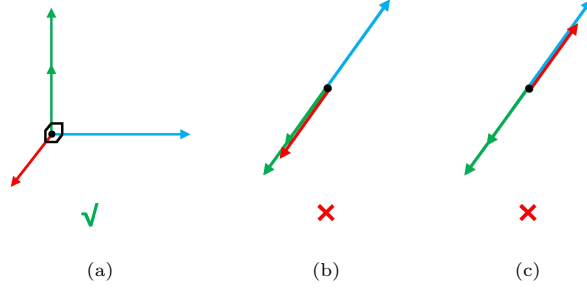


Fig. 8: Geometric illustration of different similarity metrics. Arrows of different colors represent feature vectors of different classes in a 3D Euclidean feature space. (a) Absolute cosine similarity treats opposite directions as equivalent and encourages same-class samples to lie on the same feature line while keeping different classes close to orthogonal. (b)(c) Conventional cosine similarity distinguishes the same and opposite directions, which may impose stricter directional constraints on supervised contrastive retrieval embeddings.

Conventional cosine similarity distinguishes between the same direction and the opposite direction. In contrast, ACS treats two opposite directions as equivalent by measuring the similarity between one-dimensional subspaces rather than oriented vectors. This property is useful for retrieval embeddings because class identity is determined by feature alignment rather than by the sign of the feature vector.

From a geometric perspective, ACS encourages samples from the same class to lie on the same feature line, either in the same or opposite direction, while encouraging samples from different classes to be close to orthogonal. Therefore, each class can be represented by an orthogonal line prototype in the embedding space. This sign-invariant formulation increases the flexibility of class separation in high-dimensional multi-class retrieval scenarios and empirically improves supervised contrastive retrieval learning.

3.6.2 Symmetric Cross-Entropy Loss

The ground truth matrix $GT \in \mathbb{R}^{N \times N}$ is defined by category labels:

$$GT_{ij} = \begin{cases} 1, & \text{sample } i \text{ and } j \text{ belong to the same category,} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

We use a symmetric cross-entropy loss without learnable temperature for supervised contrastive retrieval learning. The loss is formulated as:

$$L_{sce} = \frac{1}{2} [L_{ce}(GT, S) + L_{ce}(GT^T, S^T)] = -\frac{1}{N^2 - N} \sum_{i=1}^N \sum_{j=1}^N GT_{ij} \log \frac{e^{S_{ij}}}{\sum_{k \neq i}^N e^{S_{ik}}} \quad (9)$$

where, $L_{ce}(GT, S)$ and $L_{ce}(GT^T, S^T)$ denote the cross-entropy losses computed along the rows and columns of the matrix, respectively. Since 3D model retrieval is an intra-modal task, both the ground truth matrix GT and the similarity matrix S are symmetric. Therefore, the row-wise and column-wise losses are identical and can be merged into a unified term. Self-pairs are excluded because they trivially yield maximum similarity and do not provide discriminative supervision. This requires the N diagonal elements to be excluded from the mean calculation. Accordingly, the denominator in the final symmetric cross-entropy loss is set to $N^2 - N$ to accurately account for the off-diagonal elements.

3.7 Three-Stage Training Pipeline

To stabilize training and achieve optimal feature representation, we design a three-stage training pipeline with weight freezing (WF) technique, which has been proven effective in image retrieval tasks [26, 27]. The pipeline is illustrated in Fig. 9, and the details of each stage are as follows:

- **Stage I: Single-View Pre-Training.** We remove the multi-view feature aggregation network and connect a linear classification head to the ViT backbone. We train the visual Transformer backbone on a single-view classification task, using all rendered single-view images from the training set. The loss function is standard cross-entropy loss L_{ce} . This stage pre-trains the backbone to capture discriminative 2D visual features of 3D shape views, providing a solid initialization for subsequent multi-view training.
- **Stage II: Frozen Backbone Aggregation Training.** We remove the classification head and reintroduce the multi-view feature aggregation network (MVA, CAA, RCA). We freeze the weights of the pre-trained visual Transformer backbone, and only train the multi-view aggregation network using the proposed symmetric cross-entropy loss L_{sce} with contrastive learning. This stage prevents the untrained aggregation network from corrupting the well-pre-trained backbone weights, enabling stable learning of inter-view fusion strategies.
- **Stage III: Full Network End-to-End Fine-Tuning.** We unfreeze the weights of the ViT backbone, and train the entire MVAT framework end-to-end using the same contrastive learning setup as Stage II. This stage jointly optimizes the single-view feature extraction and multi-view aggregation, aligning the two components for the 3D shape retrieval task, and further boosting performance.

Each training stage runs for 50 epochs, with detailed hyperparameters provided in Section 4.2. Upon completion of the three stages, we obtain a fully trained MVAT framework for 3D shape retrieval.

3.8 Theoretical Analysis

In this section, we analyze the geometric and optimization-related properties of the proposed modules. Instead of claiming algorithmic global convergence, we focus on the geometric feasibility of the proposed absolute cosine similarity, the bounded residual scaling of RCA, and the geometry-aware property of MVA.

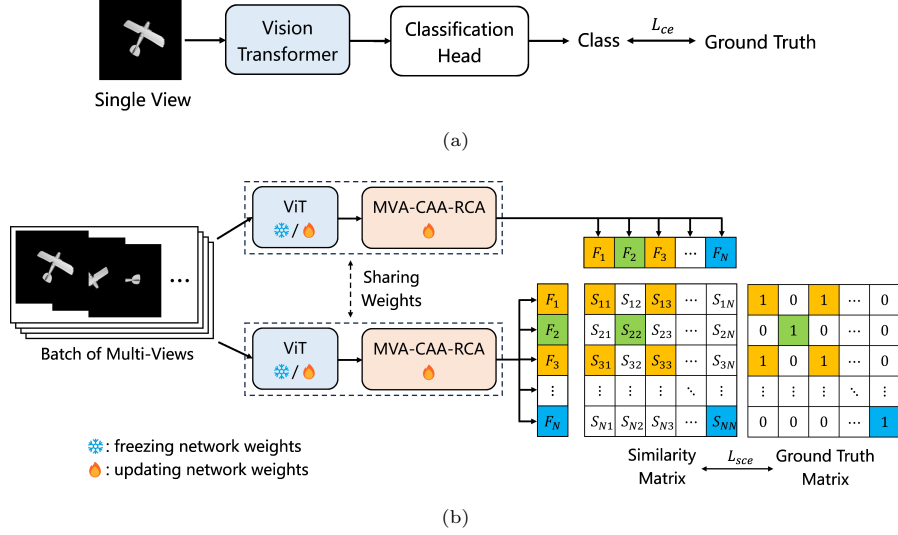


Fig. 9: Three-stage training pipeline for MVAT. (a) Pipeline of Stage I: single-view pre-training with cross-entropy loss L_{ce} . (b) Pipeline of Stage II and III: contrastive training with symmetric cross-entropy loss L_{sce} . f_i represents the global feature of the i -th sample in the batch, N represents the batch size. In the grid on the right, different colors represent different categories.

3.8.1 Geometric Feasibility of Absolute Cosine Similarity

We first analyze why absolute cosine similarity is suitable for high-dimensional multi-class retrieval embeddings.

Definition 1 (Line-Prototype Feature Configuration). For a K -class retrieval problem with feature dimension C , a line-prototype feature configuration assigns each class k to a one-dimensional subspace represented by a unit prototype vector $p_k \in \mathbb{R}^C$. Samples from class k are allowed to lie along either direction of this line, i.e., $f_i \approx p_k$ or $f_i \approx -p_k$.

Proposition 1. For absolute cosine similarity, an ideal line-prototype configuration exists when $K \leq C$.

Proof. Let $\{e_1, e_2, \dots, e_K\}$ be K orthonormal basis vectors in $\mathbb{R}^C$. Assign class k to the line spanned by e_k . For any two samples from the same class, their ideal features are either e_k or $-e_k$, and therefore:

$$|f_i^\top f_j| = |(\pm e_k)^\top (\pm e_k)| = 1. \quad (10)$$

For samples from two different classes k and l , where $k \neq l$, we have:

$$|f_i^\top f_j| = |(\pm e_k)^\top (\pm e_l)| = 0. \quad (11)$$

Thus, ACS admits an ideal feature configuration with maximum intra-class similarity and zero inter-class similarity when $K \leq C$. $\square$

This result shows that ACS provides a feasible geometric target for sign-invariant class separation. Compared with conventional cosine similarity, ACS relaxes the directional constraint by treating f and $-f$ as equivalent. This is particularly useful in supervised contrastive retrieval, where the goal is to retrieve semantically or geometrically similar shapes rather than to preserve the orientation of feature vectors.

It should be noted that Proposition 1 establishes the existence of an ideal geometric configuration, but it does not imply that gradient-based training is guaranteed to converge globally. In practice, the optimization behavior still depends on the network architecture, initialization, training data, and loss landscape. The empirical effectiveness of ACS is validated through the ablation studies in Section 5.

3.8.2 Bounded Residual Scaling of RCA

RCA refines the global feature using residual channel-wise scaling:

$$F_{\text{out}} = F_{\text{in}} \odot (\mathbf{1} + A_{\text{rca}}), \quad (12)$$

where $A_{\text{rca}} = \tanh(\text{MLP}(F_{\text{in}}))$ and $A_{\text{rca}} \in (-1, 1)$.

Therefore, the residual scaling factor satisfies:

$$\mathbf{1} + A_{\text{rca}} \in (0, 2). \quad (13)$$

This bounded scaling prevents uncontrolled channel amplification while still allowing adaptive channel suppression and enhancement. In contrast to direct feature transformation, the residual form preserves the original feature pathway and provides a stable refinement mechanism for stacked RCA blocks.

The complete gradient of RCA includes both the direct residual scaling term and the derivative of the attention branch:

$$\frac{\partial F_{\text{out}}}{\partial F_{\text{in}}} = (\mathbf{1} + A_{\text{rca}}) + F_{\text{in}} \odot \frac{\partial A_{\text{rca}}}{\partial F_{\text{in}}}. \quad (14)$$

The first term provides a bounded direct gradient pathway, while the second term corresponds to the learnable attention branch. This residual design empirically improves training stability and enables progressive feature refinement, as demonstrated by the ablation study on the number of RCA blocks.

3.8.3 Geometry-Aware Property of Hierarchical MVA

The proposed MVA module is designed according to the regular dodecahedron-based camera layout. Global Attention models dependencies among all view-orientation samples, Face Attention models local correlations within overlapping face-level neighborhoods, and Vertex Attention models orientation variations at the same spatial viewpoint.

For rotations that map the regular dodecahedron onto itself, the induced transformation corresponds to a permutation of the view sequence that preserves the vertex-level and face-level neighborhood structures. Since self-attention is permutation-equivariant when the same permutation is applied to the input sequence and the

attention mask, the hierarchical MVA module preserves the geometric neighborhood relations under such symmetry transformations. This provides a geometry-aware inductive bias for robust multi-view aggregation.

In practice, general 3D rotations may not exactly correspond to the finite symmetry group of the dodecahedron. Therefore, the proposed MVA should be understood as providing approximate rotational robustness through symmetric view coverage and geometry-aware attention, rather than exact invariance to arbitrary continuous rotations. This property is empirically evaluated in the rotation robustness experiments.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate MVAT on three mainstream 3D shape retrieval benchmarks: ModelNet40, ModelNet10, and MCB-B, covering both general 3D objects and industrial mechanical components.

4.1.1 ModelNet Datasets

The Princeton ModelNet dataset [9] is one of the most widely used benchmarks for 3D shape retrieval. We use two standard subsets:

- **ModelNet10:** Contains 4,899 CAD models across 10 categories, with 3,991 models for training and 908 for testing.
- **ModelNet40:** Contains 12,311 CAD models across 40 categories, with 9,843 models for training and 2,468 for testing.

For ModelNet datasets, we use the standard mean Average Precision (mAP) as the primary evaluation metric, as defined in [28].

4.1.2 MCB-B Dataset

The Mechanical Components Benchmark (MCB) [29] is a challenging industrial 3D shape retrieval benchmark. We use the MCB-B subset, which contains 18,038 mechanical component CAD models across 25 categories, with 14,451 models for training and 3,587 for testing.

For MCB-B, we evaluate using standard retrieval metrics: F1-score@N (F1@N), mAP, and Normalized Discounted Cumulative Gain (NDCG@N). We report both micro and macro versions of all metrics, as well as their average (micro + macro) for overall performance evaluation.

4.2 Implementation Details

Backbone and Input

Our framework uses MaxViT-Tiny [20] pre-trained on ImageNet-1K [30] as the single-view feature extractor, producing 512-dimensional features for each view. The input

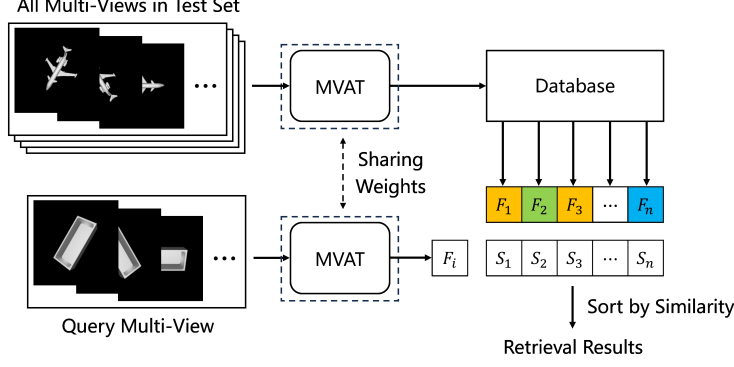


Fig. 10: Inference pipeline for MVAT. f_i represents the global feature of the i -th sample in the test set, n represents the size of the test set.

view images are resized to 224×224 with a pure black background. We stack 16 RCA blocks for global feature refinement.

View Rendering

We use Blender 3.6 LTS for offline view rendering. All 3D models are normalized to the unit sphere, with orthographic projection, camera field of view (FOV) of 60° , ambient light intensity of 0.8, no shadow casting, and pure black background. The 60 views are rendered following the regular dodecahedron camera layout described in Section 3.2.

Training Hyperparameters

We use the AdamW optimizer [31] with weight decay of 10^{-4} for all training stages. The learning rates are set to 10^{-4} for Stage I, and 10^{-5} for Stage II and III. Each training stage runs for 50 epochs, with a batch size of 128 for Stage I and 64 for Stage II and III. We use a cosine annealing learning rate scheduler with 5 epochs of warm-up for all stages. Data augmentation for Stage I includes random horizontal flip, random resized crop to 224×224 , and mild color jitter; for Stage II and III, only random horizontal flip is applied to preserve view geometric consistency.

Inference Pipeline

During testing, we use the trained MVAT to extract global features for all 3D models in the test set, and build a feature database. For each query model, we compute the absolute cosine similarity between the query feature and all features in the database, and rank the models in descending order of similarity to obtain the retrieval results. The inference pipeline is illustrated in Fig. 10.

4.3 Performance Comparison on ModelNet Datasets

We compare the performance of MVAT with state-of-the-art methods on ModelNet40 and ModelNet10 datasets, including traditional hand-crafted methods, voxel-based

methods, point-based methods, view-based methods, and multi-modal methods. The results are shown in Table 1 and Table 2.

On ModelNet40, MVAT achieves an mAP of 93.2%, surpassing the previous compared methods MVTN [10] by 0.3%, and outperforming the classic MVCNN [4] by 13.0%. On ModelNet10, MVAT achieves an mAP of 95.1%, surpassing the previous compared method SPNet [32] by 0.9%, and outperforming the attention-based VAM-IAM [14] by 1.5%. MVAT achieves competitive retrieval performance on ModelNet40 and ModelNet10, and shows a clear advantage on the mechanical component benchmark.

We also visually present the top-10 retrieval results of MVAT for randomly selected test cases from ModelNet40 in Fig. 11. The performance shows that MVAT produces accurate and stable retrieval results, and even in rare error cases, the incorrectly retrieved models still share high geometric and semantic similarity with the query model.

Table 1: Performance Comparison with SOTA Methods on ModelNet40

Method	Data Representation	mAP
SPH [1]	Hand-Crafted Descriptor	33.3%
LFD [2]	Hand-Crafted Descriptor	40.9%
3DShapeNets [9]	Voxel Grid	49.2%
DeepPano [33]	Panoramic View	76.8%
MVCNN [4]	Multi-View	80.2%
GIFT [34]	Multi-View	81.9%
PANORAMA-NN [35]	Panoramic View	83.5%
SPNet [32]	Stereographic Projection	85.2%
GVCNN [5]	Multi-View	85.7%
PANORAMA-ENN [36]	Panoramic View	86.3%
TCL [21]	Multi-View	88.0%
PT [37]	Point Cloud	88.4%
VIP-GAN [38]	Multi-View	89.2%
VNN [39]	Multi-View	89.3%
PVNet [40]	Point Cloud + Multi-View	89.5%
DSHCNet [11]	Point Cloud + Mesh	91.5%
MLVCNN [41]	Multi-View	92.8%
VAM-IAM [14]	Multi-View	92.8%
MVTN [10]	Point Cloud + Multi-View	92.9%
MVAT (ours)	Multi-View	93.2%

4.4 Performance Comparison on MCB-B Dataset

We compare the performance of MVAT with SOTA methods on the challenging MCB-B mechanical component retrieval benchmark. The results are shown in Table ??.

MVAT achieves an average F1@N of 88.6%, mAP of 95.4%, and NDCG@N of 96.0%, outperforming the strongest compared method by 12.2% in F1@N, 4.3% in

Table 2: Performance Comparison with SOTA Methods on ModelNet10

Method	Data Representation	mAP
SPH [1]	Hand-Crafted Descriptor	44.1%
LFD [2]	Hand-Crafted Descriptor	49.8%
3DShapeNets [9]	Voxel Grid	68.3%
DeepPano [33]	Panoramic View	84.2%
PANORAMA-NN [35]	Panoramic View	87.4%
VIP-GAN [38]	Multi-View	90.7%
GIFT [34]	Multi-View	91.1%
VNN [39]	Multi-View	92.8%
DSHCNet [11]	Point Cloud + Mesh	93.1%
PANORAMA-ENN [36]	Panoramic View	93.3%
VAM-IAM [14]	Multi-View	93.6%
SPNet [32]	Stereographic Projection	94.2%
MVAT (ours)	Multi-View	95.1%

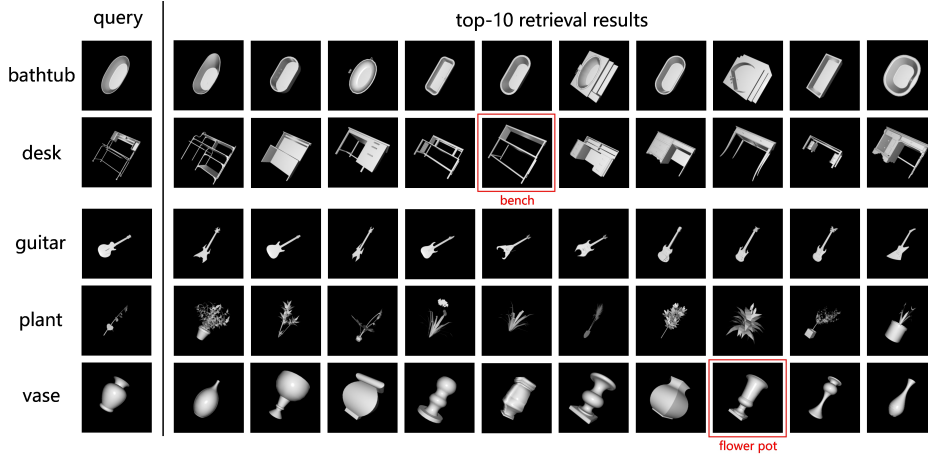


Fig. 11: Top-10 retrieval results of MVAT for test cases on the ModelNet40 dataset. In each row, the leftmost 3D model is the query, with its category label on the left; the 10 models on the right are the top-10 retrieval results. Models enclosed in highlighted frames are incorrect retrievals, with their category labels below in red.

mAP, and 6.2% in NDCG@N. This significant performance improvement demonstrates that MVAT is highly effective for retrieving standardized industrial mechanical components, indicating its potential for industrial component retrieval.

We also present the top-10 retrieval results of MVAT for MCB-B test cases in Fig. 12. The results show that MVAT achieves excellent retrieval performance even for fine-grained mechanical components with high geometric similarity.

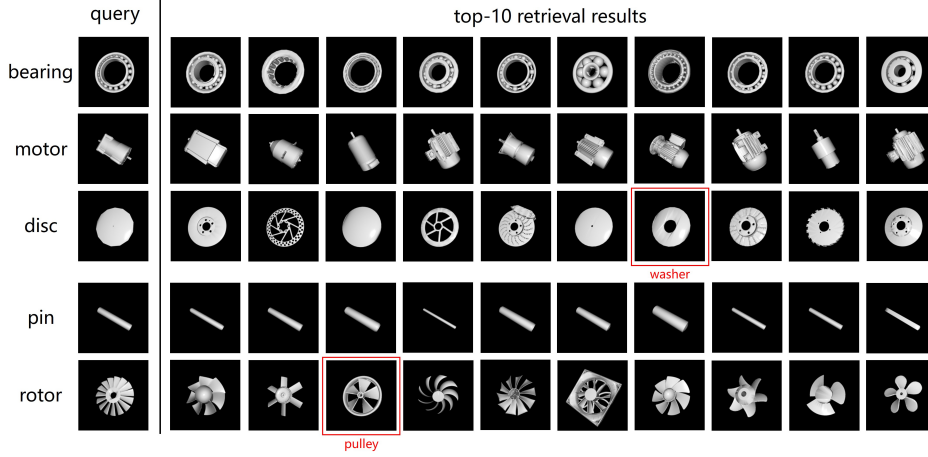


Fig. 12: Top-10 retrieval results of MVAT for test cases on the MCB-B dataset. In each row, the leftmost 3D model is the query, with its category label on the left; the 10 models on the right are the top-10 retrieval results. Models enclosed in red frames are incorrect retrievals, with their category labels below in red.

4.5 Rotation Robustness Evaluation

To verify the rotational robustness of MVAT, we conduct experiments on ModelNet40 with randomly rotated test models. We rotate each test model around a random 3D axis by $\pm 45^\circ$, $\pm 90^\circ$, $\pm 135^\circ$, and $\pm 180^\circ$ respectively, and test the MVAT trained on the original non-rotated training set. The results are shown in Table 4.

The mAP of MVAT decreases by no more than 1.2 percentage points across all rotation angles, indicating that the symmetric camera layout and hierarchical MVA module provide strong rotational robustness, even for unseen rotation angles. This is critical for practical 3D retrieval applications where 3D models may have arbitrary poses.

Table 4: Rotation Robustness Performance on ModelNet40

Rotation Angle	0°	$\pm 45^\circ$	$\pm 90^\circ$	$\pm 135^\circ$	$\pm 180^\circ$
mAP	93.2%	92.2%	92.7%	92.0%	92.8%

5 Ablation Study

We conduct comprehensive ablation studies on the ModelNet40 dataset to validate the effectiveness of each core component of MVAT. All ablation experiments use the same training and evaluation setup as the full model.

5.1 Ablation of Channel Attention Aggregation (CAA)

We first validate the effectiveness of the proposed CAA module by comparing it with traditional aggregation methods. We use a baseline model with MaxViT backbone and no MVA/RCA modules, and compare different aggregation strategies. The results are shown in Table 5.

CAA outperforms max-pooling by 3.7% and average-pooling by 0.9% in mAP, demonstrating its superiority in multi-view feature aggregation. This improvement comes from the adaptive channel attention that integrates both max and average pooling statistics, and the tanh activation that expands features to the full space for better similarity discriminability.

Table 5: Ablation of Aggregation Methods on ModelNet40

MVA	Aggregation Method	RCA	mAP
×	Max-Pooling	×	86.0%
×	Average-Pooling	×	88.8%
×	CAA (ours)	×	89.7%

5.2 Ablation of Residual Channel Attention (RCA)

We validate the effectiveness of the RCA module by varying the number of stacked RCA blocks. The baseline model uses MaxViT backbone + CAA, with no MVA module. The results are shown in Table 6.

Incorporating RCA significantly boosts retrieval performance, with 16 RCA blocks achieving the best performance, outperforming the baseline by 2.2% mAP. Stacking more than 16 blocks leads to slight performance degradation due to overfitting. These results are consistent with the bounded residual scaling analysis of RCA: the residual pathway preserves the original feature representation, while the signed channel-wise modulation progressively adjusts feature directions for improved angular discriminability in supervised contrastive retrieval.

5.3 Ablation of Multi-View Attention (MVA)

We validate the effectiveness of the hierarchical MVA module by ablating its components and their cascading order. The baseline model uses MaxViT backbone + CAA + 16 RCA blocks. The results are shown in Table 7.

Incorporating any single attention component improves performance by at least 0.8% mAP, with Vertex Attention (VA) achieving the highest single-component performance of 93.0% mAP. Among all cascading orders, the GA $\rightarrow$ FA $\rightarrow$ VA order achieves the best performance of 93.2% mAP, which follows a global-to-local hierarchical modeling process that fully aligns with the geometric symmetry of the multi-view layout. These results validate the effectiveness of the proposed hierarchical MVA module.

Table 6: Ablation of the Number of RCA Blocks on ModelNet40

MVA	CAA	Number of RCA Blocks	mAP
×	✓	0	89.7%
×	✓	1	90.0%
×	✓	2	89.7%
×	✓	4	90.2%
×	✓	8	91.0%
×	✓	16	91.9%
×	✓	32	90.7%

Table 7: Ablation of MVA Module Structures on ModelNet40

MVA Configuration	CAA	16 RCA Blocks	mAP
None	✓	✓	91.9%
GA only	✓	✓	92.7%
FA only	✓	✓	92.8%
VA only	✓	✓	93.0%
GA → FA	✓	✓	88.7%
GA → VA	✓	✓	89.5%
FA → GA	✓	✓	91.4%
FA → VA	✓	✓	90.7%
VA → GA	✓	✓	89.9%
VA → FA	✓	✓	91.5%
GA → FA → VA (ours)	✓	✓	93.2%
GA → VA → FA	✓	✓	92.1%
FA → GA → VA	✓	✓	92.9%
FA → VA → GA	✓	✓	92.8%
VA → GA → FA	✓	✓	92.6%
VA → FA → GA	✓	✓	92.6%

It is also observed that partial cascades of two attention components do not always improve performance, suggesting that incomplete hierarchical modeling may introduce inconsistent dependency patterns. The full $GA \rightarrow FA \rightarrow VA$ cascade provides the most coherent global-to-local aggregation process.

5.4 Ablation of Contrastive Learning and Loss Function

We validate the effectiveness of the proposed contrastive learning framework by comparing different loss functions. The results are shown in Table 8.

Using symmetric cross-entropy loss for contrastive learning improves performance by 2.0% mAP compared to standard cross-entropy loss without contrastive learning. It

also outperforms other contrastive loss functions, including contrastive loss and MSE loss, by 0.9% and 0.8% mAP respectively. We further visualize the feature distribution via t-SNE [47] in Fig. 13, which shows that symmetric cross-entropy loss achieves the best intra-class compactness and inter-class separability.

Table 8: Ablation of Loss Functions on ModelNet40

Loss Function	Contrastive Learning	mAP
Cross-Entropy Loss	×	91.2%
Contrastive Loss	✓	92.3%
Mean Square Error Loss	✓	92.4%
Symmetric Cross-Entropy Loss (ours)	✓	93.2%

5.5 Ablation of Absolute Cosine Similarity

We validate the effectiveness of the proposed absolute cosine similarity (ACS) by comparing it with conventional cosine similarity. The results are shown in Table 9. ACS outperforms conventional cosine similarity by 1.2% mAP. The t-SNE visualization in Fig. 14 shows that ACS leads to a radiating feature distribution with clear inter-class separation, while conventional cosine similarity leads to scattered and overlapping clusters. This validates that ACS improves feature discriminability by providing a sign-invariant geometric target for supervised contrastive retrieval learning.

Table 9: Ablation of Similarity Functions on ModelNet40

Similarity Function	mAP
Cosine Similarity	92.0%
Absolute Cosine Similarity (ours)	93.2%

5.6 Ablation of Three-Stage Training Pipeline

We validate the effectiveness of the proposed three-stage training pipeline by comparing it with simplified training strategies. The total number of training epochs is fixed at 150 for all strategies. The results are shown in Table 10.

The one-stage training method performs the worst, without single-view pre-training for discriminative feature initialization. The two-stage method without weight freezing improves the mAP to 91.4%, bringing a 4.3 percentage-point gain over the one-stage baseline. With weight freezing, the two-stage method further improves the mAP to 92.5%, suggesting that preserving the pre-trained backbone during aggregation training benefits multi-view feature learning. The proposed three-stage

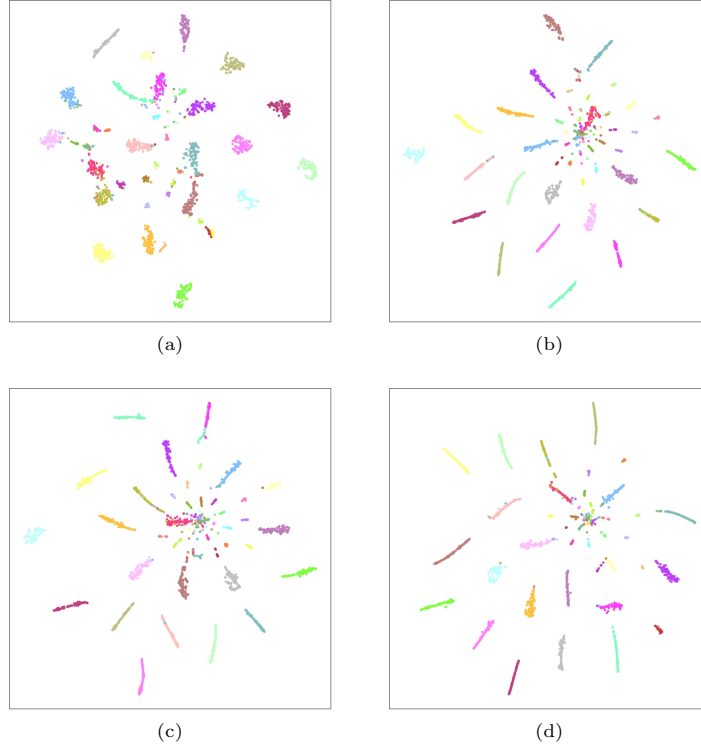


Fig. 13: t-SNE visualization of feature distributions on ModelNet40 test set. Points of different colors represent features of different categories. Features are extracted by MVAT with different loss functions: (a) cross-entropy loss, (b) contrastive loss, (c) MSE loss, (d) symmetric cross-entropy loss (ours).

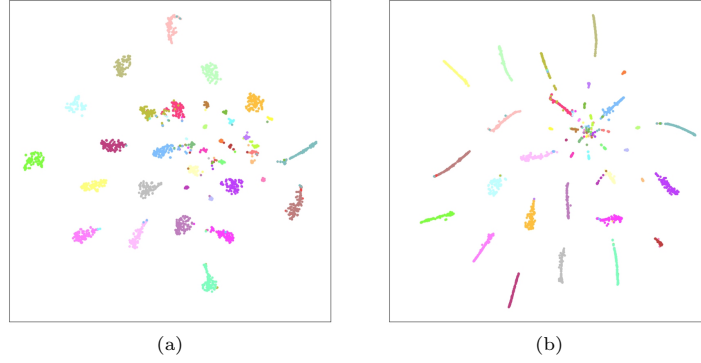


Fig. 14: t-SNE visualization of feature distributions with different similarity functions. (a) Conventional cosine similarity. (b) Absolute cosine similarity (ours).

method achieves the best performance of 93.2% mAP after end-to-end fine-tuning, with an additional improvement of 0.7 percentage points. These results validate the effectiveness of the three-stage training pipeline.

Table 10: Ablation of Training Methods on ModelNet40

Training Method	Stage I Epochs	Stage II Epochs	Stage III Epochs	mAP
One-stage	×	×	150	87.1%
Two-stage w/o Weight Freezing	75	×	75	91.4%
Two-stage w/ Weight Freezing	75	75	×	92.5%
Three-stage (ours)	50	50	50	93.2%

6 Discussion

In this section, we discuss the core advantages, limitations, and broader impacts of the proposed MVAT framework, providing a comprehensive perspective on our work.

6.1 Core Advantages of MVAT

MVAT addresses several limitations of existing view-based 3D shape retrieval methods, with three main advantages: (1) Geometry-aligned hierarchical attention. Unlike generic attention modules that ignore the structure of the camera layout, MVA explicitly incorporates the global, face-level, and vertex-level relationships induced by the regular dodecahedron-based acquisition scheme. This provides a useful geometric inductive bias for multi-view aggregation. (2) Stable and discriminative feature refinement. CAA and RCA perform signed channel-wise modulation and residual feature refinement, which improve the discriminability of angular retrieval embeddings. The bounded residual scaling in RCA helps avoid uncontrolled feature amplification during deep stacking. (3) Effective supervised contrastive retrieval learning. The proposed ACS metric relaxes the directional constraint of conventional cosine similarity by treating opposite feature directions as equivalent. This sign-invariant design admits an orthogonal line-prototype configuration for class separation and empirically improves retrieval performance.

6.2 Limitations and Future Work

Despite the advantages of MVAT, there are still some limitations to be addressed in future work: (1) Real-World Scanned Data Adaptation. While MVAT performs well on clean CAD models, its performance on real-world scanned 3D data with noise, occlusion, and clutter can be further improved. Future work will explore domain adaptation techniques and geometry-aware attention for noisy real-world data. (2) Fine-Grained Geometric Modeling. The current framework focuses on global shape retrieval, and

its ability to capture fine-grained geometric differences (e.g., small mechanical features) can be enhanced. Future work will integrate local geometric descriptors into the Transformer framework for fine-grained retrieval. (3) Multi-Modal Fusion. Future work will extend MVAT to a unified Transformer framework that fuses multi-view images, point clouds, and mesh data, leveraging complementary information from different 3D representations for better retrieval performance.

The current framework uses 60 rendered views, which improves coverage but increases computational cost. Future work will investigate adaptive view selection and lightweight aggregation.

6.3 Broader Impacts

The proposed MVAT framework provides a useful direction for Transformer-based multi-view 3D shape understanding. The hierarchical attention design aligned with geometric symmetry, the sign-invariant similarity metric with a feasible geometric interpretation, and the stable training pipeline can be extended to other 3D vision tasks, including 3D shape classification, part segmentation, pose estimation, and 3D generation. Our work also has practical value for industrial applications such as computer-aided design, digital twin, and intelligent manufacturing, where efficient and accurate 3D shape retrieval is critical.

7 Conclusion

This paper presents MVAT, a geometry-aligned Transformer-based multi-view aggregation framework for 3D shape retrieval. MVAT introduces a hierarchical Multi-View Attention module that models global, face-level, and vertex-level view correlations according to a regular dodecahedron-based camera layout. In addition, Channel Attention Aggregation and Residual Channel Attention are designed to improve multi-view feature fusion and progressive feature refinement.

We further introduce absolute cosine similarity for supervised contrastive retrieval learning. The proposed similarity metric treats opposite feature directions as equivalent and admits an orthogonal line-prototype configuration for sign-invariant class separation. This provides a feasible geometric target for high-dimensional multi-class retrieval embeddings and improves contrastive feature learning in practice.

Extensive experiments on ModelNet40, ModelNet10, and MCB-B show that MVAT achieves strong retrieval performance on both general 3D object datasets and industrial mechanical component benchmarks. Comprehensive ablation studies validate the effectiveness of the proposed hierarchical attention design, channel attention modules, absolute cosine similarity, and three-stage training pipeline. Future work will further investigate efficient view selection, stronger adaptation to noisy real-world scanned data, and multi-modal fusion with point cloud and mesh representations.

Author Contributions. Ze Zhang: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing – Original Draft. Xiaojun Wu: Supervision, Funding acquisition, Writing – Review & Editing, Project Administration. Chenxi Wu: Software, Validation, Data curation, Writing – Review & Editing. Guoyuan Liang: Validation, Writing – Review & Editing, Resources.

Declaration of competing interest. The authors declare no conflict of interest.

Data availability. Data supporting the results of this study are available from the corresponding author upon reasonable request.

Acknowledgments. This work was supported in part by the Shenzhen Science and Technology Plan Key Project (JCYJ20241202123501003), Shenzhen Science and Technology Plan Project (JCYJ20250604145547061, SYSPG20241211173609005).

References

- [1] Kazhdan, M., Funkhouser, T. & Rusinkiewicz, S. *Rotation invariant spherical harmonic representation of 3 d shape descriptors*, Vol. 6, 156–164 (2003).
- [2] Chen, D.-Y., Tian, X.-P., Shen, Y.-T. & Ouhyoung, M. *On visual similarity based 3d model retrieval*, Vol. 22, 223–232 (Wiley Online Library, 2003).
- [3] Papadakis, P., Pratikakis, I., Theoharis, T. & Perantonis, S. Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. *International Journal of Computer Vision* **89**, 177–192 (2010).
- [4] Su, H., Maji, S., Kalogerakis, E. & Learned-Miller, E. *Multi-view convolutional neural networks for 3d shape recognition*, 945–953 (2015).
- [5] Feng, Y., Zhang, Z., Zhao, X., Ji, R. & Gao, Y. *Gvcnn: Group-view convolutional neural networks for 3d shape recognition*, 264–272 (2018).
- [6] Wei, X., Yu, R. & Sun, J. *View-gcn: View-based graph convolutional network for 3d shape analysis*, 1850–1859 (2020).
- [7] Qi, C. R., Su, H., Mo, K. & Guibas, L. J. *Pointnet: Deep learning on point sets for 3d classification and segmentation*, 652–660 (2017).
- [8] Qi, C. R., Yi, L., Su, H. & Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* **30** (2017).
- [9] Wu, Z. *et al.* *3d shapenets: A deep representation for volumetric shapes*, 1912–1920 (2015).
- [10] Hamdi, A., Giancola, S. & Ghanem, B. *Mvtn: Multi-view transformation network for 3d shape recognition*, 1–11 (2021).
- [11] Huang, X., Nong, L. & Zhang, W. *A multimodal fusion network based on hypergraph for 3d shape retrieval*, 1682–1687 (IEEE, 2022).
- [12] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

- [13] Radford, A. *et al.* *Learning transferable visual models from natural language supervision*, 8748–8763 (PMLR, 2021).
- [14] Lin, D. *et al.* Multi-view 3d object retrieval leveraging the aggregation of view and instance attentive features. *Knowledge-Based Systems* **247**, 108754 (2022).
- [15] Hu, J., Shen, L. & Sun, G. *Squeeze-and-excitation networks*, 7132–7141 (2018).
- [16] Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. *Cbam: Convolutional block attention module*, 3–19 (2018).
- [17] Song, C. H., Han, H. J. & Avrithis, Y. *All the attention you need: Global-local, spatial-channel attention for image retrieval*, 2754–2763 (2022).
- [18] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
- [19] Liu, Z. *et al.* *Swin transformer: Hierarchical vision transformer using shifted windows*, 10012–10022 (2021).
- [20] Tu, Z. *et al.* *Maxvit: Multi-axis vision transformer*, 459–479 (Springer, 2022).
- [21] He, X., Zhou, Y., Zhou, Z., Bai, S. & Bai, X. *Triplet-center loss for multi-view 3d object retrieval*, 1945–1954 (2018).
- [22] Li, Z., Xu, C. & Leng, B. *Angular triplet-center loss for multi-view 3d shape retrieval*, Vol. 33, 8682–8689 (2019).
- [23] He, X., Bai, S., Chu, J. & Bai, X. An improved multi-view convolutional neural network for 3d object retrieval. *IEEE Transactions on Image Processing* **29**, 7917–7930 (2020).
- [24] Khosla, P. *et al.* *Supervised contrastive learning*, Vol. 33, 18661–18673 (2020).
- [25] Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [26] Wu, H., Wang, M., Zhou, W., Lu, Z. & Li, H. *Asymmetric feature fusion for image retrieval*, 11082–11092 (2023).
- [27] Wang, S. *et al.* *Fine-grained retrieval prompt tuning*, Vol. 37, 2644–2652 (2023).
- [28] Shilane, P., Min, P., Kazhdan, M. & Funkhouser, T. *The princeton shape benchmark*, 167–178 (IEEE, 2004).
- [29] Kim, S., Chi, H.-g., Hu, X., Huang, Q. & Ramani, K. *A large-scale annotated mechanical components benchmark for classification and retrieval tasks with deep neural networks*, 175–191 (Springer, 2020).

- [30] Deng, J. *et al.* *Imagenet: A large-scale hierarchical image database*, 248–255 (Ieee, 2009).
- [31] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [32] Yavartanoo, M., Kim, E. Y. & Lee, K. M. *Spnet: Deep 3d object classification and retrieval using stereographic projection*, 691–706 (Springer, 2018).
- [33] Shi, B., Bai, S., Zhou, Z. & Bai, X. Deeppano: Deep panoramic representation for 3-d shape recognition. *IEEE Signal Processing Letters* **22**, 2339–2343 (2015).
- [34] Bai, S., Bai, X., Zhou, Z., Zhang, Z. & Jan Latecki, L. *Gift: A real-time and scalable 3d shape search engine*, 5023–5032 (2016).
- [35] Sfikas, K., Theoharis, T. & Pratikakis, I. Exploiting the panorama representation for convolutional neural network classification and retrieval. *3DOR@Eurographics* **7** (2017).
- [36] Sfikas, K., Pratikakis, I. & Theoharis, T. Ensemble of panorama-based convolutional neural networks for 3d model classification and retrieval. *Computers & Graphics* **71**, 208–218 (2018).
- [37] Shajahan, D. A., Varma, M. & Muthuganapathy, R. Point transformer for shape classification and retrieval of urban roof point clouds. *IEEE Geoscience and Remote Sensing Letters* **19**, 1–5 (2021).
- [38] Han, Z., Shang, M., Liu, Y.-S. & Zwicker, M. *View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions*, Vol. 33, 8376–8384 (2019).
- [39] He, X., Huang, T., Bai, S. & Bai, X. *View n-gram network for 3d object retrieval*, 7515–7524 (2019).
- [40] You, H., Feng, Y., Ji, R. & Gao, Y. *Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition*, 1310–1318 (2018).
- [41] Jiang, J., Bao, D., Chen, Z., Zhao, X. & Gao, Y. *Mhvcnn: Multi-loop-view convolutional neural network for 3d shape retrieval*, Vol. 33, 8513–8520 (2019).
- [42] Brock, A., Lim, T., Ritchie, J. M. & Weston, N. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236* (2016).
- [43] Kanezaki, A., Matsushita, Y. & Nishida, Y. *Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints*, 5010–5019 (2018).

- [44] Xu, Y., Fan, T., Xu, M., Zeng, L. & Qiao, Y. *Spidercnn: Deep learning on point sets with parameterized convolutional filters*, 87–102 (2018).
- [45] Furuya, T. & Ohbuchi, R. *Diffusion-on-manifold aggregation of local features for shape-based 3d model retrieval*, 171–178 (2015).
- [46] Li, Y. *et al.* Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems* **31** (2018).
- [47] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).