ROSA-Net: Rotation-Robust Structure-Aware Network for Fine-grained 3D Shape Retrieval

Rao Fu[1,*], Yunchi Zhang[1,*], Jie Yang[1,2,†], Jiawei Sun[3], Fanglue Zhang[4], Yu-Kun Lai[5], and Lin Gao[1,2]

[1] University of Chinese Academy of Sciences
[2] Institute of Computing Technology, Chinese Academy of Sciences
[3] Beijing Jiaotong University
[4] Victoria University of Wellington
[5] Cardiff University
*Equal Contribution

Abstract. Fine-grained 3D shape retrieval aims to retrieve 3D shapes similar to a query shape in a repository with models belonging to the same class, which requires shape descriptors to represent detailed geometric information to discriminate shapes with globally similar structures. Moreover, 3D objects can be placed with arbitrary positions, orientations, and scales in real-world applications, which further requires shape descriptors to be robust to rotation and sensitive to scale. The shape descriptions used in existing 3D shape retrieval systems fail to meet the above two criteria. In this paper, we introduce a novel deep architecture, ROSA-Net, which learns rotation-robust and scale-sensitive 3D shape descriptors capable of encoding fine-grained geometric information and structural information, and thus achieve accurate results on the task of fine-grained 3D object retrieval. ROSA-Net extracts a set of compact and detailed geometric features part-wisely and discriminatively estimates the contribution of each semantic part to shape representation. Furthermore, our method can learn the importance of geometric and structural information of all the parts when generating the final compact latent feature of a 3D shape for fine-grained retrieval. We also build and publish a new 3D shape dataset with sub-class labels for validating the performance of fine-grained 3D shape retrieval methods. Qualitative and quantitative experiments show that our ROSA-Net outperforms state-of-the-art methods on the fine-grained object retrieval task, demonstrating its capability in geometric detail extraction. The code is available in the supplementary material.

1 Introduction

Recent advancements in modeling, digitizing, and visualizing physical and virtual 3D objects have resulted in an explosion of available 3D models on the internet.

† Corresponding author: yangjie01@ict.ac.cn

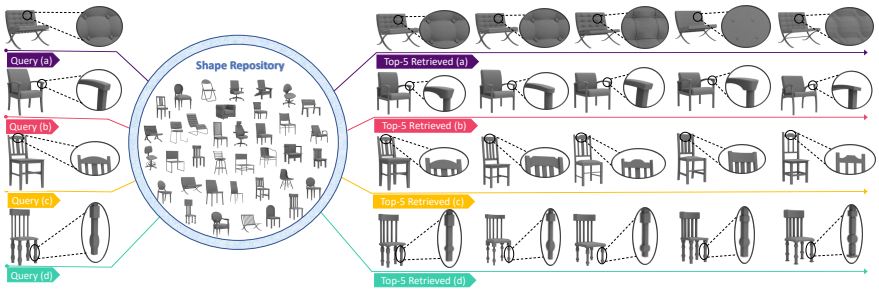


Fig. 1: Four examples of the top-5 retrieval results given query models on the shape dataset with perturbed rotations within the chair category. Our method is able to capture geometric details, learn the importance of each part, and balance the contribution of structure and geometric information in fine-grained retrieval.

As a result, effective retrieval of models from a shape repository has become an integral part of the research field of 3D shape analysis. The most popular shape retrieval methods are content-based approaches that use shape descriptors to search for similar models. However, while many methods exist to retrieve similar 3D shapes using their shape descriptors, most of these methods perform well only on inter-class retrieval, where shape search engines are tasked with retrieving shapes with the same class label among different object classes that typically have vastly different overall shapes. This often results in the retrieval of an object of the same class that does not look similar to the query due to mismatched structure or geometric details.

In contrast, fine-grained intra-class 3D shape retrieval is frequently overlooked but is essential for real-world applications. For example, in deformation-based 3D modeling [52], designers often create a new shape by deforming a similar source shape within the same class. In robot grasping [53], retrieving an object with similar fine-grained details aids in grasping novel objects. Additionally, in online shopping, users search for the ideal object within a database containing objects of the same class. In this paper, we aim to solve the task of fine-grained intra-class 3D shape retrieval, which is of critical importance in various practical scenarios.

Fine-grained intra-class 3D shape retrieval presents greater challenges compared to inter-class retrieval. One challenge of fine-grained retrieval lies in the ability of shape descriptors to distinguish between subtle differences in the geometry and structure of objects within the same class. This requires that shape descriptors capture more detailed geometric features of the shapes, such as local surface features of the chair legs. Additionally, structure-aware information must be incorporated into the shape descriptors to differentiate between objects with different underlying structures, such as chairs with different leg arrangements. Another challenge in 3D shape retrieval is that objects in a repository are typically not in a canonical form. For instance, in real-world data, objects may be placed with arbitrary positions and orientations, while in synthesized data,

it is labor-intensive and infeasible for designers to strictly canonicalize all of the designed objects. Furthermore, fine-grained intra-class retrieval requires the incorporation of scale information into the shape descriptors. For instance, a coffee table and a dining table may look similar in their geometric details and structure but differ in scale, and thus scale information is necessary for accurate retrieval. This is especially important in applications such as robot grasping, where retrieving an object with similar fine-grained details aids in grasping novel objects of varying sizes.

To tackle the above challenges, we propose a Rotation-Robust Scale-Sensitive Structure-Aware Network (ROSA-Net) for fine-grained 3D shape retrieval which extracts a 3D shape descriptor that is robust to rotation, sensitive to scale, and is also informative to indicate fine-grained shape similarity and structure information. Our method is based on the following three observations. Firstly, in both physical object manufacturing and digital 3D shape modeling, assembling interchangeable components or parts has become a practical reality to avoid considerable, and often inconvenient user interactions in 3D object design [13, 45]. It is thus natural to see part-wise differences in 3D models constructed for different functionalities. The parts of an object often vary in importance, and therefore contribute unequally to discriminating different object categories. Therefore, we extract geometric information part-wisely and then use an attention mechanism(*Part-Geo Attention*) to learn to weight the contributions of different parts to shape retrieval. Secondly, we need to ensure that the extracted descriptor can well encode the subtle difference among the same semantic part of different objects, and the descriptor should be rotation-robust and scale-sensitive. Thus, we propose a variational autoencoder(*PartVAE*) to encode reconstructive rotation-invariant and scale-sensitive mesh-based geometric information: edge lengths and dihedral angles. Intuitively, a latent feature that can be precisely reconstructed to the whole shape must have comprised all detailed geometric information of the original shape. Thirdly, there are shapes that differ in local details and shapes that differ in structure, so the descriptor should distinguish which information is more critical in the differentiation of shapes. Therefore, we propose a new paradigm(*Geo-Struct Attention*) to balance the importance of fine-grained geometric information and structural information and incorporate this information into the shape descriptors through variational autoencoder(*GlobalVAE*). With the above components, experiments show our network is able to learn 3D shape descriptors that achieve higher accuracy than previous methods when querying shapes with arbitrary poses against objects of the same class. The *Part-Geo Attention* and the *Geo-Struct Attention* are able to balance the part-wise geometry and structure information for intra-class shape retrieval.

In summary, our main contributions are as follows:

1. We propose a part-based deep model, ROSA-Netfor intra-class fine-grained 3D shape retrieval. The architecture includes a compact mesh-based descriptor that encodes rotation-robust and scale-sensitive geometric information.

2. We propose a self-supervised paradigm that balances structural and geometric information in shape discrimination.
3. We build and publish a 3D Shape dataset, ROSA-Net-Dataset, to evaluate fine-grained intra-class object retrieval methods quantitatively, which provides sub-class labels of all the 8,906 3D shapes in 6 classes.

2 Related Work

The task of 3D shape retrieval is critical in evaluating the descriptive capabilities of shape representation models. In this section, we commence by conducting a thorough review of the present research on 3D shape retrieval. Subsequently, we examine two contemporary approaches for 3D shape representation, namely mesh-based representations and rotation-robust representations. These methods are of particular relevance to our study.

2.1 3D Shape Retrieval

Shape retrieval is an indispensable aspect of numerous applications that rely on large-scale 3D shape repositories, such as shape modeling [59], template-based deformation [52], and scene modeling [60]. These applications retrieve a globally or locally similar shape from the target shape and employ it as a component or template for shape modeling, utilizing the information provided by the shape repository. Effective description of 3D shapes constitutes a crucial component of shape retrieval. Hand-crafted shape descriptors such as lightfield descriptors [6] and spherical harmonic descriptors [26] have been employed to extract global 3D features. On the other hand, local information for partial shape retrieval can be described using shape distribution [36], heat kernel diffusion [50], predefined primitives [38], bag-of-features [4, 35, 41], and shape editing distance [27]. These methods can also aggregate local information for global shape retrieval. Recently, facilitated by progress in deep neural networks, machine learning-based methods have been adopted to improve the descriptive power of 3D shape representations. Multi-view image-based approaches aim to aggregate features from multi-view images for shape representation, which aggregates image features through pooling operations [49], image matching [1, 25], or attention-based sequential view aggregation [20, 21]. Instead of projecting shapes to multi-view images, Shi et al. [46] and Steve et al. [14] proposed to project a shape to a cylinder and a unit sphere respectively, and learn features from their projection without extra aggregation operations. Other methods focused on extracting features from point or mesh based representations. [17] extracted and aggregated local features from rotation-normalized point sets. [62] transformed the point set into a volumetric representation and introduced a voxel feature encoding layer for feature extraction. [52] developed a deep embedding technique to retrieve a 3D model that can best match the query through mesh deformation.

While these methods perform well on large-scale 3D shape retrieval benchmarks [28, 42], their representative power is limited to distinguishing shapes of

different sub-classes within the same overall class, overlooking fine-grained shape features. Additionally, they focused on only geometric information or spatial information, without discerning the global semantic structure of objects. Despite some non-learning-based methods [27] utilizing both structural and geometric information, they necessitate users to manually determine the significance of these features. In contrast, ROSA-Net learns the importance through a self-supervised approach. [30]introduced a fine-grained 3D shape dataset and proposed a method to classify fine-grained 3D shapes from multiple rendered views to address the above issues. However, the process of view capturing in their approach loses geometric details. Observing their drawbacks in sub-class level retrieval, we propose ROSA-Net that focuses on fine-grained shape retrieval, which can encode both geometric and structural features and weight their importance. We also provide a dataset for quantitatively evaluating fine-grained shape retrieval.

2.2 Mesh-based Representations

There have been numerous studies that investigate how to apply convolution operations on 3D mesh-based models. Masci et al. [33] were the first to extract patches based on local polar coordinates and generalize convolution networks to non-Euclidean manifolds. Sinha et al. [48] used Convolutional Neural Networks (CNNs) to transform a general mesh model into a "geometry image" that encodes local properties of shape surfaces. Anisotropic CNN (ACNN) [3] adopted anisotropic diffusion kernels to construct patches to learn intrinsic correspondences. Monti et al. [34] further improved these ideas by parametrically constructing patch operators through vertex frequency analysis. Alternatively, methods were reported in the literature to perform convolutional operations in the spectral domain. Boscaini et al. [2] used windowed Fourier transform and proposed localized spectral convolutional networks to conduct supervised local feature learning. Xie et al. [58] learned a binary spectral shape descriptor for 3D shape correspondence. Han et al. [19] further proposed a circle convolutional restricted Boltzmann machine (CCRBM) to learn 3D local features in an unsupervised manner. In follow-up work, Hanocka et al. [22] brought up a network with unique convolution and pooling operations on the edges which connect adjacent mesh vertices. Schult et al. [44] proposed a network that applies geodesic and Euclidean convolutional operations in parallel. These methods provide fundamental building blocks for deep learning methods for geometry processing. Inspired by the above works, we propose to use generative models to automatically learn shape descriptors in the latent space.

2.3 Rotation-invariant Representations

Rotation invariance is an important attribute of shape descriptors in real-world applications of 3D shape retrieval. Although studies have been conducted by many researchers, this problem is still insufficiently explored. A common technique in this field is discrete feature aggregation. For instance, the method

[17] by Furuya et al. extracted shape descriptors from an oriented point set by aggregating processed local 3D rotation-invariant features. Similarly, Luo et al. [31] learned an orientation for each point and transformed its neighbors before aggregating neighbors' information into this point. Kanezaki et al. [25] mainly focused on how to aggregate predictions from multiple views and take a single image as input for its prediction. SeqViews2SeqLabels [21] aggregated the sequential views using an encoder-decoder Recurrent Neural Network (RNN) structure with attention. The rotation invariance has been addressed only to a very limited extent because the above methods are not able to deal with arbitrary rotations. Some recent research works suggested incorporating equivariance directly into the network architecture, because the desired equivalence of transformation can be achieved through constraining the filter structure. Thomas et al. [51] introduced tensor fields to keep translational and rotational equivariance. Zhang et al. [61] proposed to represent data by a set of 3D rotations and defined quaternion product units to operate on them. Chen et al. [7] dynamically adapted convolution kernels based on the rotation invariant relative pose information. Deng et al. [11] extended neurons from 1D scalars to 3D vectors, constructing rotation-equivariant learnable networks.

Another way to achieve equivalence is coordinate transformation. Henriques et al. [23] fixed a sampling grid according to Abelian symmetry. Also, equivariant filter orbit was the main focus of many recent works. Cohen et al. [8] proposed group convolution networks (G-CNNs) with the square rotation group. They provided the evidence for the rotational equivariance of group-convolutions. Worrall et al. [56] proposed CubeNet using KleinâĂŹs four-group on 3D voxelized data, which learns interpretable transformations with encoder-decoder networks. Some previous research works have applied functions on the icosahedron and their convolutions to achieve equivariance on the cyclic group [10] and the icosahedral group [15]. Esteves et al. [14] and Cohen et al. [9] focused on the infinite group SO(3), and used the spherical harmonic transform for the exact implementation of the spherical convolution or correlation. Esteves et al. [14] also defined several SO(3) equivariant operations on spheres to process 3D data, which can achieve better invariance and generalizes well to unseen rotations. The question remains open that how the invariance preservation mechanism can be utilized to learn a shape descriptor for fine-grained shape retrieval.

There are also several recent studies focusing on rotation invariant representations on point clouds, which learn an initial rotation to a canonical pose. Qi et al. [39] adopted an auxiliary alignment network to make model robust to affine transformations by predicting and applying such transformations to input points and features, which was then further improved to handle the variations in point density by [40]. Deng et al. [12] proposed ordering-free point pair features and a deep architecture based on PointNet to encode coordinates to transform-invariant features. Adversarial training [32] has also been used to improve model robustness to arbitrary rotations. In [55], rotation of the point cloud was regarded as an attack and rotation robustness was improved by training the classifier on inputs with adversarial rotations. Despite good results on rotation robustness,

these methods cannot be directly applied to retrieve mesh models. Moreover, they did not focus on the rotation invariance of both geometry and structure.

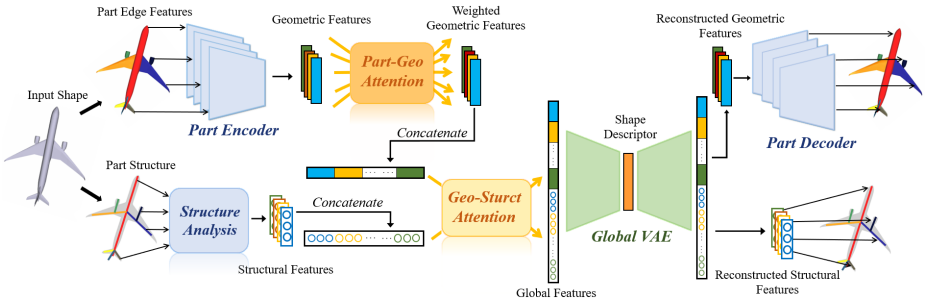


Fig. 2: Pipeline of ROSA-Netto extract shape descriptors for fine-grained intra-class shape retrieval. From left to right, i) our method first extracts part-wise base geometric features and structural features. ii) Then *PartVAE* encodes base geometric feature to latent space, and *Part-Geo Attention* learns the contribution of each geometric part. iii) Meanwhile, shape structure is analyzed and extracted as structural features. iv) After concatenation, the global geometric feature and structural feature are weighted by *Geo-Struct Attention*. v) The weighted geometric and structural features are concatenated and then encoded by a *GlobalVAE*, whose latent vector is the high-level shape descriptor of the input object.

3 ROSA-Net

Our method is inspired by the recent progress in latent vector learning and transformation-invariant feature extraction. To extract 3D shape descriptors with rich geometric details that are robust to rigid transformation, we propose to extract part-wise mesh-based features: edge lengths and dihedral angles. These features preserve rigid transformation invariant and scale-sensitive geometric details, which enable shape reconstruction from these features [16], showing that complete information is retained. Although these geometric features are descriptive, their high-dimensionality means it would be inefficient to use them directly as shape descriptors. Moreover, these features only describe low-level features of edges. It thus lacks information on the global semantic structure of the 3D shape. To address these issues, we adopt a set of variational autoencoders (VAEs) with attention mechanism to extract compact features from the base geometric features, which not only retains the translation and rotation invariance of detailed geometric features, but also balances structure and geometric information to a high-level succinct feature for retrieval tasks. In the following subsections, we first symbolize the elements of ROSA-Net, and then introduce the components of the network.

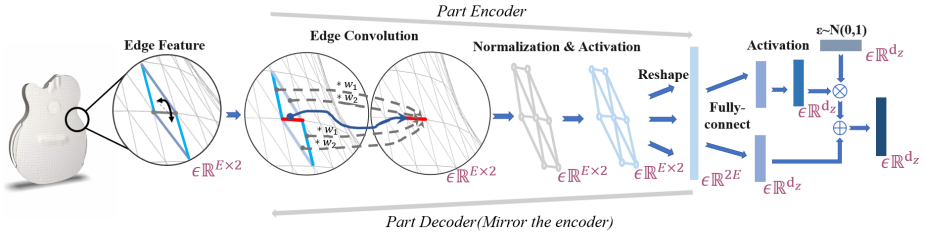


Fig. 3: The structure of one *PartVAE* that extracts rotation-invariant scale-sensitive fine-grained geometric information. *PartVAE* encodes edge features through convolutional operations on edges and their adjacent edges. E and d_z denote the number of edges of a mesh model and the dimension of the latent vector respectively.

3.1 Overview

Fig. 2 illustrates the network architecture of ROSA-Net. Given a 3D shape M_i, the input of the network is its semantically segmented parts $\{m_i{}^p, \forall p \in \{1, 2, \cdots, P\}\}$, where P is the number of parts. We extract its base geometric feature $f_i{}^p$ from each part $m_i{}^p$. Then we use a set of *partVAE*s (part-wise variational autoencoders) to encode a geometric feature set $\{f_i{}^p, \forall p \in [1, P]\}$. Each partVAE encodes the geometric feature of the corresponding part edge-wisely to a latent vector $z_i{}^p$. Furthermore, we adopt a part-geometry (*Part-Geo*) attention mechanism to weight the importance of each semantic part to amplify the effect of important parts by multiplying the latent vector of each part by the attention weight $\alpha^p{}_i$. The weighted latent vector set $\left\{z_i'{}^p, \forall p \in \{1, 2, ..., P\}\right\}$ encodes the geometric information and the contribution of each part to shape discrimination. All the vectors are then concatenated to form a global geometric feature vector gv_i, representing the geometric feature of the whole shape. Similarly, we extract the global structural feature sv_i that is robust to rigid transformation through part-based structure analysis. As the contributions of geometric and structural features to fine-grained retrieval vary in different cases, we further learn the importance of geometric and structural features respectively through geometry-structure (*Geo-Struct*) attention mechanism. In particular, we multiply the global geometric feature gv_i and the global structural feature sv_i by the learned geometry weight $w_i{}^g$ and structure weight $w_i{}^s$ respectively. Finally, the weighted geometric and structure features are concatenated to get the initial global feature vector fv_i, which is then interpreted as a low-dimensional latent vector zv_i by the global feature variational autoencoder (*GlobalVAE*). zv_i will be used as the shape descriptor of the input 3D shape for the fine-grained retrieval task. Additionally, we append triplet loss term to the original VAE loss to improve the distribution of shape features in the latent space and train our attention mechanisms in an end-to-end manner.

3.2 Geometric Feature Representation

In our observation, globally similar objects could share similar features in some semantic parts, but differ drastically in other parts, since 3D objects are often designed and assembled using different parts to satisfy various desired functions. Therefore, compared with learning features at a low level of granularity for an integrated 3D shape, learning them part-wisely is more effective. In real world applications, 3D objects may be randomly placed, thus are not always strictly aligned or zero-centred to the world coordinate system. Therefore, shape descriptors of 3D objects should be invariant to possible rigid transformations. In case of scaling transformation, the feature needs to be capable of describing the relative size of parts for shape discrimination. For example, comparing a coffee table with a dinning table that has the same panel size, the coffee table has shorter legs. Given the above observations, the part-wise geometric feature we extract should be invariant to rotation and translation but sensitive to scaling. Finally, we want the extracted features to contain as much geometric detail as possible, so that the whole 3D shape can be reconstructed from them. Therefore, we represent shapes by edge lengths and dihedral angles, which are reconstructive, scale sensitive and robust to rigid transformations [16].

In particular, the base geometric feature is defined on a set of 3D models that all have the same semantic parts with label $p \in \{1, 2, \cdots, P\}$ and E edges with the same connectivity among them, where P and E are the number of Part Semantics and edges respectively. We denote the parts with the same label p of all the models by the set $\{m_i^p, \forall i \in [1, N]\}$, where N is the total number of models. The same topological structure of all the shapes can be utilized to establish part-level correspondences. In our implementation, we use a watertight unit cube mesh with 3075 vertices as the reference model, and perform non-rigid coarse-to-fine registration [63] on the above shape part set, ensuring that all the shapes have the same connectivity as the reference. Each shape m_i^p could be described by its edge lengths and dihedral angles:

$$f_i^p = \left\{ L_i^p, \Theta_i^p \middle| \, L_i^p \in \mathbb{R}_+^E, \Theta_i^p \in [0, 2\pi)^E \right\} \tag{1}$$

where $L_i^p \in \mathbb{R}_+^E$ contains all E edge lengths, and $\Theta_i^p \in [0, 2\pi)^E$ includes dihedral angles of all edges.

The base geometric feature is not ideal for retrieval tasks because of its high dimensionality and redundancy. Therefore, after basic geometric information extraction, feature compression is a necessary step. We propose to use *partVAEs* (part-wise variational autoencoders) to extract high-level features from the base geometric features. Due to its reconstructive property, the high-level latent vector z_i^p learned by *partVAEs* will not only maintain the translation and rotation invariance, but also preserve necessary detailed geometric features for object retrieval.

Fig. 3 illustrates the structure of one *partVAE*. The input of a *partVAE* is an $E \times 2$ dimensional feature vector, f_i^p, containing above basic geometry features. In f_i^p, it concatenates the rows of an $E \times 2$ matrix, where each row

corresponds to an edge, and the two columns represent edge lengths and dihedral angles between two adjacent faces. The encoder of *partVAE* learns the posterior distribution between the input data f_i^p and the latent vector z_i^p. As the encoder learns local geometric features, convolutional operations on undirected edges are required. Thus, we adapt $MeshConv$ operation [22] to encode our reconstructive base geometric feature. In particular, the input is filtered by the first three convolutional layers, where the convolutional operation on the i^{th} edge e_i is defined as:

$$y_i = W_e * x_i + W_{n_{e,1}} * \frac{\sum_{j \in N_{e_{i,1}}} x_j}{|N_{e_{i,1}}|} + W_{n_{e,2}} * \frac{\sum_{j \in N_{e_{i,2}}} x_j}{|N_{e_{i,2}}|} + b_e \tag{2}$$

where $x_i \in \mathbb{R}^2$ is the feature (edge length and dihedral angle) of edge e_i. Edges are treated as undirectional, and each mesh part is registered from a unit cube, which is a closed, manifold mesh. Therefore, e_i is adjacent to two faces. Within each adjacent face of e_i, in the counter-clockwise order, we refer to the edge immediately after e_i as the first adjacent edge, and the edge immediately after the first adjacent edge as the second adjacent edge. Denote by $N_{e_{i,1}}$ and $N_{e_{i,2}}$ the sets of first and second adjacent edges of e_i, respectively. In our case, as each edge has 4 neighboring edges, the numbers of elements in each set $|N_{e_{i,1}}| = |N_{e_{i,2}}| = 2$. $W_e, W_{n_{e,1}}, W_{n_{e,2}} \in \mathbb{R}^{2\times 2}$ are learnable weights of convolutional operations on an edge and its adjacent edges. $b_e \in \mathbb{R}^2$ is the bias term. Additionally, all convolutional layers are appended with a batch-norm layer and a Leaky-ReLU layer with the slope for negative input $\tilde{\alpha} = 0.02$.

The output of three consecutive convolutional layers is fed into two fully-connected layers to obtain its mean and variance respectively, where the mean z_i^p is the latent vector of *partVAE*. After that, the decoder learns to reconstruct f_i^p from the latent vector z_i^p, the network structure of which is symmetric with the encoder without sharing weights. Each *partVAE* is able to extract high-level latent features of a semantic part, thus depicting geometric information part-wisely. We assume that the number of parts is fixed for all the objects in the same class. However, not all parts need to be present on a given shape. If an input model misses some parts, the input for the corresponding *partVAE*s will be zero-matrices, and the latent vectors of these parts are set as zeros.

3.3 Part Geometry Attention Mechanism

Each semantic part does not contribute equally in shape representation. For example, when measuring the similarity of two car models, car bodies may be more important than car mirrors. Therefore, we further introduce a *Part-Geo* (part geometry) attention mechanism to learn to determine the importance of each part of each shape in fine-grained retrieval.

We define an attention vector $\alpha_i = [\alpha_i^1, \alpha_i^2, ..., \alpha_i^P]$ to denote the importance of each part for object M_i. The higher the value of α_i^p, the more discriminative the part p is when recognizing shapes. Following [54], the attention vector α_i is

obtained by softmax of the dot-products of key vector K_i and query vector Q_i:

$$\alpha_i = softmax(K_i^T \cdot Q_i) \tag{3}$$

where $K_i = [K_i^{(1)}, K_i^{(2)}, ..., K_i^{(P)}]$ represents the key feature of the part i, which is a linear transformation of its latent vector: $K_i^p = W_K^p z_i^p$. The query vector Q_i is the summation of the linear transformation of the latent features of all parts: $Q_i = \sum_p W_Q^p z_i{}^p$. Here, $W_K^p, W_Q^p \in \mathbb{R}^{d_h \times d_z}$, d_z is the dimension of the latent vector, and d_h is the dimension of the key and query features. The attention vector is jointly trained with other parts of the neural network.

Thus, the output of the *Part-Geo* attention mechanism is a set of weighted geometric features, $\{\alpha_i^1 z_i^1, \alpha_i^2 z_i^2, ..., \alpha_i^P z_i^P\}$, which are concatenated and reshaped to a vector $gv_i \in \mathbb{R}^{P \times d_z}$, representing the global geometric information of the object.

3.4 Structural Information Representation

Despite the importance of structural information in shape representation, it has been neglected by existing methods for shape retrieval. Visually similar shapes often share similar structure. Objects that have parts with similar geometric features can be distinguishable when their structures are highly different. Therefore, we incorporate the global structure of an object as part of our representation.

We represent the structural information by the spatial relationships among the semantic parts. Same as geometric features, the proposed structural features are also robust to rigid transformation. For each class of 3D objects, we first define one semantic part as the body part that all models must contain. If there are more than one common semantic parts in all models, we select the part with the largest average volume. We observe that the existence of non-body parts and their relative positions to the body part are important for shape discrimination, which can be used to interpret structural information of shapes. As shown in Fig. 2, we describe the structure of objects by an 11-dimensional vector sv_i, defined as follows:

- $sv_1 \in \{0, 1\}$ denotes whether the part exists in the input 3D shape.
- $sv_2 \in \mathbb{R}$ denotes the distance from the center of the current part to the center of the body part.
- $[sv_3, sv_4, sv_5] \in [-1, 1]^3$ denotes the cosine of the angles between the first principal component of the current part and the first three principal components of the body part respectively.
- $[sv_6, sv_7, sv_8] \in [-1, 1]^3$ denotes the cosine of the angles between the second principal component of the current part and the first three principal component of the body part respectively.
- $[sv_9, sv_{10}, sv_{11}] \in \mathbb{R}^3$ denotes the unit direction from the center of the body part to the center of the current part.

3.5 Geometry-Structure Attention Mechanism

Although the objects of the same class all share similar structures, objects belonging to different classes have different diversity in the composition. For example, all guitars share the same structure, but chairs have diverse compositions. Thus geometric information and structural information could have different contribution when discriminating objects belonging to different classes. Therefore, the extracted compact global geometric and structural features gv_i and sv_i need to be re-weighted to balance their contributions to the final shape representation. To achieve this, we introduce a *Geo-Struct* (geometry and structure) attention mechanism to learn to balance the importance of structure and geometric information in shape representation. We define a score vector $w_i = [w_i^g, w_i^s] \in [0,1]^2$ for model M_i, representing the weights of geometric and structural information. The score vector is learned through two fully-connected sub-networks:

$$w_i = softmax([F(gv_i), G(sv_i))]) \tag{4}$$

where $F : \mathbb{R}^{d_z} \rightarrow [0,1]$ and $G : \mathbb{R}^{11} \rightarrow [0,1]$. In implementation, $F(\cdot)$ contains two fully-connected layers, and the dimension of output vectors of the two layers are 32 and 1 respectively. $G(\cdot)$ contains two fully-connected layers as well, where the dimension of output vectors are 16 and 1 respectively. The global geometric feature and structural features are multiplied by their corresponding weight scores respectively and then concatenated to form the global feature: $fv_i \in \mathbb{R}^{P\times(d_z+11)}$, which contains weighted geometric and structural information of the object.

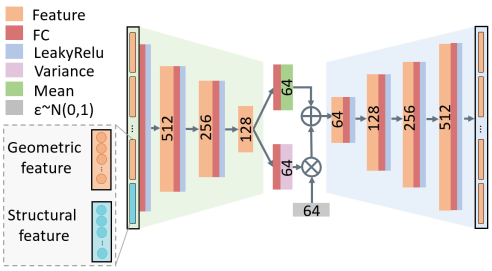


Fig. 4: Network architecture of *GlobalVAE*. GlobalVAE encodes part geometric features and global structural features through three consecutive fully-connected layers. The latent vector is used as the shape descriptor for fine-grained intra-class shape retrieval.

3.6 Global Feature Encoding

To encode both geometric and structural feature into one latent space, we further use a *GlobalVAE* (global feature variational autoencoder) to encode global geometric and structural information into a reconstructive compact representation. We use the architecture of globalVAE illustrated in Fig. 4. The *GlobalVAE* comprises three fully-connected layers, each appended with a leaky Relu layer with the slope of negative input $\tilde{\alpha} = 0.02$. The structure of *GlobalVAE* ensures that its latent vector zv_i contains the geometric information of all parts as well as the global structural information.

3.7 Losses

To optimize the network parameters of our model, we adopt three loss-terms that enable a distriminative latent space for fine-grained shape retrieval.

VAE Losses. Our optimization objective function includes a Kullback-âĂŞLeibler (KL) divergence term and a reconstruction term for each *partVAE* and the *GlobalVAE*. The KL divergence term regularizes the latent space, while the reconstruction term ensures that the input features can be explained by our autoencoders. Therefore, the loss function for all the partVAEs includes the KL divergence terms and the terms measuring the differences between all the input base geometric features and their decoding results:

$$L_{VAE}^{part} = \frac{1}{P_i}\sum_{p=1}^{P_i}(f_i^p - f'^p_i)^2 + \gamma\sum_{p=1}^{P_i} D^p_{KL_p}(q(z_i^p|f'^p_i) \parallel p(z_i^p)) \tag{5}$$

where P_i is the number of the parts of model M_i, f_i^p is the base feature of the the p^{th} part, $f_i'^p$ is the reconstructed feature of the p^{th} part. In the second term, γ is a weight that balances both terms, $p(z_i^p)$ is the prior probability distribution, $q(z_i^p|f'^p_i)$ denotes the posterior probability, and $D^i_{KL_{part}}$ denotes the KL divergence of the p^{th} partVAE. γ is a constant, which is set as 1×10^5 in our experiments. We define the loss for the *GlobalVAE* in a similar way.

Triplet Losses. Using above losses for VAEs, the distribution of the latent vectors is able to cluster the models of the repository in the feature space to some extent. However, it can be further optimized by minimizing the distance between the features of similar shapes and enforcing a margin between dissimilar shapes. Besides, we use a triplet loss [43] to optimize the final feature distribution in the latent space, which also helps the attention mechanisms to find the distinguished parts and balance the importance of structure information.

For the globalVAE, we define the term as:

$$L_{triplet}^{global} = \sum_i [D(zv_i^a, zv_i^p) - D(zv_i^a, zv_i^n) + \eta]_+ \tag{6}$$

where zv_i^a, zv_i^p and zv_i^n are the latent features of an anchor model (i.e., a chosen model in the training iteration), a positive model (i.e., a model of the same sub-class as the anchor model) and a negative model (i.e., a model of a different sub-class from the anchor model). $D(\cdot,\cdot)$ is a measure of distance between two vectors in the latent space. We use the Euclidean distance $D(v_1, v_2) = \|v_1 - v_2\|_2^2$ in our experiments. η is the threshold of the margin between the distances from the reference model to the similar and dissimilar models. In implementation, the Euclidean distance $D(\cdot,\cdot)$ between features are normalized to $[0, 1]$ in each batch. We set η to 0.3 in the following experiments.

For the set of *partVAE*s, we define a triplet loss term as:

$$L_{triplet}^{part} = \sum_i [D(gv_i^a, gv_i^p) - D(gv_i^a, gv_i^n) + \eta]_+ \tag{7}$$

We use the term to refine the distribution of the global geometric feature gv_i for the entire set of partVAEs.

Overall Loss. The overall loss of ROSA-Net is:

$$L = L_{VAE}^{part} + \lambda_1 L_{VAE}^{global} + \lambda_2 L_{triplet}^{part} + \lambda_3 L_{triplet}^{global} \tag{8}$$

where λ_1, λ_2 and λ_3 are hyper-parameters to balance the weights of different loss terms, which are defaultly set as 1×10^3, 1×10^2, and 1×10^2 in our all experiments.

3.8 Model Training and Shape Retrieval

We feed 3D shapes of all the sub-classes of the same class to ROSA-Net to learn to encode base features into a latent feature for that class. The *partVAE* set with *Part-Geo* attention mechanism first learns a high-level geometric feature set from base geometric features, using the reconstruction loss, KL losses, and triplet losses. At the same time, the high-level geometric feature set is balanced with structural feature and then fed to the *GlobalVAE* to learn to generate the final latent feature vector, where we minimize the same three types of loss terms during training. In addition, we use the objects of the same sub-class as similar shapes and objects of different sub-classes as dissimilar shapes when minimizing the triplet losses.

For an input 3D object M_i, we use the latent vector of the *GlobalVAE* gv_i as its shape descriptor. For each query shape, we rank the shapes in the repository according to the Euclidean distance between their shape descriptors. Note that the distance metric used in retrieval is the same as in the triplet loss, which is the Euclidean distance.

4 Experimental Results

We quantitatively and qualitatively compare the performance of ROSA-Net with other shape retrieval methods or shape descriptors on fine-grained intra-class 3D shape retrieval. We prove the effectiveness of the major components in ROSA-Net and show that ROSA-Nethelps retrieval on other data representation. This section is structured as follows: we first describe ROSA-dataset, which is an intra-class retrieval dataset we construct to evaluate the intra-class fine-grained shape retrieval performance. Then, we provide qualitative and quantitative comparisons between ROSA-Netand other methods on intra-class fine-grained shape retrieval. After that, we visualize and explain how Part-Geo Attention and Geo-Struct Attention automatically balance all sources of information. Also, we show that ROSA-Netis able to deal with other data representations. Lastly, we evaluate the effectiveness of the major components in our network.

4.1 ROSA-dataset

Retrieving a model against shapes within the same class but belonging to different sub-classes is a typical fine-grained shape retrieval task, where the

repository contains globally similar shapes that differ in some details. Most of the existing large-scale 3D object datasets are annotated with only class level labels, which is not suitable for fine-grained retrieval task. For example, ModelNet [57] contains 662 object categories but only a few of them are given sub-class labels. Part of ShapeNet models [5] also have intra-class semantic labels, but the labeling precision and amount are not sufficient to test intra-class retrieval methods. Although FRGC v2 dataset [37] is labeled with an intra-class manner, the dataset only focuses on facial recognition rather than common objects retrieval. FG3D [30] labels an intra-class 3D dataset, but it only contains 3 category of data - car, plane and chair. We build a new dataset with more detailed sub-class level annotations, designed for training, evaluating and comparing fine-grained shape retrieval methods, which is used to demonstrate the effectiveness of our latent descriptor for fine-grained shape retrieval. It could be used to support and evaluate future research on fine-grained 3D shape classification and retrieval task. The shapes used in our fine-grained 3D object retrieval dataset is a subset of SDM-NET data [18], containing 8,906 3D models from 6 object categories. The 6 categories are knife, guitar, car, plane, chair and table, which are further grouped into 175 sub-classes. Each sub-class of models is annotated with semantic labels, which are defined by their distinguishable features compared with other sub-classes, such as functionality, product model number and style. Take the category of guitar as an example, objects are further categorized into twelve sub-classes including double-neck guitar, acoustic, cutaway, Flying V, Gibson Explore, Gibson Les Paul, etc. These semantic labels are assigned according to their style and standard model number.

All of the intra-class labels were annotated manually, and the original annotation from ShapeNet is used as a reference for our annotators. Additionally, some unrealistic shapes that are hard to be categorized were discarded, because keeping them would confuse the retrieval method by using ambiguous sub-class labels, leading to inaccurate quantitative analysis. Please refer to our supplementary materials for more detailed information on the dataset.

We evaluated the performance of ROSA-Net in the fine-grained 3D object retrieval task on the ROSA-Dataset introduced in the previous subsection. We randomly selected 80% of objects in each sub-class as training, and the rest 20% as validation. We tested the performance of ROSA-Net on all the 6 object categories. To demonstrate the robustness of ROSA-Net to random rotation, we perturbed all models in the dataset by transforming each model with a random rotation in $SO(3)$. All experiments were conducted on the randomly rotated shape dataset.

4.2 Fine-Grained Shape Retrieval

In the fine-grained object retrieval task, we query a shape among the shapes of the same class in the dataset. As mentioned above, all shapes are randomly rotated in the experiment. ROSA-Net is trained to extract latent feature vectors for shape representation, which are then used to measure the similarity between the query and all the shapes of the same class. The similarity between shapes

Table 1: Evaluate the intra-class fine-grained 3D shape retrieval performance of ROSA-Netand other methods using five metrics on ROSA-dataset test set. ROSA-Netoutperforms other methods on all metrics.

	micro					macro				
Methods	NN	FT	ST	ndcg	mAP	NN	FT	ST	ndcg	mAP
SHD [26]	0.162	0.134	0.245	0.436	0.162	0.090	0.075	0.146	0.352	0.111
LFD [6]	0.224	0.183	0.313	0.486	0.214	0.128	0.107	0.193	0.401	0.153
RotationNet [51]	0.416	0.182	0.243	0.514	0.296	0.146	0.123	0.223	0.403	0.144
FG3D [30]	0.538	0.351	0.424	0.591	0.531	0.351	0.217	0.333	0.522	0.350
VN-DGCNN [11]	0.482	0.296	0.372	0.545	0.465	0.285	0.179	0.261	0.423	0.301
ART-DGCNN [55]	0.423	0.305	0.338	0.525	0.411	0.307	0.170	0.280	0.414	0.312
PaRI-Conv [7]	0.524	0.338	0.432	0.573	0.468	0.297	0.207	0.292	0.459	0.316
MeshCNN [22]	0.447	0.318	0.363	0.527	0.461	0.329	0.197	0.283	0.448	0.347
MeshMAE [29]	0.467	0.341	0.384	0.521	0.442	0.311	0.196	0.307	0.434	0.333
SubdivNet [24]	0.504	0.314	0.425	0.601	0.478	0.325	0.189	0.286	0.449	0.364
ROSA-Net	**0.624**	**0.556**	**0.675**	**0.770**	**0.598**	**0.588**	**0.456**	**0.551**	**0.697**	**0.502**

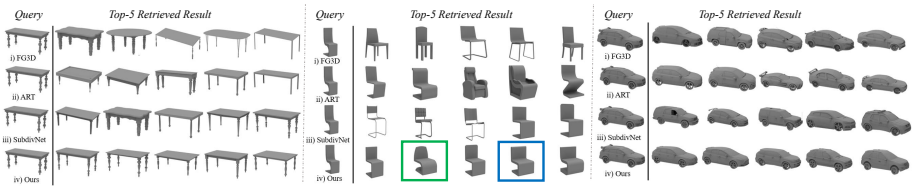


Fig. 5: Visual comparison of retrieval results among FG3D, ART, SubdivNet and our method. Our method outperforms the other methods on geometric local feature encoding and fine-grained object retrieval.

is measured by the Euclidean distance between shape descriptors. If a retrieved shape is in the same sub-class with the query shape, we denote it as a successful retrieval. Fig. 1 shows the top five retrieved results for four query shapes on chair category. Since ROSA-Net is able to find shapes with matched geometric details, it performs well in the sub-class retrieval. Specifically, as shown in the first row of Fig. 1, ROSA-Net captures that the query object has grid cotton pad on the chair back, and then retrieves shapes with similar features. Meanwhile, the retrieved results with lower rankings also show the capability of ROSA-Net. As shown in the last row, the query shape has a slat back and turned legs. ROSA-Net successfully retrieves shapes with matched features, which are in the same sub-class as the query shape. For more retrieval results, please refer to the supplementary material.

We compare ROSA-Net with other alternative approaches. We compare with other shape retrieval methods, including Spherical Harmonics descriptor (SHD) [26], LightField descriptor (LFD) [6], RotationNet [51], and FG3D [30]. We also compare rotation-equivariant shape descriptor, Vector Neuron(VN) [11], rotation-robust shape descriptor ART [55], and rotation-invariant shape

descriptor PaRI-Conv [7]. Additionally, we compare with other mesh-based shape descriptors, including MeshCNN [22], MeshMAE [29], and SubdivNet [24]. We use the best setting for all of the methods above for comparison.

Table 1 shows the quantitative comparison between ROSA-Net and other methods on ROSA-dataset. The results are evaluated using various statistical metrics: Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), Normalized Discounted Cumulative Gain (NDCG) and mean Averaged Precision (mAP). Considering the imbalance of model numbers of the sub-classes, each measure is calculated through both micro and macro average. The macro averaging computes the metric independently for each sub-class and then takes their average as the final overall metric, whereas the micro averaging is a weighted average with the weight for each sub-category proportional to the number of objects in it. Noticeably, ROSA-Net outperforms other state-of-the-art retrieval methods, rotation-robust shape descriptors, and mesh-based shape descriptors on intra-class fine-grained 3D shape retrieval, showing the effectiveness of our feature extraction strategy.

Fig. 5 provides examples of top-5 retrieved results by using FG3D, ART-DGCNN, Subdivnet, and ROSA-Net respectively. Note all of these comparison methods are robust to random shape rotation. These retrieval methods or shape descriptors can achieve fine-grained representation to some extent. However, only ROSA-Net can retrieve top-5 shapes all similar to query shape robustly. Our method is scale-sensitive as shown it prefers shapes with similar scale(green) instead of similar local detail(blue).

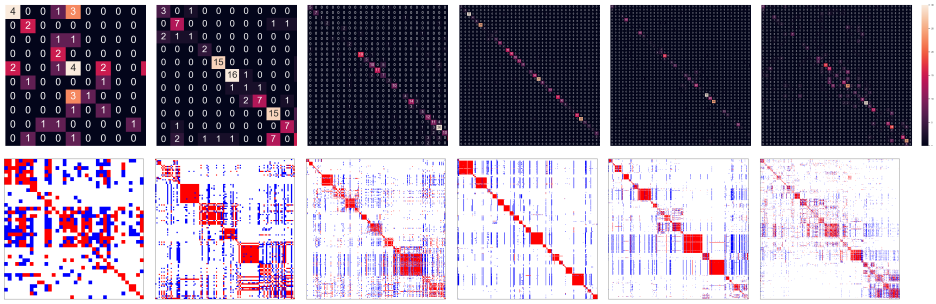
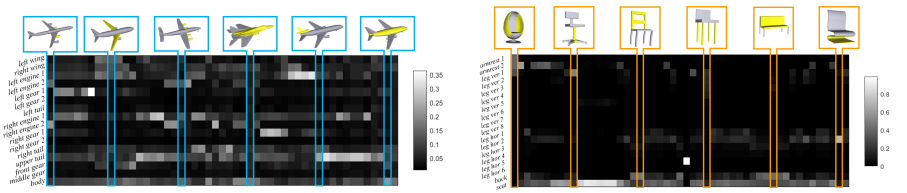


Fig. 6: Illustration of the confusion matrix and tier images of ROSA-Net on 6 classes. The confusion matrix shows all the shapes are well-classified and the tier images show that similar shapes are clustered together.

In Fig. 6, we use the confusion matrix and tier image [47] to visualize the global retrieval results of our method in each category. The confusion matrix shows shapes are classified correctly. In a tier image, each row represents a query with model j. Pixel (i, j) is filled by black, red, and blue if model i is the nearest neighbor, first tier match, and the second tier match of j respectively. Along each axis, models are grouped by sub-class, and lines are added to separate each

sub-class. Note that the tier image is not diagonally symmetric because of the imbalance shape number in each class. In each sub-class, red pixels are clustered in blocks along the diagonal, showing that models of the same sub-class are each other's first-tier matching results. Moreover, second-tier matches of each sub-class tend to congregate together in the same block, implying that ROSA-Net learns the similarity between sub-classes.

4.3 Weighted Features of Parts by Part-Geo Attention



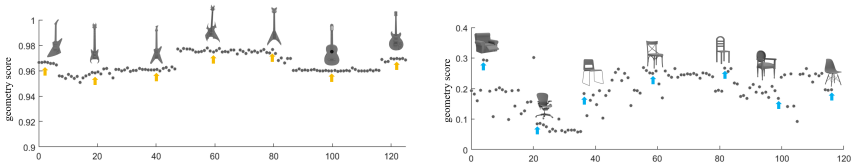
(a) plane (b) chair

Fig. 7: Part geometry attention visualization on the plane and chair datasets.

Our model utilizes the *Part-Geo* attention mechanism to balance the contributions of different parts when discerning a shape among sub-classes. Fig. 7 visualizes the learned attention information that demonstrates the importance of each part when learning the final latent descriptor, where we highlight the valid parts using the learned attention weights. Note that the weights of missing parts are set to 0. Aside from those missing parts, the discriminative parts of the shapes are successfully assigned with relatively high weights, meaning that they contribute more than the other parts to the latent feature learning. More specifically, in Fig. 7(a), the highest weight of each plane appears on the most discriminative part: engine, wing, tail or body. In Fig. 7(b), the weights of existing parts also conform to their discriminativeness. The highest weights are assigned to leg, arm, back and seat respectively.

4.4 Weighted Features by Geo-Struct Attention

The *Geo-Struct* attention mechanism balances the geometry and structure information in our shape representation. Fig. 7 visualizes the learned geometry score w^g in guitar and chair datasets. The structure score is calculated by: $w^s = 1 - w^g$. In each dataset, we randomly selected some model in the test set. Models are grouped by sub-classes along x-axis, and their geometry scores are shown by their y-coordinates. Each point represents an object instance, and we draw the 3D shapes of several representative instances for better visualization. Note that shapes of the same sub-class that share similar structure/geometry features tend to have similar structure/geometry scores. Also, the average geometry score of



(a) guitar (b) chair

Fig. 8: Visualization of geometry-structure attention information on guitar and chair datasets.

Table 2: The comparison between point cloud input and mesh input.

	micro					macro				
Methods	NN	FT	ST	ndcg	mAP	NN	FT	ST	ndcg	mAP
PaRI-Conv	0.5296	0.3479	0.4409	0.5794	0.4766	0.3056	0.2107	0.2964	0.4654	0.3216
ROSA-Net(PC)	0.5812	0.5251	0.6352	0.7166	0.5561	0.5616	0.4325	0.5256	0.6492	0.4743
ROSA-Net	0.6552	0.5955	0.7215	0.8115	0.6295	0.6305	0.4902	0.5943	0.7368	0.5369

the class of guitar is higher than chair, indicating that distinguishing guitars relies more on geometric information than chairs.

4.5 Using Other Data Representation

In this experiment, we investigate whether our approach is sufficiently practical to be combined with existing shape segmentation methods to retrieve a shape in other representations. Instead of manually segmenting the shapes into parts, we use the following way to automatically obtain the segmentation results. We randomly and uniformly sampled 2048 points from each watertight mesh, and then feed the points to PaRI-Conv [7] for segmentation. With the semantically labeled points, we align watertight mesh models to the points and assign each triangular mesh with the label of the closest point. Then we use the segmented meshes to train and test our model. Fig. 2 shows the performance of ROSA-Neton the plane subcategory, compared with the PaRI-Conv and the original method using mesh-based representation. Note that our approach is able to tolerate minor part segmentation errors, indicating its capability to be used in real-world shape retrieval applications. Also, our method performs better than the original PaRI-Conv, indicating other methods can use ROSA-Netto improve its performance on fine-grained shape retrieval.

4.6 Ablation Study

We now provide the results of a detailed ablation study that shows the contribution of each component to the overall performance, including base geometric feature selection, the feature extractor selection and the hierarchical structure.

Table 3: Ablation study proves the effectiveness of all the components.

	micro					macro				
Methods	NN	FT	ST	ndcg	mAP	NN	FT	ST	ndcg	mAP
wo Struct-info	0.5710	0.4990	0.6073	0.6932	0.5362	0.5405	0.4100	0.4961	0.6266	0.4518
wo Geo-Atten	0.5811	0.5140	0.6270	0.7171	0.5576	0.5502	0.4202	0.5103	0.6393	0.4636
wo Struct-Geo	0.6109	0.5433	0.6621	0.7546	0.5859	0.5772	0.4461	0.5391	0.6831	0.4920
ROSA-Net	**0.6237**	**0.5555**	**0.6748**	**0.7702**	**0.5977**	**0.5882**	**0.4556**	**0.5512**	**0.6969**	**0.5020**

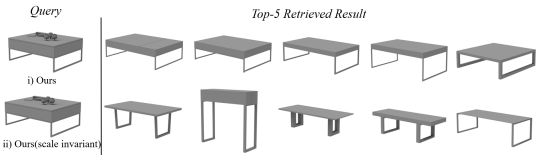


Fig. 9: Comparison between scale-invariant and scale-sensitive feature as a base geometric feature.

Geometric feature extraction. In this subsection, we show the effectiveness of the adopted scale-sensitive features and a reconstructive network to describe part geometry. We design a comparison group that replaces our base geometric feature with the 5 dimensional shape feature proposed by Hanocka et al. [22] as a scale-insensitive feature, which is denoted as "Ours(scale sensitive)". We design another comparison group by replacing *PartVAEs* with the classification network in [22], which is denoted as "Ours(CNN)".

We provide two examples for visual comparison. Fig. 9(a) is a comparison of different choices of base features. Using the scale-invariant feature as in MeshCNN [22], although the semantic parts of retrieved shapes all look similar to the parts of the query shape, the size relationship is not maintained among parts, leading to dissimilar overall shapes.

In Fig. 9, a failure case of using CNN instead of our *PartVAE* structure is provided. Using the classification network for feature extraction sometimes fails to capture geometric features from thin geometry. In contrast, with the reconstructive network of our *PartVAE*, the detailed geometric features are well-learned in the latent space of our VAE modules as shown in the experiment, showing the effectiveness of our PartVAEs.

Structure information. Finally, we compare the performance of ROSA-Net with and without structural information, without geometry attention, and without structure-geo attention in Table 3. The full model performs the best, indicating the effectiveness of all the components.

5 Conclusion

In this paper, we introduce ROSA-Net, a novel framework to extract shape descriptors for fine-grained 3D object retrieval. ROSA-Net can extract 3D shape descriptors with geometric details and global structural information, which are robust to rotation and sensitive to scale. Trained with the attention mechanisms and the dedicatedly designed losses, ROSA-Net can locate and

emphasize discriminative parts, and make a balance between structure and geometric information when representing a 3D shape. Thanks to the above designs, the Through fruitful experiments on fine-grained 3D shape retrieval, we demonstrated that ROSA-Net outperforms the state of the art on fine-grained 3D object retrieval tasks.

Limitations and Future work. Firstly, ROSA-Net encodes structural information based on the spatial relationships between semantic parts, which leads to a need for correspondence between the segmented parts of the query shape and the dataset. Incorporating unsupervised co-segmentation methods would be a natural extension of this work. Secondly, to encode fine-grained geometric details, ROSA-Net represents shapes with meshes of the same topology. However, the registration or the re-meshing process takes extra processing time. Future work could incorporate methods that are robust to meshes with different typologies. Finally, a notable limitation is the sensitivity of the shape descriptor to scale and structure variations. Consequently, descriptors for identical shapes at different scales may exhibit significant differences. As shown in Fig. 5, ROSA-Net regards the chair with a round corner with a similar scale as a similar shape(highlighted in green), rather than the chair with a sharp corner with a different scale(highlighted in blue). For future work, fine-grained sketch-based and image-based object retrieval would be a natural extension of our work. Also, with our method retrieving fine-grained similar shapes, how to utilize the retrieved shapes for modeling of new shapes is worth also exploring.

Acknowledgment

This work was supported by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), the National Natural Science Foundation of China (No. 62061136007) and the Youth Innovation Promotion Association CAS.

References

1. Bai, S., Bai, X., Zhou, Z., Zhang, Z., Jan Latecki, L.: Gift: A real-time and scalable 3D shape search engine. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5023–5032 (2016) 4
2. Boscaini, D., Masci, J., Melzi, S., Bronstein, M.M., Castellani, U., Vandergheynst, P.: Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. In: Computer Graphics Forum. vol. 34, pp. 13–23. Wiley Online Library (2015) 5
3. Boscaini, D., Masci, J., Rodolà, E., Bronstein, M.: Learning shape correspondence with anisotropic convolutional neural networks. In: Advances in neural information processing systems. pp. 3189–3197 (2016) 5
4. Bronstein, A.M., Bronstein, M.M., Guibas, L.J., Ovsjanikov, M.: Shape google: Geometric words and expressions for invariant shape retrieval. ACM Transactions on Graphics (TOG) **30**(1), 1–20 (2011) 4

5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: ShapeNet: An information-rich 3D model repository. arXiv preprint arXiv:1512.03012 (2015) 15
6. Chen, D.Y., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3D model retrieval. In: Computer graphics forum. vol. 22, pp. 223–232. Wiley Online Library (2003) 4, 16
7. Chen, R., Cong, Y.: The devil is in the pose: Ambiguity-free 3d rotation-invariant learning via pose-aware convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2022) 6, 16, 17, 19
8. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International conference on machine learning. pp. 2990–2999 (2016) 6
9. Cohen, T.S., Geiger, M., Köhler, J., Welling, M.: Spherical CNNs. arXiv preprint arXiv:1801.10130 (2018) 6
10. Cohen, T.S., Weiler, M., Kicanaoglu, B., Welling, M.: Gauge equivariant convolutional networks and the icosahedral CNN. arXiv preprint arXiv:1902.04615 (2019) 6
11. Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., Guibas, L.: Vector neurons: a general framework for so(3)-equivariant networks. arXiv preprint arXiv:2104.12229 (2021) 6, 16
12. Deng, H., Birdal, T., Ilic, S.: PPFNet: Global context aware local features for robust 3D point matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 195–205 (2018) 6
13. Duncan, N., Yu, L.F., Yeung, S.K.: Interchangeable components for hands-on assembly based modelling. ACM Transactions on Graphics (TOG) **35**(6), 1–14 (2016) 3
14. Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning so (3) equivariant representations with spherical cnns. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–68 (2018) 4, 6
15. Esteves, C., Xu, Y., Allen-Blanchette, C., Daniilidis, K.: Equivariant multi-view networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1568–1577 (2019) 6
16. Fröhlich, S., Botsch, M.: Example-driven deformations based on discrete shells. In: Computer graphics forum. vol. 30, pp. 2246–2257. Wiley Online Library (2011) 7, 9
17. Furuya, T., Ohbuchi, R.: Deep aggregation of local 3D geometric features for 3D model retrieval. In: BMVC. vol. 7, p. 8 (2016) 4, 6
18. Gao, L., Yang, J., Wu, T., Yuan, Y.J., Fu, H., Lai, Y.K., Zhang, H.: SDM-NET: Deep generative network for structured deformable mesh. ACM Transactions on Graphics (TOG) **38**(6), 1–15 (2019) 15
19. Han, Z., Liu, Z., Han, J., Vong, C.M., Bu, S., Li, X.: Unsupervised 3D local feature learning by circle convolutional restricted Boltzmann machine. IEEE Transactions on Image Processing **25**(11), 5331–5344 (2016) 5
20. Han, Z., Lu, H., Liu, Z., Vong, C.M., Liu, Y.S., Zwicker, M., Han, J., Chen, C.P.: 3d2seqviews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. IEEE Transactions on Image Processing **28**(8), 3986–3999 (2019) 4
21. Han, Z., Shang, M., Liu, Z., Vong, C.M., Liu, Y.S., Zwicker, M., Han, J., Chen, C.P.: Seqviews2seqlabels: Learning 3D global features via aggregating sequential views by RNN with attention. IEEE Transactions on Image Processing **28**(2), 658–672 (2018) 4, 6

22. Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., Cohen-Or, D.: MeshCNN: a network with an edge. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019) 5, 10, 16, 17, 20
23. Henriques, J.F., Vedaldi, A.: Warped convolutions: Efficient invariance to spatial transformations. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1461–1469. JMLR. org (2017) 6
24. Hu, S., Liu, Z., Guo, M., Cai, J., Huang, J., Mu, T., Martin, R.R.: Subdivision-based mesh convolution networks. ACM Trans. Graph. **41**(3), 25:1–25:16 (2022). https://doi.org/10.1145/3506694, https://doi.org/10.1145/3506694 16, 17
25. 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. pp. 5010–5019 (2018) 4, 6
26. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Symposium on geometry processing. vol. 6, pp. 156–164 (2003) 4, 16
27. Kleiman, Y., van Kaick, O., Sorkine-Hornung, O., Cohen-Or, D.: Shed: shape edit distance for fine-grained shape similarity. ACM Transactions on Graphics (TOG) **34**(6), 1–11 (2015) 4, 5
28. Li, B., Godil, A., Aono, M., Bai, X., Furuya, T., Li, L., López-Sastre, R.J., Johan, H., Ohbuchi, R., Redondo-Cabrera, C., et al.: SHREC'12 track: Generic 3D shape retrieval. In: 3DOR. vol. 6 (2012) 4
29. Liang, Y., Zhao, S., Yu, B., Zhang, J., He, F.: Meshmae: Masked autoencoders for 3d mesh data analysis. In: European Conference on Computer Vision (2022) 16, 17
30. Liu, X., Han, Z., Liu, Y.S., Zwicker, M.: Fine-grained 3d shape classification with hierarchical part-view attentions. IEEE Transactions on Image Processing (2021) 5, 15, 16
31. Luo, S., Li, J., Guan, J., Su, Y., Cheng, C., Peng, J., Ma, J.: Equivariant point cloud analysis via learning orientations for message passing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18932–18941 (June 2022) 6
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017) 6
33. Masci, J., Boscaini, D., Bronstein, M., Vandergheynst, P.: Geodesic convolutional neural networks on riemannian manifolds. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 37–45 (2015) 5
34. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5115–5124 (2017) 5
35. Ohbuchi, R., Furuya, T.: Distance metric learning and feature combination for shape-based 3D model retrieval. In: Proceedings of the ACM workshop on 3D object retrieval. pp. 63–68 (2010) 4
36. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Transactions on Graphics (TOG) **21**(4), 807–832 (2002) 4
37. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 947–954. IEEE (2005) 15

38. Pratikakis, I., Spagnuolo, M., Theoharis, T., Veltkamp, R.: Learning the compositional structure of man-made objects for 3D shape retrieval. In: Eurographics Workshop on 3D Object Retrieval (2010) (2010) 4
39. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 6
40. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017) 6
41. Redondo-Cabrera, C., Lopez-Sastre, R.J., Acevedo-Rodriguez, J., Maldonado-Bascon, S.: Surfing the point clouds: Selective 3D spatial pyramids for category-level object recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3458–3465. IEEE (2012) 4
42. Savva, M., Yu, F., Su, H., Aono, M., Chen, B., Cohen-Or, D., Deng, W., Su, H., Bai, S., Bai, X., et al.: SHREC16 track: largescale 3D shape retrieval from ShapeNet core55. In: Proceedings of the eurographics workshop on 3D object retrieval. vol. 10 (2016) 4
43. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015) 13
44. Schult, J., Engelmann, F., Kontogianni, T., Leibe, B.: DualConvMesh-Net: Joint geodesic and euclidean convolutions on 3d meshes. arXiv preprint arXiv:2004.01002 (2020) 5
45. Shen, W., Norrie, D.H., Barthès, J.P.: Multi-agent systems for concurrent intelligent design and manufacturing. CRC press (2003) 3
46. Shi, B., Bai, S., Zhou, Z., Bai, X.: Deeppano: Deep panoramic representation for 3-d shape recognition. IEEE Signal Processing Letters **22**(12), 2339–2343 (2015) 4
47. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The Princeton shape benchmark. In: Proceedings Shape Modeling Applications, 2004. pp. 167–178. IEEE (2004) 17
48. Sinha, A., Bai, J., Ramani, K.: Deep learning 3D shape surfaces using geometry images. In: European Conference on Computer Vision. pp. 223–240. Springer (2016) 5
49. 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. pp. 945–953 (2015) 4
50. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: Computer graphics forum. vol. 28, pp. 1383–1392. Wiley Online Library (2009) 4
51. Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., Riley, P.: Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds. arXiv preprint arXiv:1802.08219 (2018) 6, 16
52. Uy, M.A., Huang, J., Sung, M., Birdal, T., Guibas, L.: Deformation-aware 3D model embedding and retrieval. arXiv preprint arXiv:2004.01228 (2020) 2, 4
53. Vahrenkamp, N., Westkamp, L., Yamanobe, N., Aksoy, E.E., Asfour, T.: Part-based grasp planning for familiar objects. In: 2016 IEEE-RAS 16th international conference on humanoid robots (humanoids). pp. 919–925. IEEE (2016) 2
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 10

55. Wang, R., Yang, Y., Tao, D.: Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14351–14360 (2022) 6, 16
56. Worrall, D., Brostow, G.: CubeNet: Equivariance to 3D rotation and translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 567–584 (2018) 6
57. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015) 15
58. Xie, J., Wang, M., Fang, Y.: Learned binary spectral shape descriptor for 3D shape correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3309–3317 (2016) 5
59. Xie, X., Xu, K., Mitra, N.J., Cohen-Or, D., Gong, W., Su, Q., Chen, B.: Sketch-to-design: Context-based part assembly. In: Computer Graphics Forum. vol. 32, pp. 233–245. Wiley Online Library (2013) 4
60. Xu, K., Chen, K., Fu, H., Sun, W.L., Hu, S.M.: Sketch2Scene: sketch-based co-retrieval and co-placement of 3D models. ACM Transactions on Graphics (TOG) **32**(4), 1–15 (2013) 4
61. Zhang, X., Qin, S., Xu, Y., Xu, H.: Quaternion product units for deep learning on 3D rotation groups. arXiv preprint arXiv:1912.07791 (2019) 6
62. Zhou, Y., Tuzel, O.: VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2018) 4
63. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time non-rigid reconstruction using an RGB-D camera. ACM Transactions on Graphics (ToG) **33**(4), 1–12 (2014) 9