Multimedia Tools and Applications (2021) 80:973–992
https://doi.org/10.1007/s11042-020-09764-y

Deep learning for non-rigid 3D shape classification based on informative images

Li Han[1] · Jingyu Piao[1] · Yuning Tong[1] · Bing Yu[1] · Pengyan Lan[1]

Received: 17 March 2020 / Revised: 30 June 2020 / Accepted: 28 August 2020 /
Published online: 5 September 2020

Abstract
In order to enhance the discernment of features in view-based 3D shape recognition, we propose a joint convolutional neural network (CNN) learning model based on informative images. It learns deep features from intrinsic feature images and extrinsic 2D views, and generates a synthetic feature vector via weighted aggregation and refinement process, which has achieved remarkable improvement in non-rigid 3D shape classification. Our joint CNNs model contains three parts: the first part is the geometry-based feature generation unit. We provide a discriminative BoF (bag of features) image descriptor and construct CNN framework to learn the geometric features of the model. The second part is the view-based feature generation unit. We establish a parallel CNN to extract spatial features from optimized 2D views. The third part is a score generation and refinement unit, which automatically learns the weighted scores of geometric features and spatial features. Finally, the aggregated feature is refined in a CNN framework and serves as an informative shape descriptor for recognition task. The experimental results demonstrate that our deep features have the strong discerning ability. Thus, better performance and robustness can be obtained compared to state-of-the-art methods.

Keywords Shape classification · Deep learning · Non-rigid 3D shape · View-based 3D shape recognition · Bag-of-features

1 Introduction

The advancement in 3D acquisition technique leads to wider applications of 3D models in various fields, triggering a wave of research on shape analysis and classification methods. Especially, the successful application of deep learning technology in visual recognition in recent years has attracted more and more researchers to pay attention to learning-based methods for 3D shape analysis. Among them, view-based learning methods have made

✉ Li Han
hl_dlls@dl.cn

[1] School of Computer and Information Technology, Liaoning Normal University, Dalian, China

remarkable achievements in 3D shape classification and retrieval. They make full use of the successful deep learning model in the field of the 2D image, convert 3D shape into a series of 2D images, and then these 2D images are fed into a deep learning framework to extract the view features. The core work of the view-based learning method mainly includes two aspects. One is to generate valuable 2D views, such as projection views, panoramic views, geometry images, etc. [26, 31–33]. The other is to establish an effective learning mechanism to extract discriminative features from images, such as auto-encoder learning, deep belief networks (DBNs) learning, generative adversarial networks (GAN), convolutional neural network (CNN) learning, weighted CNN learning and multimodal CNN learning etc. [5, 7, 38, 42, 11, 13, 17].

However, view-based learning methods only consider the visual similarity, but ignore the intrinsic geometric information between 3D shapes. These 2D views with insufficient information directly affect the feature extraction and the accuracy of classification.

This issue prompts us to study how to construct informative images to obtain more distinctive view features. We propose a novel idea to extract discerning features by establishing a joint CNNs learning framework. This framework takes into account not only 2D projection views but also intrinsic geometry images. In particular, a score unit is introduced to automatically measure the contribution of each type of feature to recognition task. The weighted features are finally aggregated and refined in a standard CNN to serve as a final shape descriptor, which has stronger discrimination and achieves better performance in shape classification and retrieval.

The advantages of our work can be summarized in three aspects: 1) we introduce a discriminative BoF descriptor for capturing intrinsic structural properties of non-rigid 3D shape. It provides a uniform representation for the shapes with complex geometric variations. 2) We propose a joint CNNs learning model, taking both BoF images and 2D projection views as input, to extract geometric and spatial features. Through weighted fusion and refinement processing, the final feature enhances the discernment and boost the accuracy of shape classification and retrieval. 3) A joint objective function composed of cross-entropy loss function and Contrastive loss function is designed to optimize the learning process. Experiments show that our learning model is more efficient and achieves better performance than the advanced methods.

The paper is structured as follows: the related works are presented in section 2. In section 3, we introduce BoF image-based CNN learning and view-based CNN learning. The joint CNNs learning model is presented in section 4. Experimental results are analyzed and discussed in section 5. We conclude our work in section 6.

2 Related work

Excellent features will enhance the discrimination of shape representation and thus improve the classification accuracy. Therefore, most classification methods are directly related to feature extraction. According to the content of feature representation and the way of feature extraction, it can be divided into traditional geometric features and learning-based features.

Traditional geometric features, refer to low-level features, which are directly extracted from 3D model using different geometric analysis. The most popular feature is spectral feature, due to its discriminative power of isometric shape deformations, spectral feature has been extensively studied for non-rigid shape analysis, which is broadly classified into local spectral feature and global spectral feature. The local spectral features include GPS (the global point signature) [30], HKS

(Heat kernel Signature) [34], SI-HKS (scale-invariant HKS) [3], WKS (Wave Kernel Signature) [1] and SGWS (Spectral Graph Wave Signatures) [21] etc. On the other hand, global spectral features can be obtained from point signatures by integrating over the entire shape. For instance: Shape-DNA [29], ShapeGoogle [25], which presented significant performances in the applications of non-rigid shapes classification and shape retrieval. The major drawback of the low-level feature is that it cannot fully represent the 3D shapes with varied structure [20, 24]. Bag of features (BoF) [14, 16, 18, 22, 35] improved them by extracting a set of geometric features and encoding them as histograms distributed over the k-means clustering center to construct middle-level descriptor. Bronstein [3, 4] introduced a SS-BoF (spatially sensitive bag of features) that considered the spatial relationships between BoF features. Bu et al. [6] proposed the GA-BoF (geodesic-aware bags of features) by replacing the heat kernel in SS-BoF with a geodesic exponential kernel. Ye et al. [40] presented a global descriptor, which replaced geodesic distance with a reduced biharmonic distance matrix. As a global feature representation, BoF reveals well the intrinsic structural features and achieves good performance in the fields of non-rigid shape classification.

More recently, deep learning methods have achieved great success in visual recognition and natural language processing, making the study of learning-based features become a new trend for 3D shape analysis.

Learning-based features are high-level features automatically obtained from amount of data using a neural network. Earlier works take low-level features as input, such as Zernike moments [8], Geodesic moments (DeepGM [19]), Heat kernel signature (DeepSD [9], DeepShape [39]), Spectral graph wave codes (SGWC [21], Shape-aware BoF [10]) and 3D tensor (VoxNet [22]), and construct deep neural networks to implement high-level feature extraction. In recent years, the model-based learning methods directly learn features from the original representation of 3D data, and achieve effective results in shape segmentation and classification tasks. PointNet [27] proposed a novel network framework to learn deep features from point cloud data. PointNet++ [28] further solved the local structure issue by designing a hierarchical neural network, which applies PointNet recursively on a nested partitioning of the input point set. SturctureNet [41] presented a method of transforming geometry into hierarchical parts with part label [43] to implement shape segmentation and synthesis. Voxnet [23] employed CNN to learn features from voxelized shapes. FeaStNet [36] presented a deep neural network based on a novel graph convolution operator, which dynamically determines the association between filter weights and the nodes in a local graph neighborhood. However, building learning framework directly on the raw 3D data will make it sensitive to occlusion and noise, especially bringing huge storage and computational complexity.

Alternatively, view-based learning method encodes a 3D shape as a set of its rendered 2D views, and makes full use of the successful CNN framework in image recognition to learn view features, achieving satisfactory results in 3D recognition and classification [26, 33]. Shi et al. [26] transformed 3D shapes into panoramic views and build compact feature descriptor via using max pooling in the CNN. Sinha et al. [31] parameterized the 3D model to the spherical surface, and projected it onto the octahedron and then expanded into a 2D plane. A geometric image with intrinsic structure is finally obtained via the distribution of principal curvature or HKS in the plane. Bai et al. [2] employed GPU to accelerate the view features extraction. Guo et al. [12] presented a novel learning network monitored by triplet loss and classification loss. Su et al. [33] proposed a new CNN architecture to transform multiple 2D views of an object into a compact object descriptor. Bu et al. [5] proposed a multimodal learning architecture based on geometry image and projection images.

However, the deficiency of view-based learning method is that the transformation process changes the local and global structure of the 3D shape and reduces the discrimination of features. Therefore, how to effectively obtain informative images from 3D data and how to improve the discrimination of deep features are still challenging issues.

In this work, we design a joint CNNs learning model. It learns features from intrinsic BoF images and multiple 2D views in a dual-channel CNNs framework (see Fig. 1), they are weighted and refined to construct a high-level informative descriptor which provides stronger discrimination, so as to effectively identify and classify non-rigid 3D shapes.

Figure 1 illustrates the overall structure of our proposed method. First, we extract geometric features to generate a discriminative BoF image descriptor, which has better discernment against intra-class structural variations and noises. All discriminative BoF images of training data are fed into a CNN_1 framework, where the relationship between learning efficiency and CNN structure are deeply explored. Meanwhile, we establish a parallel CNN_2 to extract the extrinsic features from optimal 2D views. All the geometric and view features are weighted in a score unit. And the fused feature is finally refined in CNN_3 to boost the performance of 3D shape recognition and classification.

3 BoF image-based CNN and view-based CNN

BoF model generates codebook with low-level features through clustering method, and then represents the 3D model as an unordered set of codebook distribution frequency values. It effectively solves the problem of poor expression ability of low-level features, forms a kind of visual features between low-level and high-level semantic features. Compared with 2D images, BoF descriptors integrate a series of geometric features and effectively reveal the intrinsic structure of the model, thus achieving excellent results in the applications of non-rigid shape classification and retrieval [16, 18].

In our work, we provide a discriminative BoF image descriptor based on multiscale HKS and AGD features. The reason why these two geometric features are selected is that AGD reveals well the global topological structure of non-rigid 3D shape, while HKS captures rich local geometric information of 3D deformable shape. The combination of the two features has

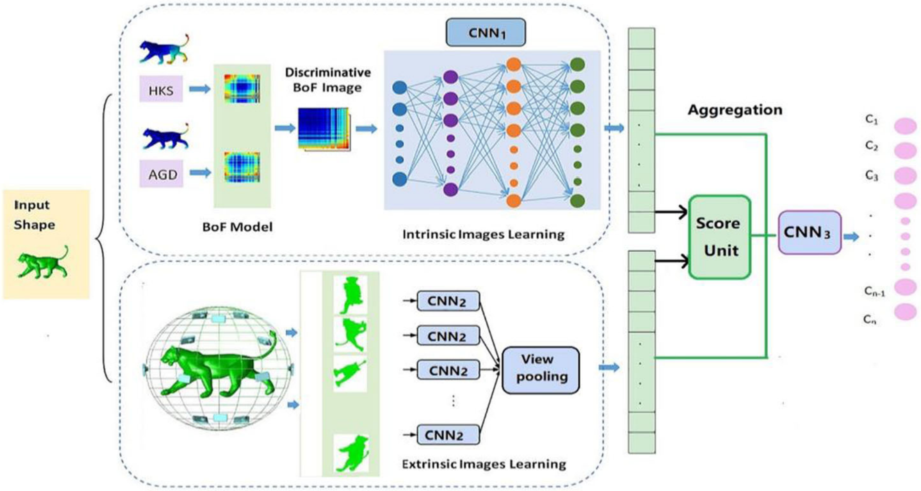


Fig. 1 Architecture of our deep learning model

shown excellent performance in the applications of shape retrieval and shape segmentation [6, 37, 39].

AGD Let $g(x_i, x_j)$ be the geodesic distance between two vertices x_i and x_j on a 3D mesh M, the average geodesic distance of $x_i$to all other vertices is defined as follows:

$$\mathrm{AGD}(x_i) = \frac{A_n(x_i)}{\sum_{x_{j\in X}} A_n(x_j)/N}$$
$$A_n(x_i) = \sqrt[n]{\frac{\sum_{x_{j\in X}} g(x_i, x_j)}{Area(\mathrm{M})}} \tag{1}$$

Where $Area$(M) denotes the total area of the mesh M.

HKS For each vertex x_i of the mesh M, its HKS is a p-dimensional feature vector as:

$$HKS(x_i) = \left(K_{t1}(x_i, x_i), \ldots, K_{tp}(x_i, x_i)\right) \tag{2}$$

The equation describes the conduction of heat on the surface with respect to time t, the heat equation $K_t(x_i, x_i)$ is called the heat kernel which is usually defined by the first $\widehat{m}$ eigenvalues λ_i and eigenvectors φ_i:

$$K_t(x_i, x_i) = \sum_{i=0}^{\widehat{m}} \exp(-\lambda_i t)\varphi_i^2(x_i) \tag{3}$$

Figure 2 shows the reconstruction results of the first 100, 150 and 200 eigenvectors of our optimal LBO [15]. It can be seen that the representation error decreases gradually with the increase of the number of basis functions, and reaches stability at 200. Therefore, in our work, we choose first 200 eigenvectors and divide time t into $p = 100$ time intervals in logarithmic scale interval$[t_{\min}, t_{\max}]$ ($t_{\min} = 4\ln 10/\lambda_{200}$, $t_{\max} = 4\ln 10/\lambda_2$) [34], so as to generate p-dimensional HKS description.

BoF descriptor 3D mesh model is defined as two geometric feature matrices $\mathbf{S}_{\mathbf{hks}}\left(s_1^h, s_2^h, \ldots, s_n^h\right)and\mathbf{S}_{\mathbf{agd}}\left(s_1^a, s_2^a, \ldots, s_n^a\right)$, where$s_i^h$, $s_i^a$are the local features of HKS and AGD at vertex x_i, and n is the number of mesh vertices.

We embed low-level local descriptors $\mathbf{S}_{\mathbf{hks}}$, $\mathbf{S}_{\mathbf{agd}}$ into vocabulary spaces. A codebook is constructed for each local descriptor by quantizing it into a certain number of codewords [3]. These codewords are usually defined by the k centers $\mathrm{V} = (v_1, v_2, .., v_k)$which are obtained via

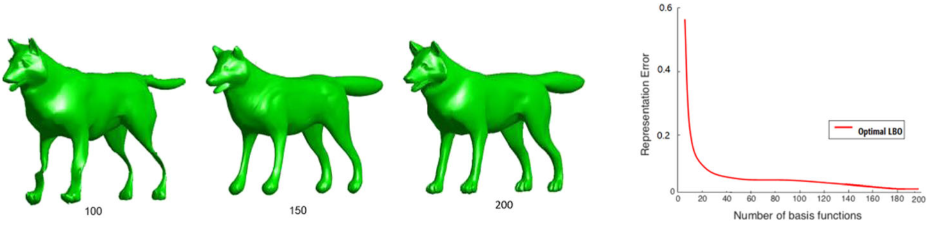


Fig. 2 Reconstruction of wolf model based on first 100,150 and 200 eigenvectors of our LBO

performing an unsupervised k-means clustering algorithm (k is also called feature dimension). Therefore, each feature$s_i$is mapped to a codeword in the codebook via the $k \times n$ cluster soft-assignment matrix $\mathrm{U} = (\mathrm{u}_1, \mathrm{u}_2, .., \mathrm{u}_n)$ whose elements are given by:

$$\mathrm{u}_i = \frac{\exp\left(-\alpha \| s_i - v_t \|_2^2\right)}{\sum_{t=1}^{k} \exp\left(-\alpha \| s_i - v_t \|_2^2\right)} \tag{4}$$

Where$\|\cdot\|_2^2$denotes the L_2-norm, and α is a smoothing parameter that controls the softness of the assignment. Each local descriptor is encoded by a k-dimensional code u_i and then constructs a $k \times n$ matrix **U,** which is denoted as BoF descriptor. The shape is eventually represented by the histogram of the codewords.

As we can see from Fig. 3 the shape descriptors of AGD and HKS of tiger model (Fig. 3a) are extracted and embedded into vocabulary space by using k-means clustering (Fig. 3b), respectively, the occurrence distribution of the codewords (k-cluster centers) are finally computed as shown in Fig. 3c.

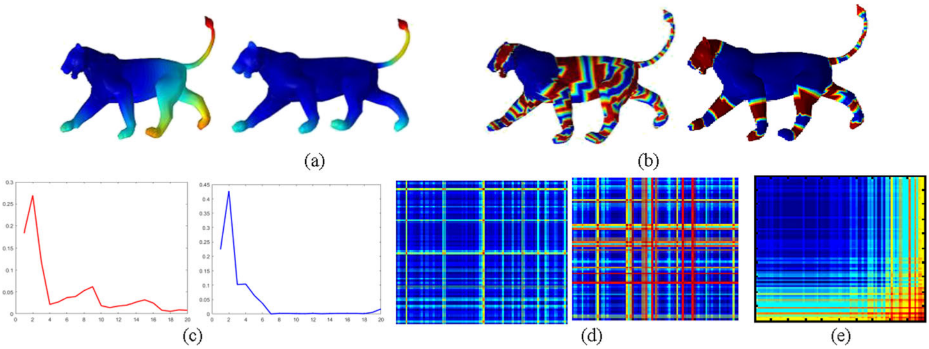
However, BoF only considers the occurrence of codewords and ignores the spatial relationship of them, which reduce its discernment ability. Inspired by SA-BoF [40], we construct a global BoF descriptor by using biharmonic distance matrix as follows:

$$\mathrm{g}_{ij} = \sum_{l=1}^{m} \frac{1}{\lambda_l^2} \left(\phi_l(i) - \phi_l(j)\right)^2; \quad \mathbf{F} = \mathbf{U}\mathbf{G}\mathbf{U}^{\mathrm{T}} \tag{5}$$

Where g_{ij} denotes the element of biharmonic distance matrix **G**, which is defined in terms of the eigenvalues λ_i and eigenfunctions φ_i of the LBO between any pair of mesh vertices $v_{\mathbf{i}}$ and $v_{\mathbf{j}}$, The resulting BoF descriptor **F** is a $k \times k$ matrix which represents global and spatial relationships of geometric features. Obviously, it is independent of the number of vertices of the model, and provides the same size representation for different shapes.

Figure 3d has shown the global BoF descriptors based on HKS and AGD features, where the color changes from blue to red as the value increases.

We compare the global BoFs of a set of shark models in SHREC 2015. However, as can be seen in Fig. 4, both HKS-BoFs and AGD-BoFs have quite different patterns, and they are not competent to recognize non-rigid shapes.


(a)

(b)

(c)

(d)

(e)

Fig. 3 **a** AGD (left) and HKS (right). **b** The quantization of features. **c** The distribution curves of AGD-BoF (left) and HKS-BoF (right). **d** The global AGD-BoF and HKS-BoF ($k = 55$). **e** Our discriminative BoF descriptor

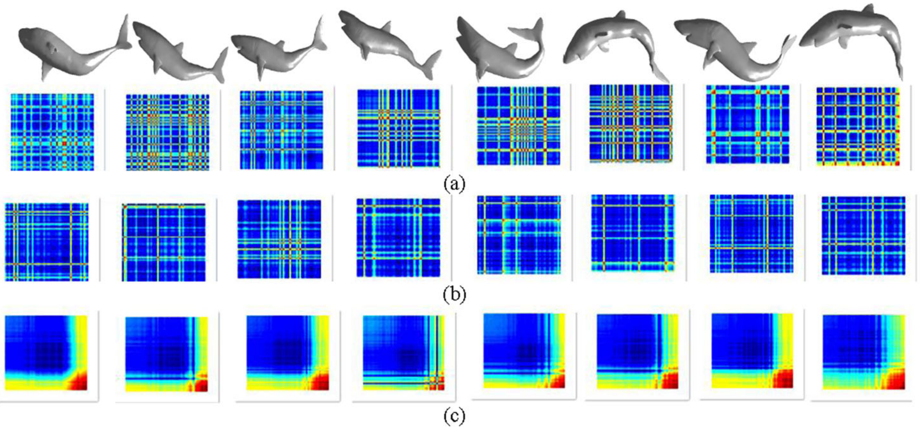


Fig. 4 The BoFs of different shark models. **a** Global HKS-BoF **b** Global AGD-BoF **c** Discriminative BoF descriptors

The key reason is that the codeword **V** generated by *k*-means clustering in vocabulary space is random, which causes the instability of constructed BoF descriptors. To improve the discernment of global BoF, we introduce a discriminative BoF descriptor (see Fig. 4c).

Discriminative BoF descriptor We concatenate two feature matrices $\mathbf{S_{hks}}$,$\mathbf{S_{agd}}$ to build a synthetic descriptor $\mathbf{S_{HA}}(s^{HA}(x_1), s^{HA}(x_2), \dots, s^{HA}(x_n))$. Each element $s^{HA}(x_i) = (S_{hks}(x_i), S_{agd}(x_i))$ is composed of *p-dimensional* HKS feature and *1-dimensional* AGD feature at vertex x_i.

And then we use *k*-means clustering to generate codebook, where the *k* cluster centers $C = (c_1, c_2, .., c_k)$ are sorted in ascending order according to their AGD features.

$$\widetilde{u}_i = \frac{\exp\left(-\alpha\|s^{HA}(x_i)-c_t\|_2^2\right)}{\sum_{t=1}^{k}\exp\left(-\alpha\|s^{HA}(x_i)-c_t\|_2^2\right)}; \quad \widetilde{F} = \widetilde{U}K\widetilde{U}^{T} \tag{6}$$

By computing the optimized visual features matrix$\widetilde{\mathbf{U}}$ *(k × n)* with sorted codewords **C** in Eq. 6, we can obtain a discriminative BoF descriptor$\widetilde{F}$, which reveals intrinsic structure of non-rigid 3D shape.

As can be seen from Fig. 5, the discriminative BoF descriptors not only reveal the intrinsic structural similarity of different shark models, but also highlight the obvious differences between different categories. Especially, it presents strong robustness to topological changes and incomplete (e.g. holes,cuts).

Our discriminative BoF descriptor $\widetilde{F}$ is invariant to the number and order of vertices on non-rigid shapes. It can provide unified matrix representation $\widetilde{F}$*(k × k)* for models with different resolutions.

Let $U^* = \widetilde{U}\cdot P$, $K* = P^{T}KP$,$U^{*T} = P^{T}\cdot\widetilde{U}^{T}$, where P is a permutation matrix that $P^{T}\cdot P = 1$, then we have:

$$F^* = U^*K^*U^{*T} = \widetilde{U}\cdot P\cdot P^{T}KP\cdot P^{T}\cdot\widetilde{U}^{T} = \widetilde{U}\cdot K\cdot\widetilde{U}^{T} = \widetilde{F} \tag{7}$$

Therefore, we can convert BoF matrix into BoF image to build uniform image representation of complex non-rigid model.

BoF image-based CNN learning Taking the discriminative BoF images as input, we design a BoF image-based CNN learning model to extract deep features.

Our BoF image-based CNN learning model use classic AlexNet structure which consists of five convolutional layers $C_{1...5}$ followed by three fully connected layers $FC_{1...3}$. Each convolutional layer has ReLU activation, and we use the max-pooling operator P() in our framework (see Table 1). The output features of the penultimate layer FC_2 (after ReLU non-linearity, 4096-dimensional) are taken as shape descriptors, and the last fully connected layer FC_3 is a Softmax classification layer.

In the practice, we take discriminative BoF image (224*224 pixels) as input and learn high-level intrinsic features by training hidden layers individually in an unsupervised manner. Figure 6 displays the step plots of deep features learned from BoF images of spider and human models.

BoF images effectively capture intrinsic geometric features, but may ignore the spatial correlation information of 3D shapes. Therefore, we further establish an optimized view-based CNN learning framework to provide extrinsic feature description.

View-based CNN In our view-based CNN framework, we follow the successful network architecture of MVCNN [33] (see Table 1). It replicates CNN branches for learning multiple views and aggregates them in a view-pooling layer to provide a compact and informative shape descriptor.

As we all know, more cameras will capture more detailed spatial information, but also cause more time consumption in deep learning. Instead of setting 12 cameras on the equator or 20 cameras on the icosahedral vertex in MVCNN [33], we present an improved theme, which need not assume that the vertical direction of the shape is consistent, but uses an appropriate number of cameras to achieve better projection views.

As shown in Fig. 7a, we distribute 12 cameras evenly cross a unit sphere enclosing the 3D shape. Six rendered views are from the cameras on the equator using the azimuth θ_i^z and elevationθ_i^e, which are set as$\theta_i^z = i \times 60^\circ (i = 0, 1, 2, ...5)$,$\theta_i^e = 0$. Other three rendered views

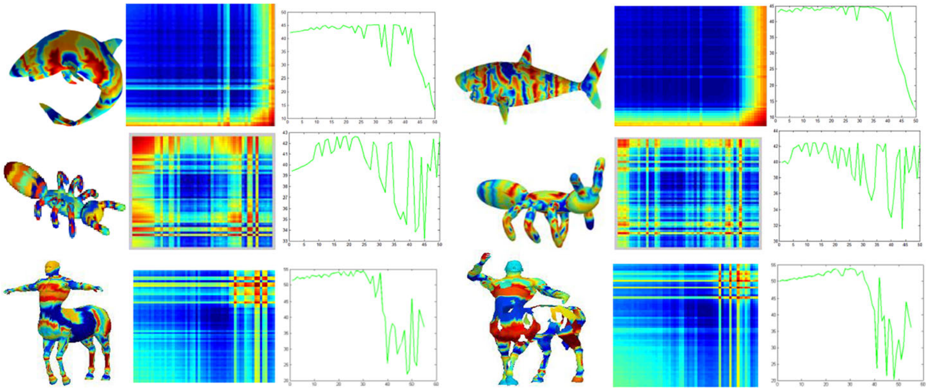


Fig. 5 Discriminative BoFs and the distribution curves of non-rigid shapes

are generated from$\theta_i^z = (i \times 90^\circ)\ +45^\circ(i = 0, 1, 2)$,$\theta_i^e = 45^\circ$ and the rest three rendered views are from three cameras setting$\theta_i^z = (i \times 90^\circ)\ +45^\circ(i = 0, 1, 2)$,$\theta_i^e = 135^\circ$. All cameras point towards the centroid of the mesh model.

The projection views pass through parallel CNN with shared parameter and generate an informative shape descriptor via an element-wise maximizing in a view-pooling unit.

The classification performance comparison of our 2D view-based CNN and MVCNN [33] on three classic non-rigid shape datasets is shown in Fig. 7b. It can be seen that our CNN network has achieved better results and the average accuracy is improved by nearly 0.65%.

4 Joint CNNs learning model

In this work, we establish a joint CNNs framework to generate shape descriptor with rich information, by weighting and fusing the intrinsic BoF image and extrinsic projection views, and further refine them to achieve deep features for recognition task.

Our goal is to obtain an informative shape descriptor F, which is extracted from geometric image learning F_{Bof} and spatial view learning F_{View}. Let J be the extraction process, and then the F can be defined as:

$$F = J\left(F_{Bof}, F_{View}\right) = \mathrm{W}_B \cdot F_{Bof} \oplus \mathrm{W}_V \cdot F_{View} \tag{8}$$

This work focuses on exploring the efficient combination of two features to improve the discrimination of shape descriptor, we introduce a weight learning unit to give each feature a score value, and further use Hadamard product (·) and concatenation (⊕) in the aggregation unit to achieve effective feature fusion.

As shown in Fig. 8, our joint CNNs learning model contains three pipelines: one is the BoF image-based learning in CNN_1, the other is the view-based learning, in which multiple views are passed through parallel CNN_2 and processed via an element-wise maximum operation. In the third pipeline, two kinds of deep features are evaluated in score unit, and the weighted and aggregated feature is refined by CNN_3 as discriminative shape descriptor for classification task. In our network, we design an objective function composed of Softmax loss function and Contrastive loss function to optimize the learning process.

Score unit Different from traditional pooling or concatenation, we pay attention to the contribution of each type feature to the recognition task. We propose a score unit based on a two-layer neural network (CNN_3). First, the first layer is used to align the input features. To be specific, the input features come from the second fully connected layer of CNN_1 and CNN_2. Then, the output of the first layer with different features are concatenated and sent to the second layer for learning the weight coefficient. The formula is as follows:

Table 1 Framework of BoF image-based CNN and our view-based CNN

BoF image-based CNN_1	C_1(96)P(3)-C_2(256)P(3)-C_3(384)-C_4(384)-C_5 (256)P(3)-FC_1(4096)-FC_2(4096)-FC_3(Softmax)
view-based CNN_2	C(64)P(2)-C(64)P(2)-C(128)P(2)-C(128) P(2)-C(256)P(2)-C(256)P(2)-C(256) P(2)-C(512)P(2)-C(512)P(2)-C(512)P(2)-C (512)P(2)-C(512)P(2)-C(512)P(2)-FC(4096)

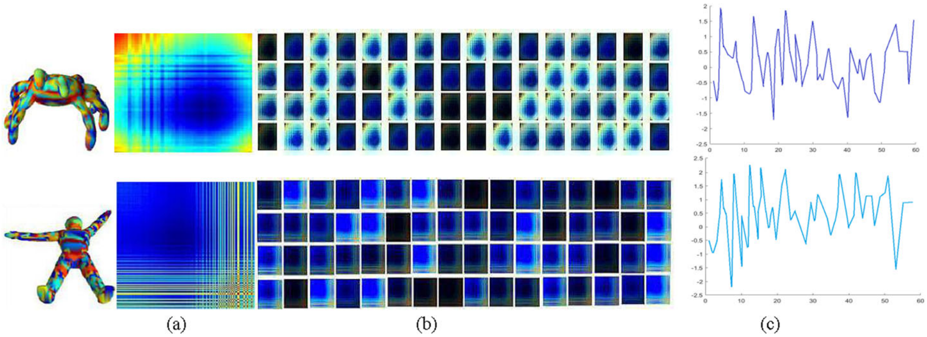


Fig. 6 BoF image-based CNN learning. **a** The BoF images. **b** The feature map after 1st layer. **c** The final deep features

$$\begin{bmatrix} \mathbf{h}_{bof} \\ \mathbf{h}_{view} \end{bmatrix} = \tanh\left(\mathbf{w}_n \begin{bmatrix} f_i^{bof} \\ f_i^{view} \end{bmatrix} + \mathbf{b}_n \right) \tag{9}$$

$$\mathbf{h_a} = \left[\mathbf{h}_{bof}, \mathbf{h}_{view}\right]; \mathbf{a} = \mathrm{softmax}(\mathbf{W_a h_a} + \mathbf{b_a}) \tag{10}$$

Where f_i^{bof}, f_i^{view} are the input features, and $\mathbf{W}_n$,$\mathbf{b}_n$ are the weight and bias terms of the first layer. Here tanh activation function is adopted to distinguish different input features, $\mathbf{h}_{bof}$,$\mathbf{h}_{\mathrm{view}}$ are the output features, those are aggregated and sent to the second layer to learn the weight coefficient **a**, where $\mathbf{W_a}$ and $\mathbf{b_a}$ are the weight and bias terms of the second layer, whose output is converted into weight coefficient through Softmax.

Then, in the aggregation unit, we use the Hadamard product of the input features and the weight coefficients to obtain the fused feature:

$$\mathrm{v}' = \mathbf{a}\cdot \begin{bmatrix} f_i^{bof} \\ f_i^{view} \end{bmatrix} \tag{11}$$

Finally, the aggregated feature is further refined in CNN$_3$ with AlexNet architecture to generate compact shape descriptor for a recognition task.

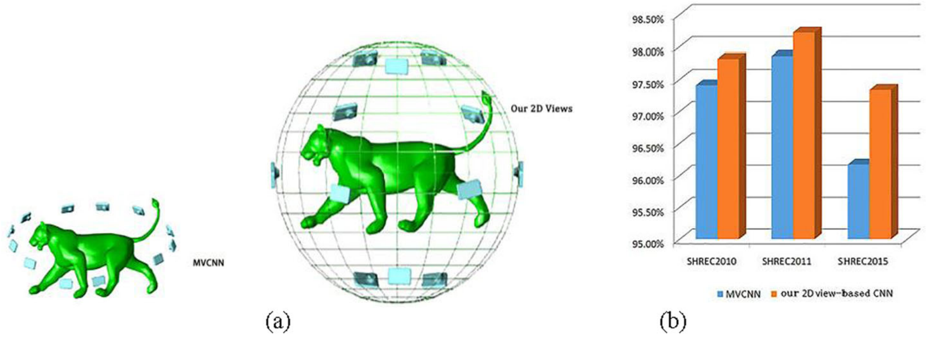


Fig. 7 The comparison of our 2D view-based CNN and MVCNN [33]. **a** Different camera settings. **b** Classification performances on three datasets

Optimization Our CNNs network is jointly supervised by cross-entropy loss and Contrastive loss, to achieve the optimization goal of maximum inter-class distance and minimum intra-class margin.

$$\min(L) = \min(L_s + L_c) \tag{12}$$

$$L_c = \sum_{i=1}^{m/2}\left(\alpha {\mathrm{D}_W}^2 + (1-\alpha)\max(T_r - \mathrm{D}_W, 0)^2\right) \tag{13}$$

where

$$D_W = \|Y_{2i-1} - Y_{2i}\|_2 = \sqrt{\sum_{l=1}^{c}\left(y_{2i-1}^{(l)} - y_{2i}^{(l)}\right)^2} \quad (i = 1, 2, \ldots, m/2)\, s.t. y_i^{(l)} = e^{w_l^T X_i + b_l}; \tag{14}$$

Where L_s is the cross-entropy loss function and L_c is the Contrastive loss function. D_W represents the L_2 normalization of a paired shape features Y_{2i-1}andY_{2i}, and αdenotes the similarity between them, if they are matched, α is set to 1, otherwise set to 0. T_r represents the distance between shape features of different categories, we only consider Euclidean distance between 0 and T_r for dissimilar features.

The cross-entropy loss function L_s improves the feature separability by making the distance of inter-class farther; the Contrastive loss function L_c, on the other hand, expresses the matching degree of paired samples that improves the cohesion of features by narrowing the intra-class distance. To fit data into Eq. 13, we input the shapes into our CNN framework in pairs. First, we calculate the Contrastive loss and Softmax loss through forward propagation, and then update the parameters through backward propagation with stochastic gradient descent as follows:

$$\begin{aligned}
&\frac{\partial L}{\partial W} = \frac{\partial L_s}{\partial W} + \frac{\partial L_c}{\partial W}\\
&\frac{\partial L_s}{\partial W} = -\sum_{i=1}^{m} P\left(\mathrm{I} - \frac{Y_i}{\sum_{j=1}^{m} Y_i}\right) X_i\\
&\frac{\partial L_c}{\partial W} = 2\sum_{i=1}^{m/2}\left(1 + (1+\alpha)\frac{T_r}{D_{iw}}\right)(Y_{2i-1} - Y_{2i})^T X_i
\end{aligned} \tag{15}$$

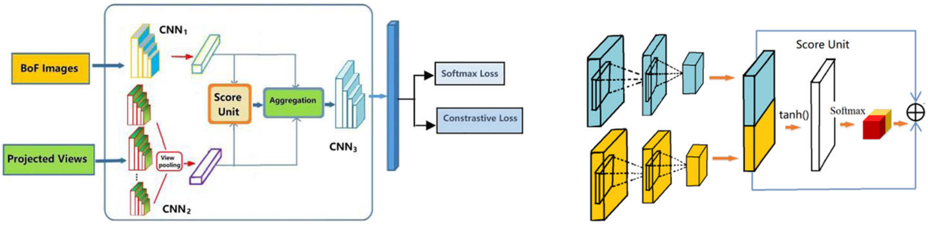


Fig. 8 The joint CNNs learning model and Score Unit

The algorithm of our joint CNNs learning model is summarized as follows:

```
Algorithm1   Joint CNNs –based 3D shape classifier
Input: Non-rigid dataset $D=\{M_1,...,M_m\}$ of $m$ shapes
Output: deep feature vector $\hat{y}$ containing predicted class label for each shape $M_i$
For $i=1$ to $m$ do
 1:  BoF image generation
   1.1 compute the $p \times n$ HKS matrix $S_i^h$ for 3D shape $M_i$, where $n$ is the number of vertices and $p$ is the dimension of HKS.
   1.2 compute the $1 \times n$ AGD matrix $S_i^a$ for shape $M_i$
   1.3 construct the $(p+1) \times n$ conjunction matrix $S_i^{HA}$
   1.4 compute the sorted codewords $\mathrm{V}_\mathrm{i}^*$ and $k \times n$ BoF matrix $\mathrm{U}_\mathrm{i}^*$, where $k$ is the number of $k$-means cluster center
   1.5 compute the biharmonic distance matrix $\mathrm{G}_i$ and generate $k \times k$ discriminative BoF matrix $\mathrm{F}_i$.
   1.6 Convert discriminative BoF matrices $\mathrm{F}_i$ to BoF images descriptor $\mathrm{B_i}$
 2:  2D views generation
   For $i$=1 to 12 do
       $V_i = Views\,(\theta_i^z,\ \theta_i^e)$ ;
   Endfor
Endfor
 3:  Build data set $\mathrm{Y}=\{\mathrm{B},\mathrm{V}\}$
 4: Input training data $y \in \mathrm{Y}$ into CNN$_1$, CNN$_2$.
    Initialize parameters, learning rate, maximum number of iterations T

 5: While t< T do
      a)  BoF image-based CNN$_1$ learning intrinsic feature vector $y_b$;
      b)  View-based CNN$_1$ Learning extrinsic feature vector $y_v$;
      c)  Using Score Unit and Aggregation Unit to generate fused feature $\hat{y}$;
      d)  Generate informative descriptor $\hat{\hat{y}}$ in CNN$_3$;
      a)  Predicted class labels $l_y$ in Softmax classifier
      e)  Calculate the Contrastive loss and Softmax loss through forward propagation;
      b)  Update parameters in three CNNs through back propagation with Formula 15.
      c)  t++;
End while
```

5 Experimental results

We conduct extensive experiments to test the performance of our proposed approach on classic non-rigid shape benchmarks: SHREC2010, SHREC2011, SHREC2015 and SHREC2016.

SHREC10 contains 200 non-rigid models with different postures from 10 classes. SHREC-2011 is a dataset of 3D shapes consisting of 600 watertight mesh models, which are obtained from transforming 30 original models. The SHREC-2015 is a dataset of 1200 watertight mesh models from 50 classes [34]. We also test our learning model on the incomplete non-rigid shapes using SHREC2016, which includes 8 categories with 400 incomplete models. In the experiment, we train our CNNs model by randomly selecting 50% models, and extract 1000 times to ensure the training accuracy.

Our CNNs model is pre-trained on ImageNet from 1 k categories and then fine-tuned on all the 3D shapes in the training set. In the training phase, we organize the training data into pairs to fit our joint loss function, generate 12 rendered views and 1 discriminative BoF image for each data. All rendered views of each pair of data are fed to the CNN$_2$ framework and then generate deep extrinsic feature through max pooling. Meanwhile, each pair of discriminative BoF images is sent to CNN$_1$ to extract deep intrinsic features. And after weighted fusion, the deep feature is refined in CNN$_3$ and input a Softmax classifier.

Our network parameters are optimized using the joint objective function and stochastic gradient descent. We set learning rate of 0.03 and learning rate decay of 0.95, dropout probability of 0.5, regularization weight of 5×10^{-4}.

In this section, we first discuss the robustness of our discriminative BoF image descriptor against different resolution and Gaussian noises, and then we study the contribution of each part of CNN framework to the recognition task. Finally, the performance of our joint CNNs learning model on non-rigid shape datasets is analyzed by comparing to state-of-the-art methods.

5.1 Performance of discriminative BoF image descriptor

In our work, we combine HKS and AGD features to generate BoF image descriptor, which provides stronger discriminative representation. Fig. 9a shows the distribution curves of our BoFs of two David models. Compared with those of HKS-BoF and AGD-BoF in Fig. 9b, c, our BoF descriptor reveals better the similarity. Moreover, it provides a more stable and excellent classification performance over three datasets.

In Fig. 9d, we compare the classification performance of different BoF descriptors in three non-rigid databases. As we can see that AGD-BoF is sensitive to the topological changes and unable to recognize complex models, such as octopus (category no.5 in SHREC-2010, category no.19 in SHREC-2011 and category no.20 in SHREC-2015). HKS-BoF has a good effect on most models but it is difficult to classify the unstructured models, such as snakes, sharks (category no. 7 in SHREC-2010 and category no. 24 in SHREC-2011 and SHREC-2015);

Our BoF feature combines AGD with HKS features, which effectively enhance the discernment of features. The average classification accuracy based on our discriminative BoF image descriptor is 10.38% higher than that of AGD-BoF and 7.73% higher than that of HKS-BoF (see Table 2).

We further evaluate the robustness of BoF image descriptor to noise and multi-resolutions. Taking the cat model as an example, the distribution curves and classification performances under different resolutions (2 K, 1 K) and Gaussian noises ($\delta = 0.05$, $\delta = 0.1$) and topological variations (holes and cuts) are compared. As shown in Fig. 10a, b the BoF distribution curves of test groups lightly change which shows the strong stability of BoF descriptors.

Meanwhile, we test the discrimination ability of our BoF descriptor on non-rigid shape datasets SHREC2010 and SHREC2011 (Fig. 10c). The red curve describes the classification accuracy of our BoF image-based CNN learning with clean models, while the other two curves represent the classification accuracy with noisy models. The average precision is 93.8% and 95.5%, even in the noisy models, the impact on performance is very small, within the range of 0.25%.

The BoF feature dimension k is defined by k-clustering. Although larger k can reveal geometric characteristics of 3D shape in more detail (Fig. 11a), it will also generate larger matrix, thus increasing the computation burden. We observe that the accuracy improves when k ranges from 50 to 200, after which the curve changes smaller (Fig. 11b). The performance remains stable when it reaches 1000 epochs (Fig. 11c). Therefore, we use $k = 200$ and $T = 1000$ for the remaining experiments.

Our joint CNNs learning model effectively combines BoF image-based CNN and view-based CNN to generate informative and compact shape descriptors, thus improving the

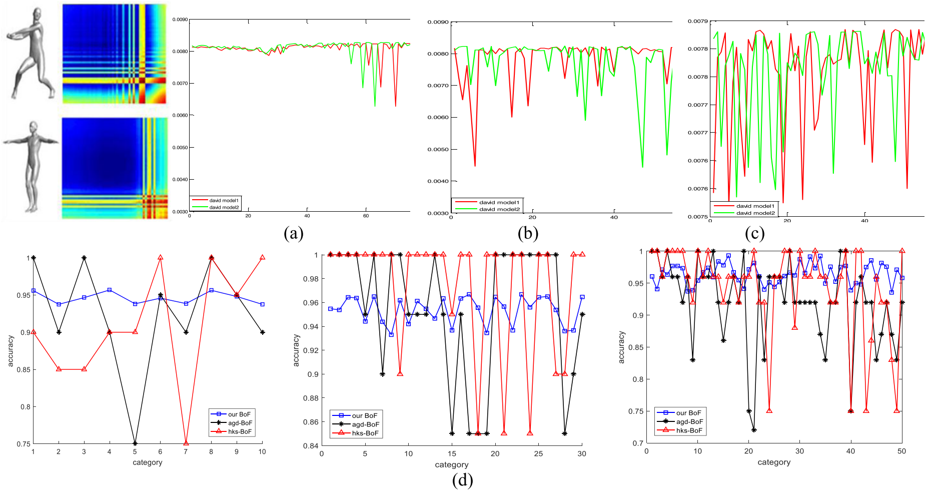


Fig. 9 **a** The distribution curves of discriminative BoF image descriptor. **b** of HKS-BoF descriptor. **c** of AGD-BoF descriptor. **d** The classification accuracy on SHREC2010 (left), SHREC 2011(middle) and SHREC 2015 (right)

performance of shape classification. We compare it with each pipeline running independently.

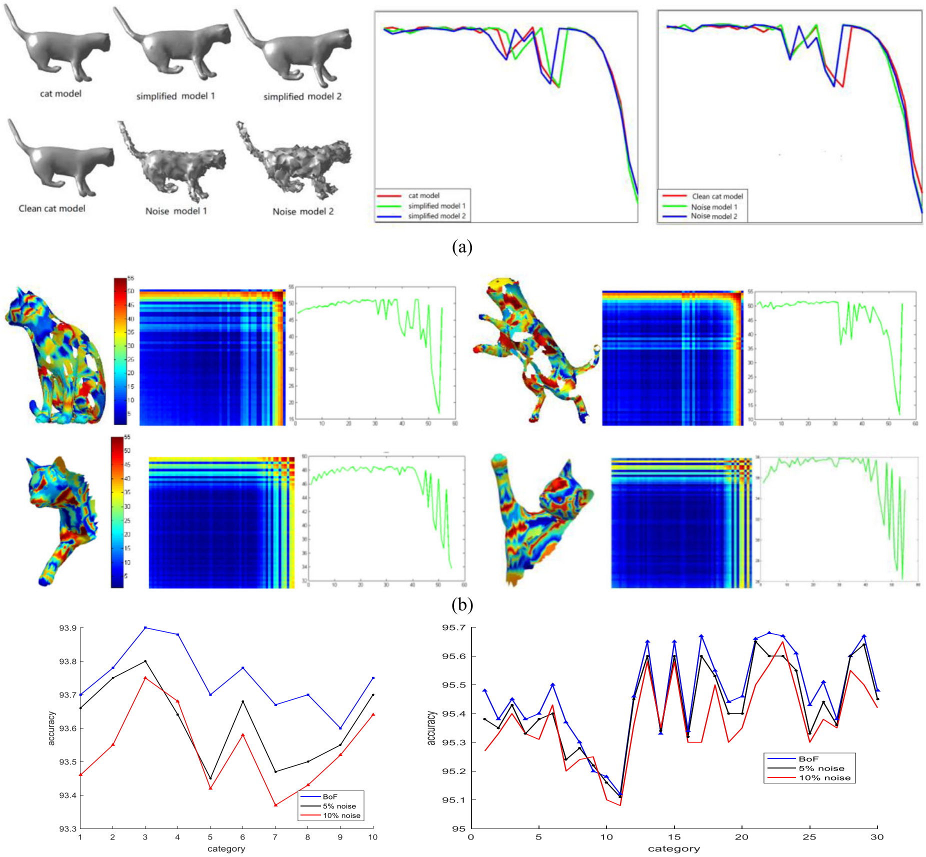
Table 3 shows the classification performance on four databases. It can be seen that the deep feature extracted from our joint CNNs learning has the best discrimination ability, while the BoF image-based CNN learning has lowest performance compared with the view-based CNN learning. Although our joint CNNs consumes more time in training process, it greatly improves the convergence due to the use of joint loss function, so a better trade-off between the consuming time and the accuracy is achieved.

Figure 12a, b, c shows the comparison of deep features extracted from the BoF images, the 2D views and the informative images, respectively. We can see that the deep feature curves with our informative images present the best matching results between two crabs and two bears (Fig. 12a). For these two kinds of models, the accuracy based on BoF image learning is 96.53%, and that of 2D views learning is 97.06%, while the accuracy of our joint CNNs model reaches 98.14%.

Fig. 12d presents the best classification performance of our joint CNNs model on SHREC2015. Fig. 12e shows the training process based on three different CNN frameworks. We can see that our joint CNNs model can obtain more stable and higher training accuracy in a certain number of iteration (200 times) comparing with the other two CNN models. Figure 12f shows the comparison of our objective function with other two loss functions by taking the SHREC2010 dataset as an training example, we can see that our joint loss function curve is more stable and presents higher convergence.

Table 2 The classification accuracy (%) of different BoFs on three datasets

Dataset	SHREC2010	SHREC2011	SHREC2015
AGD-BoF	83.26	84.53	85.78
HKS-BoF	87.65	88.10	90.36
Discriminative BoF	93.75	95.50	95.47



(a)

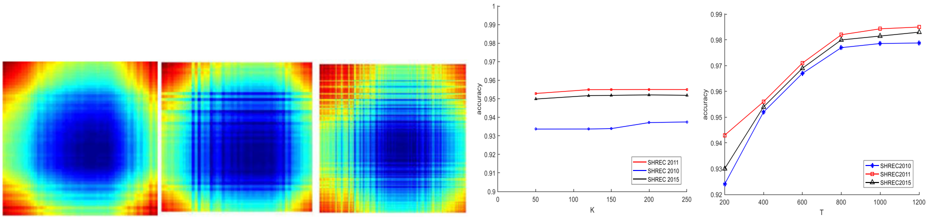
(b)

(c)

Fig. 10 **a** The comparison of distribution curves under different simplified and noised models. **b** The comparison of different incomplete shapes on SHREC2016. **c** And the classification accuracy on SHREC 2010 and SHREC 2011

5.2 Comparison to state-of-the-art methods

Since our method combines BoF image-based CNN learning with 2D view-based CNN learning, we are interested in knowing how each learning framework improves the classification performance. Therefore, we will discuss our learning model by comparing it with BoF



(a)

(b)

(c)

Fig. 11 **a** The BoF images of crab models under different feature dimension k ($k = 50{,}100,\ 200$). **b**, **c** Classification performance under different feature dimension k and training epoch T

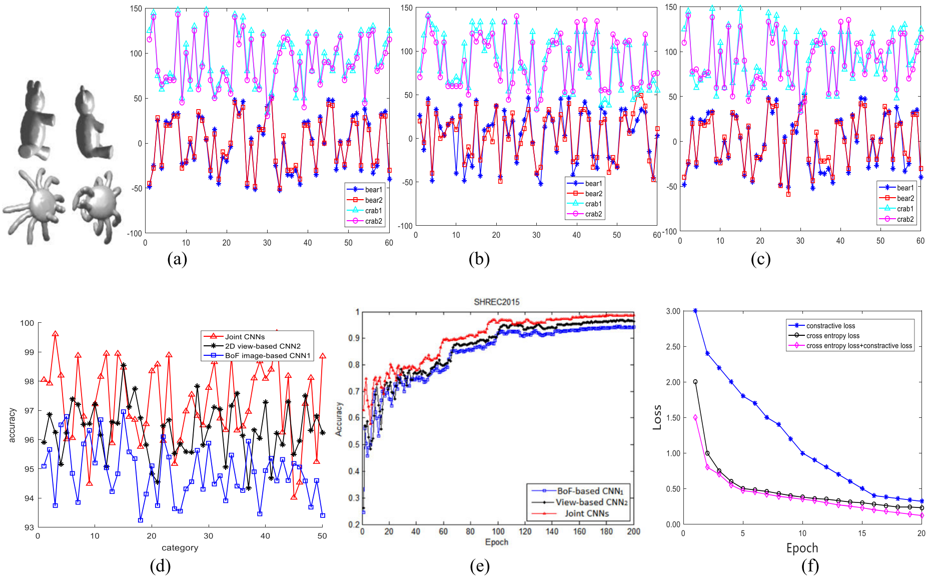
Table 3 The performance of our joint CNNs framework

Architecture	Accuracy (%)				Training Time (m) cross-entropy loss/ Joint loss function
	SHREC10	SHREC 11	SHREC 15	SHREC 16	
BoF image-based CNN$_1$	93.75	95.50	95.47	89.4	92 / 72
2D view-based CNN$_2$	97.61	98.04	96.49	87.6	150 / 123
Joint CNNs	**98.38**	**98.53**	**97.82**	**91.3**	**196 / 178**

methods [3, 4, 6, 21, 40], feature learning methods [9, 19, 36, 39] and view-based learning method [33, 44].

ShapeGoogle [4] is one of the representative works, it embeds HKS into a vocabulary space, and extracts a frequency histogram of geometric words as BoF descriptor, which is robust to structural variations and has achieved good results in non-rigid shape retrieval. The GA-BoF [6] adopts scale invariant heat kernel and AGD as low-level descriptor, and constructs global BoF using geodesic exponential kernel instead of heat kernel to avoid the influence from time scale and size, the deep features are learned in a two-layer deep belief networks (DBN). While SA-BoF [40] employs spectral graph wavelets and learned high-level features in a deep auto-encoder. SGWC BoF [21] uses spectral graph wavelet signatures to construct middle-level features and implement the classification task by multiclass SVM.

Our BoF image-based CNN learning framework uses multiscale HKS and AGD to construct discriminative BoF image and learns intrinsic features inside a CNN framework optimized by a joint objective function. The comparison with typical BoF methods is shown in Table 3. We can see that our performance has significantly improved by an average of 3% to10% compared to GA-BoF and SA-BoF, and it is slightly lower than SGWC [21].



(a) (b) (c)

(d) (e) (f)

Fig. 12 The learned deep features of deformable shapes **a** from our informative images. **b** from BoF images. **c** from multiple 2D views (**d**) (**e**) the comparison of the classification performance and the training process based on different CNNs frameworks on SHREC2015. **f** the comparison of loss curves during learning

Table 4 The comparison of performances (%) with deep learning methods

Dataset	SHREC2010	SHREC2011	SHREC2015	mAP
ShapeGoogle [4]	90.96	92.89	83.16	89.15
GA-BoF [6]	86.02	93.20	72.93	84.11
SA-BoF [40]	91.83	98.00	84.27	91.37
SGWC-BoF [21]	95.66	97.66	92.54	95.28
BoF image-based CNN	**93.75**	**95.50**	**95.47**	**94.91**
DeepShape [39]	95.5	96.53	92.87	94.96
DeepGM [19]	96.33	97.89	93.03	95.75
FeaStNet [36]	97.84	98.12	96.35	97.43
FVCNN [44]	97.66	98.05	96.53	97.41
MVCNN [33]	97.42	97.86	96.17	97.15
Our 2Dview-based CNN	**97.61**	**98.04**	**96.49**	**97.38**
Our joint CNNs	**98.38**	**98.53**	**97.82**	**98.24**

Considering BoF images may ignore the spatial correlation of original 3D shape, we construct informative images by taking BoF images and 2D views as input and learn in a joint CNNs learning framework. The comparison of our joint CNNs learning model with the feature learning methods (DeepGM [19], DeepShape [39], FeaStNet [36],FVCNN [44]) and view-based learning method (MVCNN [33]) is shown in Table 4.

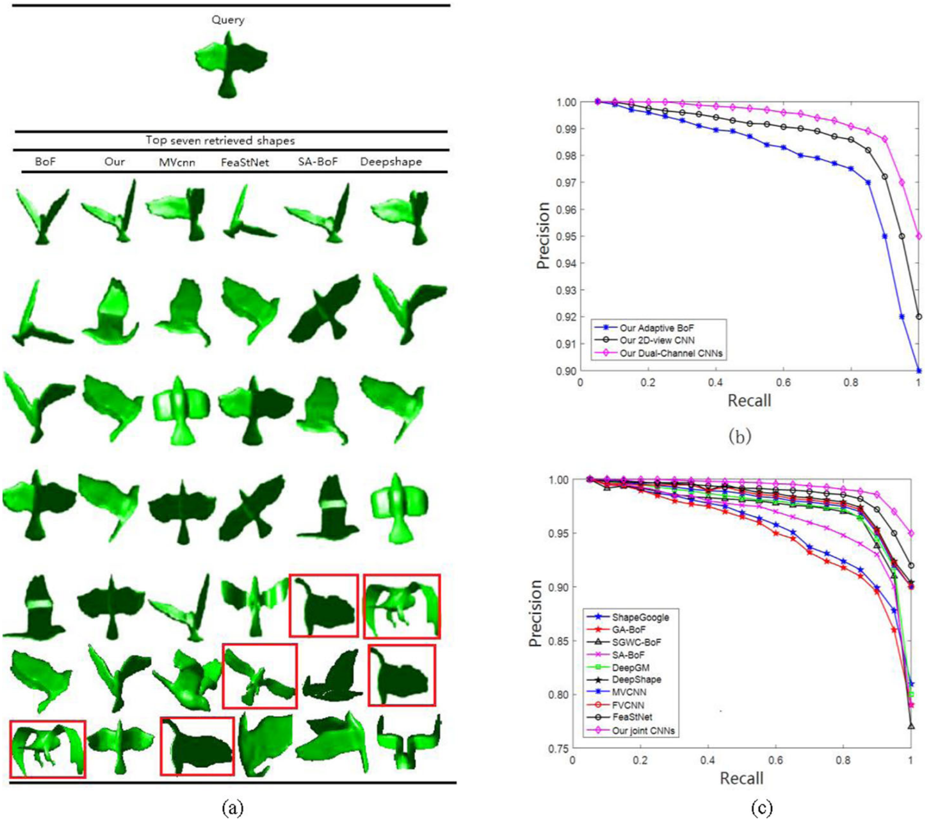


Fig. 13 **a** Top 7 retrieved shapes using baseline methods on SHREC2015. **b** The comparison of P-R curves by using different features in our learning model. **c** The comparison of state-of-the-art methods

DeepShape [39] takes heat kernel shape descriptor (HeatSD) as input and learns the deep features by a many-to-one encoder neural network. While DeepGM [19] learns deep features in an auto-encoder learning framework by taking geodesic moments as input. FeaStNet [36] uses a dynamic graph convolution operator in a local neighborhood and learns local shape properties with the raw 3D shape coordinates as input instead of 3D shape descriptors. FVCNN [44] proposes a feature coding module, the features of the 3D shape are coded as the pixel values on the view plane from which the views are generated and then learned by a fusion module based on CNNs. MVCNN [33] converts 3D shape into multiple projection 2D views and adopts max pooling to aggregate the multi-view features extracted by VGG-M.

It can be seen that the integration with view features and BoF image features effectively promotes the classification performance. It achieves 2.5%, 3.3% and 0.81% higher than DeepGM [19], DeepShape [39] and FeaStNet [36] on average, and 1.09% and 0.83% higher than MVCNN [33] and FVCNN [44],respectively.

The retrieval results and P-R curves using different methods are further illustrated in Fig. 13. Taking a pigeon in SHREC2015 as the query sample, we compare the retrieved results of top 7 models with several baseline methods. As can be seen in Fig. 13a our method is able to correctly retrieve relevant pigeon shapes, while other methods get eagle models more than once by mistake (the red box shows the wrong eagle model). The comparisons of the precision-recall graphs of state-of-the-art methods is shown in Fig. 13c. As we can see that GA-BoF has the lowest precision compared with the typical BoF methods and feature learning methods, DeepGM and DeepShape present comparable performance, while our method provides the best retrieval accuracy, which is 0.78%, 0.83% and 0.86% higher than FeaStNet, FVCNN and MVCNN, respectively.

We also study the impact of each pipeline in our framework on shape retrieval task (Fig. 13b, it further verifies that the feature extracted from the joint CNNs learning presents better discernment and boosts the retrieval performance.

6 Conclusions and future work

In this paper, we proposed a CNNs framework to deal with non-rigid 3D shapes classification task. We aim to efficiently boost the classification performance from two perspectives of improving informative image representation and CNN learning mechanism.

In the first stage, the low-level 3D shape descriptors based on HKS, AGD are extracted and used to construct discriminative BoF images. We adopt a standard CNN framework to extract intrinsic deep features from BoF images. Meanwhile, we learn extrinsic spatial features from projected 2D views within a parallel view-based CNN model. Then, a score unit is designed to automatically evaluate different deep features. Finally, the weighted and aggregated feature is refined to perform 3D shape classification. All the training processes are monitored by a joint objective loss function which effectively improves the convergence and the accuracy.

We showed that our deep features are robust and stable, which achieve significantly better performance than state-of-the-art methods.

However, our informative images capture global features rather than semantic structural features; it is thus still difficult to implement partial recognition and structure understanding tasks. Therefore, it is necessary to research a novel deep learning model that can directly extract hierarchical structural features and build symmetry-aware and structure-aware learning mechanism. In addition, our learning model has a constraint on the topological connectivity of data. We will further extend our work to learn localized and structural features in large-scale point cloud data.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments. The research presented in this paper is supported by a grant from NSFC (61702246), grants from research project of Liaoning province (2019lsktyb-084, 2020JH4/10100045) and a fund of Dalian Science and Technology (2019J12GX038).

References

1. Aubry M, Schlickewei U, Cremers D (2011) The wave kernel signature: A quantum mechanical approach to shape analysis. In: Proc. Computational Methods for the Innovative Design Electrical Devices, pp 1626–1633
2. Bai S, Bai X, Zhou Z, Zhang Z, Latechi LJ (2016) GIFT: A real-time and scalable 3D shape search engine. In: Proc. CVPR, pp. 5023–5032
3. Bronstein M. Kokkinos, I. (2010) Scale-invariant heat kernel signatures for non-rigid shape recognition. In: Proceedings of the CVPR, pp 1704–1711
4. Bronstein A, Bronstein M, Guibas LJ, Ovsjanikov M (2011) Shape Google: geometric words and expressions for invariant shape retrieval. ACM Trans Graph 30(1):1–22
5. Bu S, Cheng S, Liu Z, Han J (2014) Multimodal feature fusion for 3d shape recognition and retrieval. IEEE Multimed 21(4):38–46
6. Bu S, Liu Z, Han J, Wu J, Ji R (2014) Learning high-level feature by deep belief networks for 3-D model retrieval and recognition. IEEE Trans Multimed 24(16):2154–2167
7. Cai W, Wei Z (2020) PiiGAN: generative adversarial networks for pluralistic image. IEEE Access 8:48451–48463
8. Chen D-Y, Tian X-P, Shen Y-T, Ouhyoung M (2003) On visual similarity based 3d model retrieval. Comput Graph Forum 22:223–232. Wiley Online Library
9. Fang Y, Xie J, Dai G, Wang M, Fan Z, Xu T, Wang E (2015) 3D deep shape descriptor. In: Proc. of the 28th IEEE Conf. On CVPR, pp.2319–2328
10. Ghodrati H, Hamza AB (2016) Deep shape-aware descriptor for nonrigid 3D object retrieval. Int J Multimed Inf Retr 3:1–14
11. Ghodrati H, Hamza AB (2017) Nonrigid 3D shape retrieval using deep auto-encoders. Appl Intell 47:44–61
12. Guo H, Wang J, Gao Y, Li J, Lu H (2015) Graph-based characteristic view set extraction and matching for 3D model retrieval. Inf Sci 320:429–442
13. Guo H, Wang J, Gao Y et al (2016) Multi-view 3d object retrieval with deep embedding network. IEEE Trans Image Process 25(12):5526–5537
14. Han Z, Liu Z, Vong CM, Liu YS, Bu S, Han J, Chen CLP (2017) BoSCC: bag of spatial context correlations for spatially enhanced 3Dshape representation. IEEE Trans Image Process 26(8):3707–3720
15. Han L, Liu S, Yu B, Xu S (2020) Orientation-preserving spectral correspondence for 3D shape analysis. J Imaging Sci Technol 64(1):1–13
16. Laga H, Schreck T, Ferreira A, et al. (2011) Bag of words and local spectral descriptor for 3D partial shape retrieval. Proc. of the 4thEurographics Conf. on 3D Object Retrieval, Llandudno, April 10, 41–48
17. Leng B, Cheng Z, Zhou XC (2018) Learning discriminative 3D shape representations by view discerning networks. IEEE Trans Vis Comput Graph
18. Litman R, Bronstein A, Bronstein M, Castellani U (2014) Supervised learning of bag-of-features shape descriptors using sparse coding. Comput Graph Forum 33(5):127–136
19. Luciano L, Hamza AB (2017) Deep learning with geodesic moments for 3D shape classification. Pattern Recog Lett
20. Masoumi M, Hamza AB (2017) Spectral shape classification: a deep learning approach. J Vis Commun Image Represent 43:198–211
21. Masoumi M, Li C, Hamza AB (2016) A spectral graph wavelet approach for nonrigid 3D shape retrieval. Pattern Recogn Lett 83:339–348
22. Matsuda T, Furuya T, Ohbuchi R (2015) Lightweight Binary Voxel Shape Features for 3D Data Matching and Retrieval. In: Multimedia BigData, pp. 100–107
23. Maturana D, Scherer S (2015) VoxNet: a 3D convolutional neural network for real-Time object recognition. In: Proc. International Conference on Intelligent Robots & Systems (IROS)

24. Mohamed W, Hamza AB (2016) Deformable 3D shape retrieval using a spectral geometric descriptor. Appl Intell 45(2):2213–2229
25. Ovsjanikov M, Bronstein A, Bronstein, M., Guibas LJ (2009) Shape Google: A computer vision approach to isometry invariant shape retrieval. In: Proc. 2009 IEEE 12th Int Conf Comput Vis Workshops, pp. 320–327
26. Papadakis P, Pratikakis I, Theoharis T, Perantonis S (2010) Panorama: a 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. J Comput Vis 89(2):177–192
27. Qi CR, Su H, Mo K, Guibas LJ (2017) Pointnet: deep learning on point sets for 3d classification and segmentation. In Proc. CVPR
28. Qi CR, Yi L, Su H, Guibas LJ (2018) PointNet++: deep hierarchical feature learning on point sets in a metric space. In Proc. CVPR
29. Reuter M, Wolter F, Peinecke N (2006) Laplace-Beltrami spectra as ‘Shape-DNA’ of surfaces and solids. Comput Aided Des 38(4):342–366
30. Rustamov R (n.d.) Laplace-Beltrami eigenfunctions for deformation invariant shape representation. In: Proc. Symp. Geometry Processing, pp 225–233.
31. Shi BG, Bai S, Zhou Z et al (2015) DeepPano: deep panoramic representation for 3D shape recognition. IEEE Signal Process Lett 22(12):2339–2343
32. Sinha A, Bai J, Ramani K (2016) Deep learning 3D shape surfaces using geometry images. In: Proceedings of the European Conference on Computer Vision. Amsterdam, 223–240.
33. Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multi-view convolutional neural networks for 3d shape recognition. In: Proc.ICCV
34. Sun J, Ovsjanikov M, Guibas L (2009) A concise and provably informative multi-scale signature based on heat diffusion. Comput Graph Forum 28(5):1383–1392
35. Toldo R, Castellani U, Fusiello A (2009) Visual vocabulary signature for 3D object retrieval and partial matching. In: Proc. 2nd Eurograph Conf 3D Object Retrieval, pp. 21–28
36. Verma N, Boyer E, Verbee J (2018) FeaStNet: Feature-Steered graph convolutions for 3D shape analysis. In: Proc. CVPR, pp. 2598–2606
37. Wan L, Zou C, Zhang H (2017) Full and partial shape similarity through sparse descriptor reconstruction. Vis Comput 33(12):1497–1509
38. Wang Z, Zou C, Cai W (2020) Small sample classification of Hyperspectral remote sensing images based on sequential joint Deeping learning model. IEEE Access 8:71353–71363. https://doi.org/10.1109/ACCESS.2020.2986267
39. Xie J, Fang Y, Zhu F (2016) Deep Shape: deep Learned shape descriptor for 3D shape matching and retrieval. Comput Vis Pattern Recog
40. Ye J, Yu Y (2015) A fast modal space transform for robust non rigid shape retrieval. Vis Comput 32(5):553–568
41. Yi L, Zhao W, Wang H, Sung M, Guibas L (2019) StructureNet: hierarchical graph networks for 3D shape generation. In Proc. Siggraph Asia
42. You H, Tian S, Yu L, Lv Y (2020) Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Trans Geosci Remote Sens 58(2):1281–1293
43. Yu F, Liu K, Zhang Y, Zhu C, Xu K (2019) PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proc. CVPR
44. Zhou Y, Zeng F, Qian J, Xiang Y, Feng Z (2019) FVCNN: Fusion View Convolutional Neural Networks for Non-rigid 3D Shape Classification and Retrieval, International Conference on Image and Graphics, 566–581, Beijing, P.R. China, 8.23–8.25

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.