Deep Spatiality: Unsupervised Learning of Spatially-Enhanced Global and Local 3D Features by Deep Neural Network With Coupled Softmax

Zhizhong Han, Zhenbao Liu, *Member, IEEE*, Chi-Man Vong, *Senior Member, IEEE*, Yu-Shen Liu, *Member, IEEE*, Shuhui Bu, *Member, IEEE*, Junwei Han, *Senior Member, IEEE*, and C. L. Philip Chen, *Fellow, IEEE*

***Abstract*— The discriminability of the bag-of-words representations can be increased via encoding the spatial relationship among virtual words on 3D shapes. However, this encoding task involves several issues, including *arbitrary mesh resolutions*, *irregular vertex topology*, *orientation ambiguity on 3D surface*, *invariance to rigid*, and *non-rigid shape transformations*. To address these issues, a novel unsupervised spatial learning framework based on deep neural network, *deep spatiality* (DS), is proposed. Specifically, DS employs two novel components: *spatial context extractor* and *deep context learner*. Spatial context extractor extracts the spatial relationship among virtual words in a local region into a *raw spatial representation*. Along a *consistent circular direction*, a *directed circular graph* is constructed to encode relative positions between pairwise virtual words in each face ring into a *relative spatial matrix*. By decomposing each relative spatial matrix using singular value decomposition, the raw spatial representation is formed, from which deep context learner conducts unsupervised learning of the global and local features. Deep context learner is a deep neural network with a novel model structure to adapt the proposed *coupled softmax layer*, which encodes not only the discriminative information among local regions but also the one among global shapes. Experimental results show that DS outperforms state-of-the-art methods.**

***Index Terms*— Deep spatial, spatially-enhanced 3D features, directed circular graph, coupled softmax.**

Manuscript received September 14, 2017; revised February 13, 2018; accepted March 8, 2018. Date of publication March 16, 2018; date of current version March 29, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61472202, Grant 61573284, and Grant 61672430, in part by the NWPU Basic Research Fund under Grant 3102016JKBJJGZ08, and in part by the University of Macau under Grant MYRG2016-00134-FST. *(Corresponding author: Yu-Shen Liu.)*

Z. Han is with the School of Software, Tsinghua University, Beijing 100084, China, and also with Northwestern Polytechnical University, Xi'an 710072, China (e-mail: h312h@mail.nwpu.edu.cn).

Z. Liu and S. Bu are with the Shool of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liuzhenbao@nwpu.edu.cn; bushuhui@nwpu.edu.cn).

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: jhan@nwpu.edu.cn).

C.-M. Vong is with the Department of Computer and Information Science, University of Macau, Macau 99999, China (e-mail: cmvong@umac.mo).

Y.-S. Liu is with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: liuyushen@tsinghua.edu.cn).

C. L. P. Chen is with the Faculty of Science and Technology, University of Macau, Macau 99999, China (e-mail: philip.chen@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2816821

I. Introduction

BAG of words (BoW) representations have been widely used for natural language processing [1], computer vision [2] and 3D shape analysis [3]–[5]. BoW originally represents a text, such as a sentence or a document, by an occurrence frequency histogram of words over a predefined dictionary. BoW is further enabled for the representation of an image and a 3D shape using "virtual words", learned over all local features around pixels (image), vertices or faces (3D shape), respectively. Although BoW is simple and effective, its discriminability is significantly limited due to the lack of spatial relationships between words (or virtual words). Over the past few decades, various methods have been proposed to encode the spatial relationships for texts and images, such as the n-gram model for texts [1] and the spatial pyramid for images [2]. These methods only encode the spatial relationships as absolute positions, such as front and back or up and down [2], [6], [7]. However, such kind of absolute positions is not an appropriate way of spatial encoding for 3D shapes, which is easily corrupted under rigid (rotation) and non-rigid (the arm bending) 3D shape transformations, orientation ambiguity on 3D surface, irregular vertex topology and arbitrary mesh resolutions [8]–[10]. Therefore, these spatial relationships are useless in 3D shapes. To date, it remains challenging to encode the discriminative spatial relationship among virtual words into the BoW representation for 3D shapes, that are invariant to rigid and non-rigid transformations.

Some methods have been proposed to tackle this challenge. Instead of encoding absolute positions, one strategy is to encode the relative positions between local BoW representations which are calculated from locally segmented regions, such as patches [11], regions between concentric spheres centered at the barycenter [12], and regions between intrinsic isocontours for articulated shapes [13]. However, this strategy only encodes coarse spatial relationships because local BoW representation disregards the spatial information in local region. To encode more detailed spatial relationships, another strategy utilizes pairwise distances, such as geodesic distance or heat diffused distance between two virtual words

on 3D articulated shape [14] or distance between local feature and the barycenter of artificial 3D shape [15]. Although the strategy uses up all distances between pairwise virtual words to overcome the obstacles of 3D shapes (e.g., arbitrary mesh resolution, irregular vertex topology and orientation ambiguity on 3D surface), another issue called spatial ambiguity comes up that often fails to encode the spatial relationship among multiple pairs of virtual words. In fact, the aforementioned two strategies adopt a global perspective which enables the direct encoding of global spatial information. To simultaneously resolve the issues of both strategies, a local perspective is recently proposed by our previous work [16], which indirectly encodes the global spatial information by patterns of local spatial information. However, the indirectly encoded global information is still with limited discriminability, because the patterns are learned merely from local spatial information but without further considering the discriminative information among different shapes. Thus, for various 3D shape analysis applications, the challenge becomes even tougher when more detailed spatial relationships among virtual words are encoded for higher discriminative 3D features.

To remedy these issues, an unsupervised spatial learning framework, named *deep spatiality* (DS), is proposed based on deep neural network. Although deep learning models have been already widely used [17], [18], DS is the first deep learning model to encode spatiality on 3D mesh in an unsupervised way. DS employs two steps to learn the global and local spatial information simultaneously. First, DS adopts the local perspective proposed in our previous work [16] to indirectly encode the global spatial information, i.e., learning the patterns of spatial relationships among virtual words in local regions. Then, DS utilizes a novel structure of deep neural network to learn the patterns of local spatial information through encoding i) the discriminative information among local regions and ii) the one among global shapes. In this way, DS has the ability to simultaneously learn spatially-enhanced global and local 3D features with high discriminability. In detail, there are two novel components based on virtual words in DS, namely, *spatial context extractor* and *deep context learner*, which are summarized as follows.

With striding the aforementioned obstacles of 3D shapes, spatial context extractor extracts the spatial relationship among virtual words in a local region, to be encoded as a *raw spatial representation*. Given the central face of a local region, a central virtual word is assigned to this face. The spatial relationship in the local region is then regarded as the spatial context around the central virtual word. Under a novel *spatial context formalization*, the spatial context is formalized as all relative locations between pairwise virtual words in each neighboring face ring of the central virtual word. In order to effectively encode the spatial context, a novel *directed circular graph* is constructed using the virtual words in each neighboring face ring along a *consistent circular direction*. Subsequently, in each directed circular graph, the relative positions between pairwise virtual words are encoded into a *relative spatial matrix*, from which the spatial context is extracted by the singular value decomposition (SVD). Finally, all spatial context extracted from each neighboring face ring is concatenated together, which forms a raw spatial representation.

Deep context learner simultaneously learns the spatially-enhanced global and local 3D features from these raw spatial representations. It is a specially designed deep neural network with a novel model structure, in which a *coupled softmax layer* is proposed to simultaneously learn global and local 3D features. The learning process takes place in the coupled softmax layer via encoding the discriminative information among local regions and the one among global shapes. The essence of coupled softmax layer is to keep the layer output consistent with unsupervised truth, i.e. i) which type of virtual word a spatial context surrounds, and ii) which shape a local region is from. As a result, the deep context learner has the ability to learn highly discriminative spatially-enhanced 3D features in an unsupervised manner. The significant contributions of our work are summarized as follows.

1) An unsupervised spatial learning framework, i.e. DS, is proposed to learn spatially-enhanced 3D features with high discriminability. DS encodes the global spatial information from the local one based on BoW representations.
2) A spatial context extractor is proposed to capture the spatial relationship among virtual words in a local region, which strides the obstacles of 3D shapes, such as arbitrary vertex number, irregular vertex topology and orientation ambiguity on 3D surface.
3) A deep context learner is proposed as a deep neural network with a novel coupled softmax layer, which enables the learning of spatially-enhanced global and local 3D features simultaneously in an unsupervised manner.

The rest of our paper is organized as follows. Section II reviews the related work of BoW and spatially-enhanced BoW in 3D domain. The overview of DS is introduced in Section III. Spatial context extractor and deep context learner are detailed in Section IV and Section V, respectively. Experimental setup and results with analysis are described in Section VI. Finally, the conclusion and discussions are given in Section VII.

II. Related Work

Two classes of related work are briefly reviewed in this section. First, the methods related to BoW for 3D shapes are introduced. Second, the methods of encoding spatial relationships into BoW representations for 3D shapes are described in order to show the significance of our proposed DS.

A. BoW for 3D Shapes

BoW represents 3D shape using different types of virtual words which are learned from the local features around vertices or faces. The vertices or faces on 3D shape are labeled by the indices of their nearest types of virtual words in the feature space, respectively. Consequently, BoW represents a 3D shape as a frequency histogram of different types of virtual words, which performs statistics for occurred virtual words that are assigned to the vertices or faces on the 3D shape.

BoW has been widely employed for 3D shape recognition and retrieval [12], [19], [20]. The difference among these studies mainly lies in the local features for the learning of different types of virtual words. For example, Lian *et al.* [20] adopted SIFT features to represent 2D views captured from 3D local regions. Li and Godil [12] and Liu *et al.* [19] and employed spin image features of sampled points on the surface. With the help of powerful spectral descriptors, such as heat kernel signature [21] and local spectral descriptor [3], BoW representations obtained better results in [3] and [14]. Recently, Tabia *et al.* [5] introduced covariance matrix to enable efficient fusion of different types of local features and modalities for learning virtual dictionary.

B. Spatially-Enhanced BoW for 3D Shapes

The disadvantage of BoW is the disregard of spatial relationships among virtual words. Similar to spatial pyramid matching (SPM) for images [2], one class of approaches employs spatial relationships among local BoW representations computed from local 3D regions, such as patches [11], regions between concentric spheres centered at the barycenter [12], and regions between intrinsic isocontours for articulated shapes [13].

Specifically, concentric BoW [12] was proposed as a spatially-enhanced BoW representation for 3D shapes. Firstly, concentric spheres adopted by spherical harmonic descriptor [22] was utilized, which separates a shape into multiple regions. Then, the spatial relationships between neighboring regions were encoded via concatenating their corresponding local BoW representations. However, the encoded spatial relationship cannot resist the non-rigid shape transformation, such as bending of the arm, since the distribution of virtual words in each region changes upon non-rigid shape transformation. The same problem also exists in the spatially enhanced BoW [11]. Instead of regions between concentric spheres, local BoW representations were computed from patches which were directly segmented on 3D surface. However, the spatial relationship among these patches were merely encoded by their pairwise Euclidean distance, and therefore, it was not invariant to non-rigid shape transformation. Inspired by SPM [2], intrinsic spatial pyramid matching (ISPM) [13] employed isocontours of the second eigenfunction of Laplace-Beltrami operator to cut the articulated shapes into parts. Although the spatial information encoded by ISPM can resist non-rigid shape transformation, it is only suitable for articulated shapes. This is because the rigid shape cannot be cut into consistent parts by isocontours of its second eigenfunction of Laplace-Beltrami operator.

In a nutshell, the common problem of the aforementioned approaches is that local BoW representation disregards the spatial information in each separated region, leading to unsatisfactory discriminability.

To encode more detailed spatial information, spatial sensitive bag-of-features (SSBoF) employed pairwise geodesic distance or heat diffused distance which is invariant to rigid and non-rigid shape transformations [14]. SSBoF provided a simple 3D shape representation which was the summation of all pairwise distance-weighted features of virtual words. However, the spatial relationship is encoded merely by pairwise distances on 3D surface, which is ambiguous and restricted. This is because the location of one virtual word (determined by a specific distance to another virtual word) can be anywhere on a circle but rather than a fixed location. Therefore, the relative location of two virtual words becomes uncertain and hence causes spatial ambiguity. In addition, the ambiguity makes SSBoF fail to encode the spatial relationship among multiple pairs of virtual words. As a result, the spatial information in a local region cannot be fully captured, which is in fact a significant element for discriminating shapes. The same issue also exists when employing the Euclidean distance between local feature and the barycenter in [15], where the Euclidean distance is only meaningful for rigid 3D shapes and cannot resist non-rigid shape transformation.

All the aforementioned approaches encoded spatial information from a global perspective. However, in this way, the directly encoded global information often leads to high dimension and coarse spatial information in the shape representation, which is incompact for subsequent processing. To resolve this problem, bag of spatial context correlation (BoSCC) [16] was proposed based on a novel local perspective. BoSCC indirectly encoded the global spatial information via learning the patterns of spatial relationships among virtual words in local regions. However, BoSCC merely learns the patterns from local spatial information and does not contain the discriminative information among different shapes, which significantly limits the discriminability of learned global features.

To simultaneously learn highly discriminative global and local 3D features, DS is proposed to capture both the discriminative information among local regions, and the one among global shapes by the novel deep context learner. Although DS also adopts the local perspective as BoSCC, it is able to capture more detailed spatial information than BoSCC and other methods by the novel spatial context extractor. Moreover, DS is also able to learn more discriminative and compact features than the global perspective based methods.

III. Overview of Deep Spatiality

DS is composed of four parts: virtual word learner, local region sampler, spatial context extractor, and deep context learner. The overview of DS is illustrated in Fig. 1.

A. Virtual Word Learner

The first part of DS is virtual word learner, and it learns different types of virtual words over low-level features of faces from all 3D shapes. Given a set of 3D shapes $\mathbf{M} = \{\boldsymbol{M}^m | m \in [1, \mathcal{M}]\}$ as briefly shown in Fig. 1 (a), the low-level feature $\boldsymbol{f}_j^m$ is first computed for the j-th face F_j^m on the m-the shape $\boldsymbol{M}^m$, where $\boldsymbol{M}^m = \{F_j^m | j \in [1, \mathcal{N}^m]\}$. $\boldsymbol{f}_j^m$ is constructed via concatenating the features obtained by many informative local descriptors. Similar to [23], the descriptors used in our method include shape diameter, multi-scale surface curvature, singular values extracted from principal component analysis of local regions, distances from medial surface points, average geodesic distances, shape contexts, and spin images.

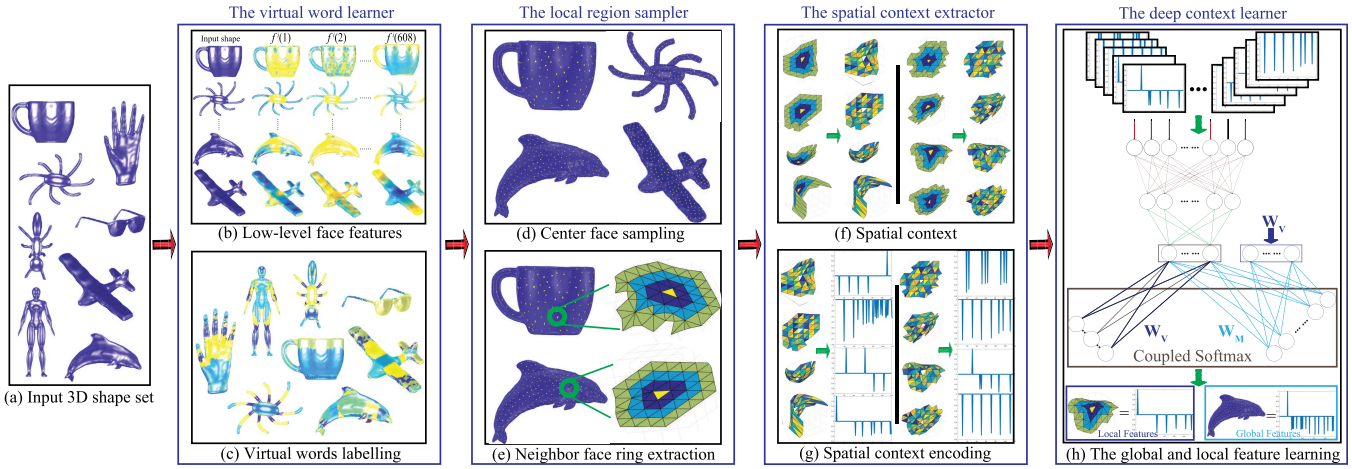

Fig. 1. The overview of deep spatiality, which consists of virtual word learner ((b) and (c)), local region sampler ((d) and (e)), spatial context extractor ((f) and (g)) and deep context learner (h). The input 3D shapes are briefly shown in (a).

As a result, $\boldsymbol{f}_j^m$ is formed as a vector which is illustrated in Fig. 1 (b). Then, $\mathcal{V}$ types of virtual words are learned over all $\boldsymbol{f}_j^m$ from $\mathbf{M}$ by K-means clustering, where each cluster center is regarded as a *type of virtual word*. The low-level features of all types of virtual words are denoted as $\mathbf{F}_V = \{\boldsymbol{f}_\upsilon | \upsilon \in [1, \mathcal{V}]\}$. Finally, each face F_j^m is assigned to its nearest type of virtual word whose index is denoted as L_j^m to label F_j^m. The labelled result is briefly shown in Fig. 1 (c).

B. Local Region Sampler

The second part of DS is local region sampler, and it uniformly samples $\mathcal{K}$ local regions from each shape $\boldsymbol{M}^m$. To preserve the spatial relationship among virtual words in an arbitrary local region G, G is modeled as the combination of a central face F and its $\mathcal{R}$ neighboring face rings n_r, where $r \in [1, \mathcal{R}]$. To sample $\mathcal{K}$ local regions, $\mathcal{K}$ central faces are determined via uniformly sampling $\mathcal{K}$ vertices by farthest geodesic sampling method [24]. For each sampled vertex, a central face is randomly selected from the faces that surround the sampled vertex. In Fig. 1 (d) and (e), the sampled central faces and their corresponding local regions are shown, respectively. In our experiments, $\mathcal{K}$ is typically set to 500.

C. Spatial Context Extractor

The third part of DS is spatial context extractor, and it extracts the spatial information from each local region. With handling various mesh resolution, irregular vertex topology and orientation ambiguity in a local region G, spatial context extractor C encodes the spatial relationship among virtual words in G into a raw spatial representation $C(G)$. The spatial relationship among the virtual words in G is regarded as the spatial context of central virtual word L assigned to the central face F of G. The spatial context of L is represented by the virtual words in each neighboring face ring n_r of F. This procedure is shown in Fig. 1 (f) and Fig. 1 (g), respectively.

D. Deep Context Learner

The fourth part of DS is deep context learner, and it performs the learning of global and local 3D features. The raw spatial representations of all sampled regions are provided to deep context learner for learning spatially-enhanced global and local 3D features, as illustrated in Fig. 1 (h). Deep context learner employs the novel coupled softmax layer to guide the learning procedure in an unsupervised manner. Spatial context extractor and deep context learner are detailed in the following two sections, respectively.

IV. Spatial Context Extractor

A. An Overview of Spatial Context Extractor

With striding obstacles of 3D shapes, spatial context extractor C is proposed to capture the spatial relationship among virtual words in a local region G. This is achieved via extracting the spatial context around the central virtual word L of G.

An overview of spatial context extractor is illustrated in Fig. 2. For G shown in Fig. 2 (a), a novel spatial context formalization is applied to formalize the spatial context of L in the neighboring face ring n_r, i.e., as all relative locations between pairwise virtual words in n_r. Each relative location is encoded by the circular distance between pairwise virtual words along a consistent circular direction. As shown in Fig. 2 (b), this novel spatial context formalization makes it feasible to capture the spatial relationship among virtual words on 3D surface. This is because the issues of various mesh resolution and irregular vertex topology can be resolved by measuring circular distance in a ring-by-ring manner. Furthermore, the orientation ambiguity on 3D surface is eliminated with the help of consistent circular direction.

Based on the spatial context formalization, a directed circular graph is constructed in each n_r as shown in Fig. 2 (c), where the spatial context of L in n_r can be encoded into a relative spatial matrix $\boldsymbol{S}_r$ as shown in Fig. 2 (d). $\boldsymbol{S}_r$ is constructed by relative locations between all pairwise virtual words in n_r in terms of circular distances. To extract the spatial

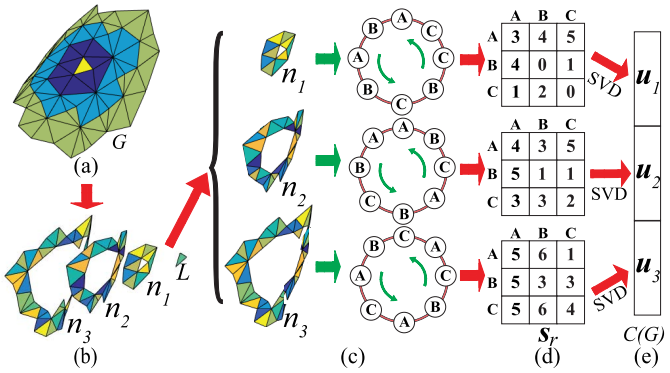

Fig. 2. The overview of spatial context extractor. (a) A local region G. (b) The spatial context of central virtual word L in G. (c) The directed circular graph constructed by virtual words in each neighboring face ring n_r. (d) The relative spatial matrix $\boldsymbol{S}_r$ representing each n_r. (e) The raw spatial representation of G, $C(G)$, formed by concatenating singular vector $\boldsymbol{u}_r$ of $\boldsymbol{S}_r$ under the SVD.

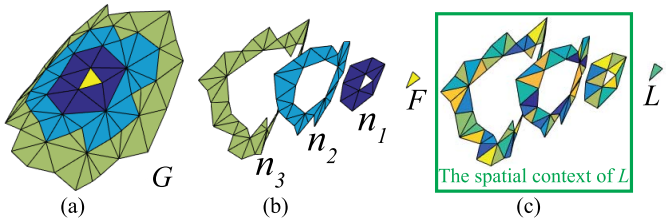

Fig. 3. The modeling of local regions and the spatial context formalization. (a) A 3D local region. (b) Neighboring face rings. (c) The spatial context.

context of L in $\boldsymbol{S}_r$, the left singular vector $\boldsymbol{u}_r$, corresponding to the maximum singular value, is employed via decomposing $\boldsymbol{S}_r$ by the SVD. Finally, the raw spatial representation $C(G)$ is formed by concatenating $\boldsymbol{u}_r$ obtained from each n_r, such that $C(G) = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$, where $r \in [1, \mathcal{R}]$, as shown in Fig. 2 (e).

Compared to the Markov model based methods which merely employ the co-occurrence frequency of neighboring virtual words [16], our SVD based spatial context encoding is with superior performance because the higher order spatial relationship can be captured by all pairwise relative locations in each n_r.

B. Spatial Context Formalization

No matter how many vertices in the m-th shape $\boldsymbol{M}^m$ and how these vertices are connected, a 3D local region G from $\boldsymbol{M}^m$ (as shown in Fig. 3 (a)) can be always formed by a central face F and its neighboring face rings n_r. As illustrated in Fig. 3 (b), colors are used to distinguish the faces in different n_r. Based on the modelling of G, the spatial relationships between virtual words in G is regarded as the spatial context of L which is composed by surrounding virtual words in each n_r. The spatial context of L is shown in Fig. 3 (c), where the types of virtual words are indicated by different colors. Finally, the spatial context of L is formalized as relative positions between all pairwise virtual words in each n_r. Based on the spatial context formalization, the spatial relationships extracted from different local regions are comparable.

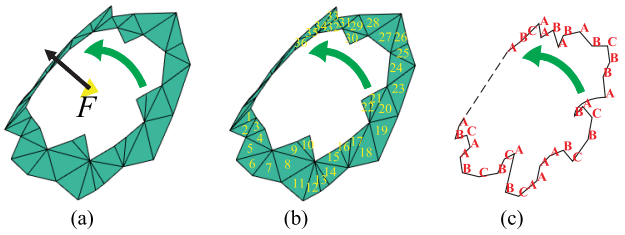

Fig. 4. The consistent circular direction and the directed circular graph. (a) The circular direction. (b) A sequential face ring. (c) A directed circular graph.

C. Construction of Directed Circular Graph

In each n_r, to encode the relative location between pairwise virtual words, a directed circular graph is constructed via ordering the virtual words along a consistent circular direction.

The circular direction is established by the right-hand corkscrew-rule, that is, the direction of wrapping one's right hand when the thumb is pointing in the same direction as the norm of central face F. As shown in Fig. 4 (a), the circular direction (green arrow) is about the norm of F (black arrow). This circular direction is consistent across different 3D local regions, which increases the discriminability via encoding directed relative locations.

Along the circular direction, the faces in n_r can be expressed as a sequential face ring via analysing the face by face neighboring relationship, as shown in Fig. 4 (b), where the sequential face order is indicated by the yellow numbers. The directed circular graph is finally constructed via placing the virtual word L_j^m assigned to each F_j^m in n_r using the same sequential face order. An example including three types of virtual words (A, B and C) is shown in Fig. 4 (c). In a directed circular graph, a node represents a virtual word and an edge connects the pairwise neighboring virtual words.

Based on the directed circular graph, the relative location between pairwise virtual words L_j^m and $L_{j'}^m$ is encoded by the circular distance $d(L_j^i, L_{j'}^i)$ which is measured along the circular direction. To handle various mesh resolutions and irregular vertex topology, $d(L_j^i, L_{j'}^i)$ is set to a value of one if L_j^m and $L_{j'}^m$ are neighbors. Otherwise, it is set to the length of path from L_j^m to $L_{j'}^m$ along the circular direction. Due to the circular direction, the circular distance eliminates the spatial ambiguity between two virtual words. Taking the directed circular graph in Fig. 5 (a) for example, three types of virtual words (A, B and C) are used to label four faces whose labels are denoted as $L_1, \ldots, L_4$, respectively, where the green arrow indicates the circular direction. The circular distances starting from L_1 are illustrated in Fig. 5 (b). In this example, the circular distance between two neighbors L_1 and L_2 is $d(L_1, L_2) = 1$, and the path length from L_1 to L_3, $d(L_1, L_3) = 2$. Similarly, $d(L_1, L_4) = 3$.

D. Spatial Context Extraction

In each n_r, all pairwise relative locations are encoded into a relative spatial matrix $\boldsymbol{S}_r$, which represents spatial relationships instead of merely considering the neighboring

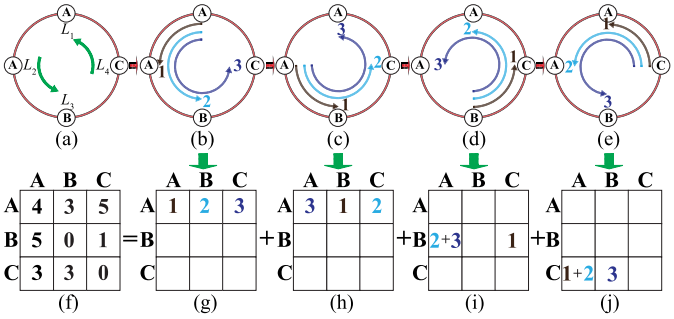

Fig. 5. The procedure of encoding all relative locations between pairwise virtual words in a circular directed graph into a relative spatial matrix.

co-occurrence frequency in BoSCC [16]. $\boldsymbol{S}_r$ is a $\mathcal{V} \times \mathcal{V}$ matrix, in which each entry s_{pq} represents the spatial relationships between the p-th type and the q-th type of virtual words along the circular direction. Therefore, s_{pq} is calculated via summing up all circular distances of pathes from the p-th type to the q-th type of virtual words.

For a clearer description, all virtual words with the same υ-th type form a set in each n_r, as denoted by $\mathbb{L}_\upsilon^r = \{L_j^m | L_j^m = \upsilon, F_j^m \in n_r\}$. As defined in Eq. (1), s_{pq} summaries all pairwise circular distances from L_i to $L_{i'}$, where $L_i \in \mathbb{L}_p^r$ and $L_{i'} \in \mathbb{L}_q^r$.

$$s_{pq} = \sum_{L_i \in \mathbb{L}_p^r} \sum_{L_{i'} \in \mathbb{L}_q^r} d(L_i, L_{i'}). \tag{1}$$

This procedure can be further illustrated in Fig. 5. The circular distances starting from L_1, L_2, L_3 and L_4 are shown from Fig. 5 (b) to Fig. 5 (e), respectively, and the corresponding circular distances (in the same color) are obtained as shown in the matrices from Fig. 5 (g) to Fig. 5 (j). Finally, $\boldsymbol{S}_r$ is the summation of all the corresponding circular distances, as shown in Fig. 5 (f). To handle various number of virtual words in the directed circular graph, $\boldsymbol{S}_r$ is normalized by the summation of each row, such that $s_{pq} := s_{pq} / \sum_{q \in [1,\mathcal{V}]} s_{pq}$.

Since all $\mathcal{V}$ types of virtual words may not appear altogether in each n_r of G, $\boldsymbol{S}_r$ is not always a full rank matrix. As a result, the sorted eigenvalues or the eigenvector corresponding to a specific eigenvalue are unable to extract the spatial context of L in n_r from $\boldsymbol{S}_r$. It is because the eigenvalues or the eigenvectors of $\boldsymbol{S}_r$ are not always real numbers. To resolve this issue, the SVD is used to decompose $\boldsymbol{S}_r$, whose left singular vector corresponding to the maximum singular value, $\boldsymbol{u}_r$, is employed to represent the spatial context of L in n_r. Through the SVD, $\boldsymbol{u}_r$ does not degenerate the encoding of spatial context captured by $\boldsymbol{S}_r$ while reducing the dimensionality from $\mathcal{V} \times \mathcal{V}$ to $1 \times \mathcal{V}$.

V. Deep Context Learner

A. An Overview of Deep Context Learner

To benefit from the powerful learning ability, deep context learner is specially designed based on the deep neural network, which aims to simultaneously learn the spatially-enhanced 3D global and local features from the raw spatial representations of local regions.

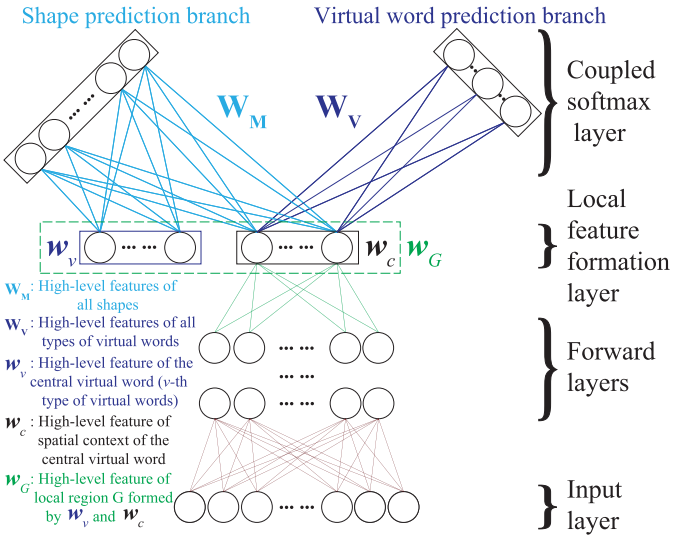

Fig. 6. The structure of deep context learner is composed of input layer, forward layers, local feature formation layer, and coupled softmax layer. The local region G which is centered by the υ-th type of virtual word is used to illustrate the learning of the high-level features $\mathbf{W}_M$ of all $\mathcal{M}$ shapes and the high-level features $\mathbf{W}_V$ of all $\mathcal{V}$ types of virtual words.

Deep context learner is designed according to the following three aspects. First, since deep context learner adopts the local perspective, both global and local features are learned based on the spatial context patterns. To capture the spatial context patterns, several forward layers are employed. Second, a local feature formation layer is designed to form the feature of local regions, which is used to predict the shape that a local region comes from. Third, coupled softmax layer is proposed to guide the unsupervised learning in deep context learner. It also encodes both the discriminability information among local regions and the one among global shapes.

As shown in Fig. 6, deep context learner consists of four parts, namely, input layer, forward layers, local feature formation layer, and coupled softmax layer, which are detailed in the following.

B. Structure of Deep Context Learner

1) Input Layer: The raw spatial representations of all sampled local regions are used to train deep context learner. Each raw spatial representation is fed into deep context learner from input layer.

2) Forward Layers: Forward layers are $\mathcal{H}$ layers of neural networks, and each layer contains $\mathcal{D}^l$ nodes, where $l \in \{1, \mathcal{H}\}$. Forward layers aim to capture the spatial context patterns from raw spatial representations by hierarchical abstraction. In Fig. 6, the high-level feature of spatial context, denoted by $\boldsymbol{w}_c \in \mathbb{R}^{1 \times \mathcal{T}}$, is formed by the spatial context patterns captured from forward layers. Under $\boldsymbol{w}_c$, the effect of outlier virtual words caused by non-rigid shape transformations can be alleviated effectively.

3) Local Feature Formation Layer: This layer is specially designed to form the high-level feature of a local region G, $\boldsymbol{w}_G$, as shown in Fig. 6. Since G is formalized by a central virtual word L and its spatial context (where L belongs to the

υ-th type virtual word), $\boldsymbol{w}_G$ is formed via concatenating the high-level features of L and its spatial context, $\boldsymbol{w}_\upsilon$ and $\boldsymbol{w}_c$, such that $\boldsymbol{w}_G = [\boldsymbol{w}_\upsilon, \boldsymbol{w}_c]$.

4) Coupled Softmax Layer: This layer aims to learn the high-level features of all types of virtual words and the global features of 3D shapes, denoted as $\mathbf{W}_V = \{\boldsymbol{w}_\upsilon | \upsilon \in [1, \mathcal{V}]\}$ and $\mathbf{W}_M = \{\boldsymbol{w}_m | m \in [1, \mathcal{M}]\}$, respectively, as illustrated in Fig. 6. This layer guides the whole learning process of deep context learner in an unsupervised manner via simultaneously answering the following two prediction questions.

1) Given the high-level feature of its spatial context, which type of virtual word does a central virtual word belong to?
2) Given the high-level feature of a local region, which shape does the local region come from?

The first prediction encodes semantic similarities between different types of virtual words into $\mathbf{W}_V$, which makes $\mathbf{W}_V$ more discriminative than $\mathbf{F}_V$. The semantic similarities imply that two types of virtual words are similar if their surrounding spatial contexts are always similar. For example, if two types of virtual words are always surrounded by similar spatial contexts, then the distance d_w between their $\boldsymbol{w}_\upsilon$ is drawn closer than the distance d_f between their $\boldsymbol{f}_\upsilon$; otherwise, d_w is separated farther than d_f. Furthermore, $\boldsymbol{w}_\upsilon$ is helpful to increase the discriminability of $\boldsymbol{w}_G$, since two local regions can be still similar even if their central virtual words are different. Similarly, the second prediction embeds semantic similarities between different shapes into $\mathbf{W}_M$, that is, two 3D shapes are similar if they always contain similar local regions.

The semantic similarities can be encoded via addressing the aforementioned questions separately, which is an alternative way of learning $\mathbf{W}_V$ and $\mathbf{W}_M$. In other words, $\mathbf{W}_V$ and $\boldsymbol{w}_c$ of each spatial context are learned by a neural network with a softmax layer to answer the first prediction question. Subsequently, $\mathbf{W}_M$ is learned from the fixed $\mathbf{W}_V$ and $\boldsymbol{w}_c$ by another neural network with a softmax layer to answer the second prediction question. However, this alternative does not consider the discriminative information between shapes when learning $\mathbf{W}_V$ and $\boldsymbol{w}_c$, which limits the discriminability of the learned features. To resolve this issue, coupled softmax layer is proposed to address these questions in a coupled way, which integrates the learning of global and local features together to increase their discriminability.

Coupled softmax layer not only encodes the semantic similarities but also enhances them in the learning process. It first encodes the semantic similarities among $\boldsymbol{w}_m$ into $\mathbf{W}_V$. Next, the semantic similarities among $\boldsymbol{w}_\upsilon$ are encoded into $\mathbf{W}_M$. These two encoding procedures are mutually enhanced by one another iteratively. Taking two types of virtual words with similar spatial context for example, their high-level features $\boldsymbol{w}_\upsilon$ are drawn much closer if they always appear in two similar shapes. In contrast, for two types of virtual words with dissimilar spatial context, their corresponding $\boldsymbol{w}_\upsilon$ are separated much farther if they always appear in two dissimilar shapes. Furthermore, if two shapes contain many similar types of virtual words, their corresponding $\boldsymbol{w}_m$ are drawn much closer; otherwise, their corresponding $\boldsymbol{w}_m$ are separated much farther.

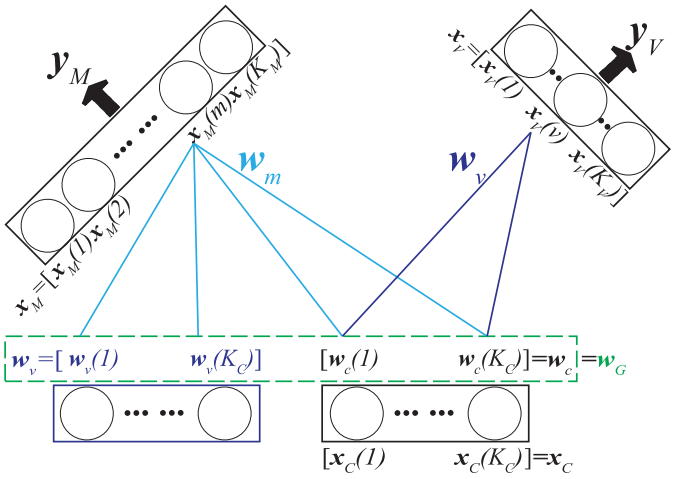

Fig. 7. The parameters involved in the gradient inference within deep context learner are illustrated, where the high-level feature $\boldsymbol{w}_m$ of the m-th shape and the high-level feature $\boldsymbol{w}_\upsilon$ of the υ-th type of virtual word are learned from the high-level feature $\boldsymbol{w}_G$ of local region G.

In coupled softmax layer, the two questions are simultaneously addressed by two coupled branches as shown in Fig. 6, namely *virtual word prediction branch* and *shape prediction branch*, respectively. The two branches employ conditional probability to evaluate the prediction results. Virtual word prediction branch provides $P(L = \upsilon | \boldsymbol{w}_c)$ to explain how likely the central virtual word L is with the υ-th type ($L = \upsilon$) when given the high-level feature $\boldsymbol{w}_c$ of the spatial context of L. Similarly, shape prediction branch uses $P(O = m | \boldsymbol{w}_G)$ to explain how likely a local region G expressed by its high-level feature $\boldsymbol{w}_G$ is from the m-th shape $\boldsymbol{M}^m$ ($O = m$).

C. The Learning of Deep Context Learner

1) The Modeling of Coupled Softmax Layer: Deep context learner is trained over the raw spatial representation of region G_i from the O_i-th shape, where $i \in [1, I]$ and I is the number of sampled regions.

Virtual word prediction branch conducts the prediction in terms of the probability defined as follows,

$$P(L_i = q | \boldsymbol{w}_{i,c}) = \frac{exp(\boldsymbol{w}_{i,c}^T \boldsymbol{w}_q + b_q)}{\sum_{\upsilon=1}^{\mathcal{V}} exp(\boldsymbol{w}_{i,c}^T \boldsymbol{w}_\upsilon + b_\upsilon)}. \tag{2}$$

In Eq. (2), given the high-level feature $\boldsymbol{w}_{i,c}$ of the spatial context of L_i, the probability that the central virtual word L_i in G_i is the q-th type of virtual word ($L_i = q$) is provided. $\boldsymbol{w}_q$ and b_q are the high-level feature and the bias of the q-th type of virtual word, respectively, moreover, $\boldsymbol{w}_\upsilon$ and b_υ are the ones of the υ-th type of virtual word. As illustrated in Fig. 7, $\boldsymbol{w}_{i,c}$ is computed through sigmoid function from the input of local feature formation layer $\boldsymbol{x}_{i,C}$, such that $\boldsymbol{w}_{i,c} = 1/(1 + exp(-\boldsymbol{x}_{i,C} + \boldsymbol{c}))$, where $\boldsymbol{c}$ is the bias vector of local feature formation layer. For clarity in Fig. 7, the subscript $\{i,\}$ in the involved parameters is omitted. Furthermore, $\boldsymbol{x}_{i,V}$ is used to denote the input of virtual word prediction branch. Each entry $\boldsymbol{x}_{i,V}(\upsilon)$ of $\boldsymbol{x}_{i,V}$ is the similarity between $\boldsymbol{w}_{i,c}$ and $\boldsymbol{w}_\upsilon$,

such that $\boldsymbol{x}_{i,V}(\upsilon) = \boldsymbol{w}_{i,c}^T \boldsymbol{w}_\upsilon$ and $\upsilon \in [1, \mathcal{V}]$. Note that sigmoid function is employed as the activation function in deep context learner.

In addition, shape prediction branch conducts the prediction in terms of the probability defined as follows,

$$P(O_i = p|\boldsymbol{w}_{i,G}) = \frac{exp(\boldsymbol{w}_{i,G}^T \boldsymbol{w}_p + a_p)}{\sum_{m=1}^{\mathcal{M}} exp(\boldsymbol{w}_{i,G}^T \boldsymbol{w}_m + a_m)}. \quad (3)$$

In Eq. (3), given the high-level feature $\boldsymbol{w}_{i,G}$ of G_i, the probability that the region G_i is from the p-th shape ($O_i = p$) is provided. $\boldsymbol{w}_{i,G} = [\boldsymbol{w}_{L_i}, \boldsymbol{w}_{i,c}]$, where the central virtual word of G_i is with the L_i-th type. $\boldsymbol{w}_p$ and a_p are the high-level feature and the bias of the p-th shape, respectively, moreover, $\boldsymbol{w}_m$ and a_m are the ones of the m-th shape. Similar to $\boldsymbol{x}_{i,V}$, the input of shape prediction branch is denoted by $\boldsymbol{x}_{i,M}$, as illustrated in Fig. 7. Each entry $\boldsymbol{x}_{i,M}(m)$ of $\boldsymbol{x}_{i,M}$ is the similarity between $\boldsymbol{w}_{i,G}$ and $\boldsymbol{w}_m$, such that $\boldsymbol{x}_{i,M}(m) = \boldsymbol{w}_{i,G}^T \boldsymbol{w}_m$, and $m \in [1, \mathcal{M}]$.

Both Eq. (2) and Eq. (3) employ an inner product to evaluate the prediction, followed by normalization based on the inner products among all other candidates. Note that the length of $\boldsymbol{w}_\upsilon$ is with the same dimension as $\boldsymbol{w}_{i,c}$, and the length of $\boldsymbol{w}_m$ is twice the length of $\boldsymbol{w}_\upsilon$.

In addition to $\mathbf{W_M}$ and $\mathbf{W_V}$, $\mathbf{W}$ is used to denote the other parameters involved in deep context learner, including the weights between input layer and forward layers, the weights in forward layers, the weights between forward layers and local feature formation layer, and the weights in local feature formation layer. The overall parameters involved in deep context learner is denoted as $\boldsymbol{\Theta} = [\mathbf{W}, \mathbf{W_M}, \mathbf{W_V}]$.

Deep context learner is trained by finding $\boldsymbol{\Theta}$ that maximizes the log-likelihood defined in Eq. (4),

$$E = \frac{1}{I} \sum_{i=1}^{I} (\log P(L_i|\boldsymbol{\Theta}) + \log P(O_i|\boldsymbol{\Theta})). \quad (4)$$

The parameters $\boldsymbol{\Theta}$ can be learned by back propagation algorithm, where the error between the prediction and the ground truth is determined in coupled softmax layer. In practice, stochastic gradient ascent is employed to perform the learning process. In the following, the gradient inference in deep context learner is presented.

2) The Gradient Inference in Deep Context Learner: The essence of gradient inference in deep context learner lies on three partial derivatives of cost function E, with respect to $\boldsymbol{w}_m$, $\boldsymbol{w}_\upsilon$ and $\boldsymbol{x}_{i,C}$, i.e., $\partial E/\partial \boldsymbol{w}_m$, $\partial E/\partial \boldsymbol{w}_\upsilon$ and $\partial E/\partial \boldsymbol{x}_{i,C}$. Then, the partial derivatives of cost function E with respect to other parameters in $\boldsymbol{\Theta}$ can be inferred based on $\partial E/\partial \boldsymbol{x}_{i,C}$ as the traditional feedforward network.

Firstly, $\partial E/\partial \boldsymbol{w}_m$, $\partial E/\partial \boldsymbol{w}_\upsilon$ and $\partial E/\partial \boldsymbol{x}_{i,C}$ are inferred, where the gradient inferring process can also be illustrated in Fig. 7.

$\partial E/\partial \boldsymbol{w}_m$ is derived in Eq. (5), where $\boldsymbol{O}_i$ is the shape label vector which is a one-hot vector with only the O_i-th entry set to one

$$\begin{aligned}\frac{\partial E}{\partial \boldsymbol{w}_m} &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_m} + \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_m} \\ &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_m} = \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,M}} \frac{\partial \boldsymbol{x}_{i,M}}{\partial \boldsymbol{w}_m} \\ &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,M}} \boldsymbol{w}_{i,G} = \frac{1}{I}\sum_{i=1}^{I} (\boldsymbol{y}_{i,M} - \boldsymbol{O}_i)\boldsymbol{w}_{i,G},\end{aligned} \quad (5)$$

where $\partial \log P(L_i|\boldsymbol{\Theta})/\partial \boldsymbol{w}_m$ is equal to $\mathbf{0}$, since $\boldsymbol{w}_m$ does not contribute to virtual word prediction branch; similar to softmax classifier, $\partial \log P(O_i|\boldsymbol{\Theta})/\partial \boldsymbol{x}_{i,M}$ can be derived as the error between the prediction of shape prediction branch $\boldsymbol{y}_{i,M}$, and the ground truth $\boldsymbol{O}_i$. Correspondingly, the partial derivatives of E with respect to $\boldsymbol{w}_{i,G}$ is derived by

$$\begin{aligned}\frac{\partial E}{\partial \boldsymbol{w}_{i,G}} &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,G}} + \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,G}} \\ &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,G}} = \frac{1}{I}\sum_{i=1}^{I} \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,M}} \frac{\partial \boldsymbol{x}_{i,M}}{\partial \boldsymbol{w}_{i,G}} \\ &= \frac{1}{I}\sum_{i=1}^{I}\sum_{m=1}^{\mathcal{M}} \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,M}} \boldsymbol{w}_m \\ &= \frac{1}{I}\sum_{i=1}^{I}\sum_{m=1}^{\mathcal{M}} (\boldsymbol{y}_{i,M} - \boldsymbol{O}_i)\boldsymbol{w}_m,\end{aligned} \quad (6)$$

where $\partial \log P(L_i|\boldsymbol{\Theta})/\partial \boldsymbol{w}_{i,G}$ is equal to $\mathbf{0}$, since $\boldsymbol{w}_{i,G}$ does not directly contribute to virtual word prediction branch.

$\partial E/\boldsymbol{w}_\upsilon$ is derived by Eq. (7) which includes two terms. The first term represents the error back propagated from virtual word prediction branch, which contains the discriminative information among different types of virtual words. The second term represents the error back propagated from shape prediction branch, which correspondingly contains the discriminative information among different shapes.

$$\frac{\partial E}{\partial \boldsymbol{w}_\upsilon} = \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_\upsilon} + \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_\upsilon}. \quad (7)$$

The first term of Eq. (7) is further derived by Eq. (8),

$$\frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_\upsilon} = \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,V}} \frac{\partial \boldsymbol{x}_{i,V}}{\partial \boldsymbol{w}_\upsilon} = (\boldsymbol{y}_{i,V} - \boldsymbol{L}_i)\boldsymbol{w}_{i,c}, \quad (8)$$

where $\boldsymbol{L}_i$ is the virtual word label vector which is a one-hot vector with only the L_i-th entry set to one; similar to softmax classifier, $\partial \log P(L_i|\boldsymbol{\Theta})/\partial \boldsymbol{x}_{i,V}$ can be derived as the error between the prediction of virtual word prediction branch $\boldsymbol{y}_{i,V}$ and the ground truth $\boldsymbol{L}_i$; moreover, $\partial \boldsymbol{x}_{i,V}/\partial \boldsymbol{w}_\upsilon = \boldsymbol{w}_{i,c}$.

There are two cases for the second term in Eq. (7), since only the υ-th type of virtual word is employed as the central virtual word of G_i. Therefore, one is $L_i = \upsilon$, the other is $L_i \neq \upsilon$. For the first case, $\partial \log P(O_i|\boldsymbol{\Theta})/\partial \boldsymbol{w}_\upsilon$ is the first half of $\partial E/\partial \boldsymbol{w}_{i,G}$ whose indices are $\langle 1 : \mathcal{T} \rangle$, since

$\boldsymbol{w}_{i,G} = [\boldsymbol{w}_{L_i}, \boldsymbol{w}_{i,c}]$. While $\partial \log P(O_i|\boldsymbol{\Theta})/\partial \boldsymbol{w}_\upsilon$ is equal to $\mathbf{0}$ when $L_i \neq \upsilon$, since only $\boldsymbol{w}_\upsilon$ has to be updated. In summary, the second term in Eq. (7) can be detailed below

$$\frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_\upsilon} = \begin{cases} \dfrac{\partial E}{\partial \boldsymbol{w}_{i,G}}\langle 1:\mathcal{T}\rangle; & \text{if } L_i = \upsilon \\ \mathbf{0}; & \text{otherwise.} \end{cases} \tag{9}$$

According to Eq. (7), Eq. (8) and Eq. (9), $\partial E/\partial \boldsymbol{w}_\upsilon$ is finally obtained as

$$\frac{\partial E}{\partial \boldsymbol{w}_\upsilon} = \frac{1}{I}\sum_{i=1}^{I} \begin{cases} (\boldsymbol{y}_{i,V} - \boldsymbol{L}_i)\boldsymbol{w}_{i,c} + \dfrac{\partial E}{\partial \boldsymbol{w}_{i,G}}\langle 1:\mathcal{T}\rangle; & \text{if } L_i = \upsilon \\ (\boldsymbol{y}_{i,V} - \boldsymbol{L}_i)\boldsymbol{w}_{i,c}; & \text{otherwise.} \end{cases} \tag{10}$$

Besides the parameters in coupled softmax layer, other parameters in $\boldsymbol{\Theta}$ can be updated based on $\partial E/\partial \boldsymbol{x}_{i,C}$ as derived by

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{x}_{i,C}} &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,C}} + \frac{\partial log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,C}} \\ &= \frac{1}{I}\sum_{i=1}^{I} \frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,c}}\frac{\partial \boldsymbol{w}_{i,c}}{\partial \boldsymbol{x}_{i,C}} + \frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,c}}\frac{\partial \boldsymbol{w}_{i,c}}{\partial \boldsymbol{x}_{i,C}} \\ &= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\partial \log P(L_i|\boldsymbol{\Theta})}{\partial \boldsymbol{x}_{i,V}}\frac{\partial \boldsymbol{x}_{i,V}}{\partial \boldsymbol{w}_{i,c}} + \frac{\partial \log P(O_i|\boldsymbol{\Theta})}{\partial \boldsymbol{w}_{i,c}}\right)\frac{\partial \boldsymbol{w}_{i,c}}{\partial \boldsymbol{x}_{i,C}} \\ &= \frac{1}{I}\sum_{i=1}^{I}\left(\sum_{\upsilon=1}^{\mathcal{V}}(\boldsymbol{y}_{i,V} - \boldsymbol{L}_i)\boldsymbol{w}_\upsilon + \frac{\partial E}{\partial \boldsymbol{w}_{i,G}}\langle \mathcal{T}+1:end\rangle\right) \\ &\quad \times \frac{\partial \boldsymbol{w}_{i,c}}{\partial \boldsymbol{x}_{i,C}}, \end{aligned} \tag{11}$$

where $\langle \mathcal{T}+1:end\rangle$ represents the indices of the second half of $\partial E/\partial \boldsymbol{w}_{i,G}$.

Since the activation function is sigmoid function, $\partial \boldsymbol{w}_{i,c}/\partial \boldsymbol{x}_{i,C}$ is equal to $\boldsymbol{w}_{i,c}(\mathbf{1} - \boldsymbol{w}_{i,c})$. Then, $\partial E/\partial \boldsymbol{x}_{i,C}$ is finally obtained as

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{x}_{i,C}} &= \frac{1}{I}\sum_{i=1}^{I}\sum_{\upsilon=1}^{\mathcal{V}}(\boldsymbol{y}_{i,V} - \boldsymbol{L}_i)\boldsymbol{w}_\upsilon \boldsymbol{w}_{i,c}(\mathbf{1} - \boldsymbol{w}_{i,c}) \\ &\quad + \frac{1}{I}\sum_{i=1}^{I}\frac{\partial E}{\partial \boldsymbol{w}_{i,G}}\langle \mathcal{T}+1:end\rangle \boldsymbol{w}_{i,c}(\mathbf{1} - \boldsymbol{w}_{i,c}). \end{aligned} \tag{12}$$

Using the aforementioned gradient inference, stochastic gradient ascent on the neural network are performed to iteratively update $\boldsymbol{\Theta}$ as defined in Eq. (13), where ε is the learning rate.

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} + \varepsilon\frac{\partial E}{\partial \boldsymbol{\Theta}}. \tag{13}$$

VI. Experimental Results and Analysis

In this section, the performance of DS is evaluated and analyzed. The setup of parameters involved in DS is firstly discussed. These parameters are tuned to demonstrate how they affect the discriminability of global features learned by DS for global shape retrieval. Then, DS is compared with state-of-the-art methods under different shape benchmarks for several typical 3D shape analysis applications, including global shape retrieval, shape classification, partial shape retrieval and shape correspondence. The shapes employed in the experiments are with various mesh resolutions, irregular vertex topology and different rigid and non-rigid shape transformations.

TABLE I

The Measures of Shape Retrieval Obtained by BoW With Different $\mathcal{V}$ Under McGill Dataset

$\mathcal{V}$	Retrieval measures					
	NN	FT	ST	E	DCG	aPR
40	0.7743	0.4156	0.5588	0.3955	0.7883	8.1665
80	0.8053	**0.4461**	**0.5720**	**0.4180**	0.8005	8.4145
120	0.7898	0.4340	0.5676	0.4110	0.7944	8.2775
160	**0.8186**	0.4460	0.5650	0.4169	**0.8015**	**8.5131**
200	0.8031	0.4161	0.5438	0.3954	0.7897	8.2592

A. The Setup of Parameters for Deep Spatiality

The setup of parameters involved in DS is conducted through global shape retrieval under McGill 3D shape benchmark [25]. These parameters include the size of virtual dictionary $\mathcal{V}$, the number of nodes $\mathcal{D}^l$ in the l-th forward layer, the number of neighboring face rings $\mathcal{R}$, and the number of forward layers $\mathcal{H}$. DS with different parameters are all trained in 3500 epoches and under a learning rate of 0.0001.

1) The Size of Virtual Dictionary $\mathcal{V}$: The first experiment is conducted to obtain an optimal virtual dictionary based on the results of BoW in Table I. Then, other results of DS in this section are obtained based on this virtual dictionary to show the spatiality encoding ability of DS. Five candidate numbers are considered to compare the impact of increasing $\mathcal{V} \in \{40, 80, 120, 160, 200\}$. In Table I, the results obtained by BoW are comprehensively compared in terms of various measures [26], including Nearest Neighbor (NN), First-Tier (FT), Second-Tier (ST), E-Measures (E), Discounted Cumulated Gain vector (DCG) and area under PR curves (aPR). The bold numbers show that the performance with $\mathcal{V} = 160$ is a little better than the one with $\mathcal{V} = 80$, which is the best among all the results. In the following experiments in this subsection, $\mathcal{V} = 160$ is employed as the size of virtual dictionary.

2) The Number of Nodes in the First Forward Layer $\mathcal{D}^1$: Several candidate values are used to evaluate the impact of $\mathcal{D}^1$ on the performance of DS, such that $\mathcal{D}^1 \in \{50, 100, 200, 300, 400, 500, 600, 700\}$. The deep context learner employs only one forward layer to learn from the spatial context extracted from three neighboring face rings. The performance comparison is shown in Table II, where BoW with $\mathcal{V} = 160$ is regarded as a base line. Benefited from the effectively encoded spatial information, DS under different $\mathcal{D}^1$ are all better than BoW. The best result is obtained with $\mathcal{D}^1 = 400$, and hence, $\mathcal{D}^1 = 400$ is adopted in the subsequent experiments. In addition, due to overfitting, the performance of DS cannot be further improved with larger $\mathcal{D}^1$, which is indicated by the worse results with $\mathcal{D}^1 = 500, 600, 700$.

3) The Number of Neighboring Face Rings $\mathcal{R}$: The impact of $\mathcal{R}$ on the performance of DS is analyzed in this experiment. To keep sampled regions local, the performances with $\mathcal{R} = \{1, 2, 3\}$ are compared, and no larger $\mathcal{R}$ is explored. In addition, the comparison is comprehensively

TABLE II

THE MEASURES OF SHAPE RETRIEVAL WITH DIFFERENT $\mathcal{D}^1$ UNDER MCGILL DATASET. $\mathcal{V} = 160$, $\mathcal{R} = 3$, $\mathcal{H} = 1$

$\mathcal{D}^1$	Retrieval measures					
	NN	FT	ST	E	DCG	aPR
BoW	0.8186	0.4460	0.5650	0.4169	0.8015	8.5131
50	0.8628	0.5403	0.6928	0.5125	0.8593	9.6268
100	0.8717	0.5461	0.7034	0.5171	0.8626	9.7181
200	0.8584	0.5371	0.6896	0.5074	0.8571	9.6049
300	0.8783	0.5440	0.7028	0.5154	0.8613	9.6503
400	**0.8827**	**0.5519**	**0.7152**	**0.5276**	**0.8666**	**9.7609**
500	0.8562	0.5134	0.6762	0.4894	0.8518	9.5377
600	0.8650	0.5093	0.6624	0.4813	0.8471	9.4514
700	0.8407	0.4633	0.6190	0.4457	0.8245	8.9339

TABLE III

THE AREA UNDER PR CURVES WITH DIFFERENT $\mathcal{R}$ UNDER MCGILL DATASET. $\mathcal{D}^1 = 400$, $\mathcal{H} = 1$

$\mathcal{R}$ \ $\mathcal{V}$	40	80	120	160	200
BoW	8.1665	8.4145	8.2775	8.5131	8.2592
1	8.4120	8.5725	8.6903	9.3209	8.6806
2	8.5541	9.0109	9.1015	9.4721	9.1472
3	**8.7018**	**9.2464**	**9.2908**	**9.7609**	**9.3597**

TABLE IV

THE MEASURES OF SHAPE RETRIEVAL WITH DIFFERENT $\mathcal{H}$ UNDER MCGILL DATASET. $\mathcal{V} = 160$, $\mathcal{D}^1 = 400$, $\mathcal{R} = 3$

$\mathcal{D}^2$	Retrieval measures					
	NN	FT	ST	E	DCG	aPR
$\mathcal{D}^1 = 400$	**0.8827**	**0.5519**	**0.7152**	**0.5276**	**0.8666**	9.7609
50	0.8805	0.5454	0.6954	0.5139	0.8639	**9.7867**
100	0.8805	0.5333	0.6822	0.5030	0.8564	9.6196
150	0.8695	0.5378	0.6866	0.5064	0.8591	9.7141

conducted under different sizes of virtual dictionary, $\mathcal{V} = \{40, 80, 120, 160, 200\}$. The results with different $\mathcal{V}$ are compared in terms of aPR in Table III, where BoW is regarded as a base line. These results imply that the spatial information encoded by DS is able to significantly increase the discriminability of BoW features with different $\mathcal{V}$ and $\mathcal{R}$. In addition, the more neighboring face rings employed, the more spatial information encoded, which obtains the better results. Thus, $\mathcal{R} = 3$ is adopted in the subsequent experiments.

4) The Number of Forward Layers $\mathcal{H}$: In this experiment, only $\mathcal{H} = 2$ is explored, where results with three candidate numbers of nodes in the second forward layer are compared, $\mathcal{D}^2 = \{50, 100, 150\}$. This is because overfitting is observed with $\mathcal{H} = 2$. As shown in Table IV, the best result is achieved with only one forward layer ($\mathcal{D}^1 = 400$). Although the results do not keep improved by the increasing $\mathcal{H}$ in this experiment, it would be helpful to increase the learning ability of DS for larger dataset via increasing $\mathcal{H}$. In the following experiments, $\mathcal{H} = 1$ and $\mathcal{D}^1 = 400$ are adopted to establish forward layers.

B. Global Shape Retrieval

In the following global shape retrieval experiments, DS is compared with several state-of-the-art methods of encoding the spatial relationship among virtual words for 3D shapes,

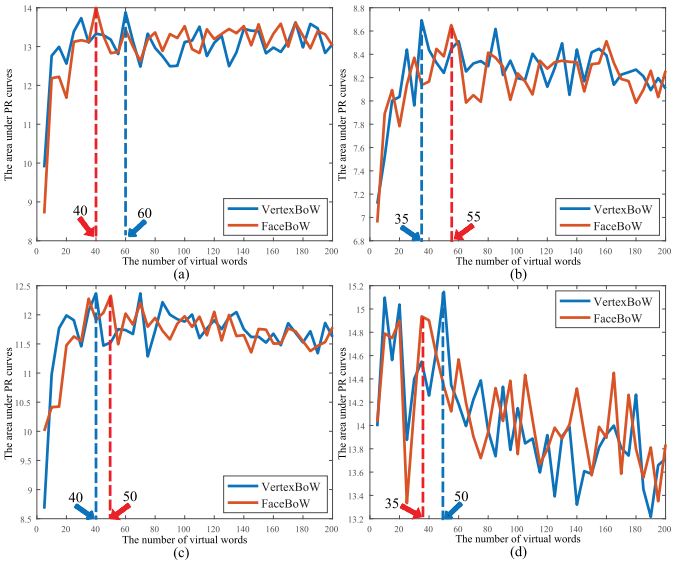

Fig. 8. The comparison of area under PR curves for choosing the optimal vertex and face virtual dictionaries under different 3D shape benchmarks. (a) LabeledPSB. (b) McGill. (c) SHREC2007. (d) SHREC2010.

including SSBoF [14], ISPM [13], bag of spatial context (BoSC) [16] and BoSCC [16]. In addition, BoW is also regarded as a baseline.

The experiments are carried out under several well-known 3D shape benchmarks, including LabeledPSB [23], McGill 3D shape benchmark, SHREC2007 dataset [27] and SHREC2010 dataset [28].

The compared methods are based on different kinds of virtual dictionaries. In SSBoF, ISPM, BoSC and BoSCC, the virtual dictionary is learned from low-level features of vertices. In the proposed DS, the virtual dictionary is learned from low-level features of faces. For a fair comparison, each method employs the optimal vertex virtual dictionary or the optimal face virtual dictionary that can be obtained in each benchmark. Specifically, the optimal vertex virtual dictionary is selected from some candidate ones in terms of the aPR obtained by BoW based on these candidate ones under global shape retrieval, where the candidate ones are learned from vertex features with different sizes, respectively. In the same way, the optimal face virtual dictionary is obtained. For clarity, BoW based on the optimal vertex or face virtual dictionary is named as "VertexBoW" or "FaceBoW", respectively.

The candidate virtual dictionaries are learned with $\mathcal{V} \in \{5, 10, 15, \ldots, 200\}$, respectively. In Fig. 8, the aPRs of VertexBoW and FaceBoW under different $\mathcal{V}$ are shown in each benchmark, where the blue and red dash lines indicate the sizes of optimal vertex and face virtual dictionaries, respectively.

With the optimal virtual dictionaries in each benchmark, DS is compared with the state-of-the-art methods in different 3D shape benchmarks as shown in Table V, where the VertexBoW and FaceBoW are denoted as V.BoW and F.BoW for short, respectively. ISPM with hard pooling and soft pooling are compared separately, and they are denoted as ISPM.H and ISPM.S for short. In addition, the median distance among distances between pairwise types of virtual words is regarded as the standard deviation of Gaussian function

TABLE V

THE MEASURES OF SHAPE RETRIEVAL FOR DIFFERENT METHODS UNDER DIFFERENT 3D SHAPE BENCHMARKS

Benchmarks	Methods	Retrieval measures				
		NN	FT	ST	E	DCG
LabeledPSB	V.BoW	0.9184	0.6089	0.7260	0.5159	0.8764
	F.BoW	0.9237	0.6198	0.7380	0.5229	0.8783
	SSBoF	0.6526	0.3424	0.5094	0.3399	0.7490
	ISPM.H	0.9261	0.6587	0.7670	0.5457	0.9029
	ISPM.S	0.8496	0.5296	0.6824	0.4714	0.8511
	BoSC	0.8895	0.6101	0.7332	0.5173	0.8787
	BoSCC	0.9000	0.6137	0.7532	0.5276	0.8843
	DS	**0.9316**	**0.6798**	**0.8096**	**0.5731**	**0.9110**
McGill	V.BoW	0.8119	0.4524	0.6038	0.4320	0.8112
	F.BoW	0.8319	0.4643	0.6012	0.4351	0.8133
	SSBoF	0.5708	0.3864	0.5492	0.3819	0.7514
	ISPM.H	0.9290	0.5866	0.7432	0.5510	0.8862
	ISPM.S	0.8758	0.5090	0.6496	0.4761	0.8435
	BoSC	0.8872	0.5199	0.6718	0.4892	0.8524
	BoSCC	0.8695	0.5270	0.6856	0.5001	0.8559
	DS	**0.9314**	**0.5880**	**0.7520**	**0.5512**	**0.8907**
SHREC2007	V.BoW	0.8775	0.5337	0.6574	0.4607	0.8454
	F.BoW	0.8800	0.5372	0.6618	0.4614	0.8446
	SSBoF	0.6775	0.4205	0.5736	0.3922	0.7754
	ISPM.H	0.8897	0.5872	0.7140	0.5005	0.8734
	ISPM.S	0.8471	0.5076	0.6628	0.4589	0.8407
	BoSC	0.8725	0.5612	0.7042	0.4882	0.8653
	BoSCC	0.8700	0.5650	0.6994	0.4898	0.8646
	DS	**0.9250**	**0.6095**	**0.7464**	**0.5250**	**0.8879**
SHREC2010	V.BoW	0.9394	0.7152	0.8780	0.6130	0.9356
	F.BoW	0.9444	0.6896	0.8598	0.5915	0.9317
	SSBoF	0.9141	0.6302	0.8358	0.5775	0.9091
	ISPM.H	0.9695	0.7866	0.9416	0.6593	0.9606
	ISPM.S	0.9543	0.7273	0.8570	0.6080	0.9370
	BoSC	0.9545	0.7216	0.8902	0.6226	0.9418
	BoSCC	0.9596	0.7358	0.8908	0.6252	0.9445
	DS	**0.9708**	**0.7886**	**0.9436**	**0.6608**	**0.9650**

TABLE VI

THE MEASURES OF SHAPE RETRIEVAL FOR DIFFERENT METHODS UNDER SHREC2007 DATASET

Methods	Retrieval measures			
	NN	FT	ST	DCG
DS	**0.9325**	0.6053	**0.7486**	**0.8896**
Tabia14	0.930	**0.623**	0.737	0.864
Lavoue12	0.918	0.590	0.734	0.841
Tabia11	0.853	0.527	0.639	0.719

involved in ISPM.S and SSBoF. Moreover, the high-level virtual dictionary employed by BoSC and BoSCC is learned with the same size of the optimal vertex virtual dictionary. The bold numbers show that DS outperforms all the state-of-the-art methods of encoding relationships among virtual words for 3D shapes under all shape benchmarks.

DS is also compared with some other state-of-the-art methods under SHREC2007, including the hybrid BoW of Lavoue [29], the curve based method of Tabia *et al.* [30], the BoW method of Toldo *et al.* [31] and covariance descriptors of Tabia *et al.* [5]. With $\mathcal{D}^1 = 300$, we obtain the better result than the one shown in Table V. As shown in Table VI, DS outperforms all other methods.

C. Classification

The performance of DS is further evaluated for shape classification under the same four shape benchmarks involved in global shape retrieval. Specifically, binary support vector machine (SVM) is employed to resolve the multi-class classification problem using the one-versus-one coding design with 10-fold cross validation. For this task, the classification accuracy is measured using different numbers of training samples, such as 50%, 60%, 70%, 80% and 90% shapes which are randomly sampled from each class, and accordingly, the remaining shapes in each class are regarded as the testing samples.

TABLE VII

THE CLASSIFICATION RESULTS FOR DIFFERENT METHODS UNDER DIFFERENT 3D SHAPE BENCHMARKS

Benchmarks	Methods	Training shape ratios				
		50%	60%	70%	80%	90%
LabeledPSB	V.BoW	81.6	76.3	78.9	69.7	84.2
	F.BoW	75.3	75.7	74.6	77.6	76.4
	SSBoF	52.1	44.1	43.9	46.1	47.4
	ISPM.H	83.2	86.1	87.7	78.9	89.5
	ISPM.S	75.8	76.8	75.4	75.0	78.9
	BoSC	71.6	71.1	68.4	71.1	68.4
	BoSCC	74.2	72.4	72.8	76.3	79.0
	DS	**92.6**	**90.1**	**94.7**	**90.8**	**92.1**
McGill	V.BoW	34.8	42.7	49.7	41.3	54.9
	F.BoW	30.5	39.8	48.4	41.3	48.2
	SSBoF	12.4	13.1	13.6	10.6	13.6
	ISPM.H	48.5	55.9	65.7	64.9	68.5
	ISPM.S	80.7	85.5	83.9	81.4	87.0
	BoSC	45.4	55.4	65.8	57.6	67.1
	BoSCC	41.5	48.5	51.8	48.4	65.7
	DS	**90.1**	**93.0**	**93.7**	**86.6**	**96.3**
SHREC2007	V.BoW	64.0	65.0	63.3	73.8	70.0
	F.BoW	68.5	65.0	67.5	65.0	70.0
	SSBoF	36.5	39.4	30.8	37.5	37.5
	ISPM.H	80.5	86.8	88.2	83.8	90.0
	ISPM.S	72.5	74.8	75.6	71.3	67.5
	BoSC	72.5	76.3	77.5	82.5	80.0
	BoSCC	68.5	68.8	71.7	76.3	77.5
	DS	**93.0**	**90.6**	**93.3**	**92.5**	**92.5**
SHREC2010	V.BoW	77.8	81.4	81.2	78.5	80.3
	F.BoW	79.8	77.5	74.4	78.5	75.8
	SSBoF	74.7	78.8	69.3	75.9	75.8
	ISPM.H	86.7	80.0	84.7	82.1	80.0
	ISPM.S	83.7	77.5	84.7	82.1	80.0
	BoSC	80.8	81.4	85.0	80.7	84.8
	BoSCC	80.8	81.4	77.8	81.0	80.3
	DS	**99.0**	**96.3**	**91.7**	**97.5**	**100.0**

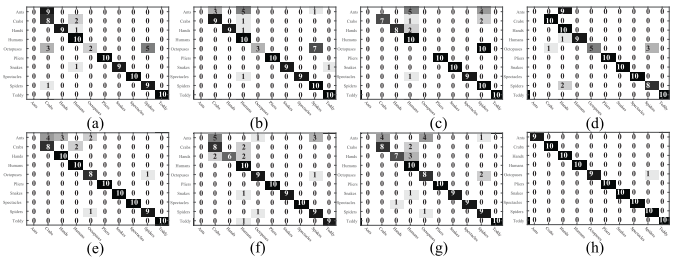

Fig. 9. The comparison shown by the confusion matrices with 50% training samples under SHREC2010 benchmark for (a) VertexBoW, (b) FaceBoW, (c) SSBoF, (d) ISPM.H, (e) ISPM.S, (f) BoSC, (g) BoSCC and (h) DS.

The classification accuracy listed in Table VII indicates that the performance of DS is the best among state-of-the-art methods. In Fig. 9, the confusion matrices obtained with 50% training shapes under SHREC2010 benchmark are compared to further highlight the discriminability of DS.

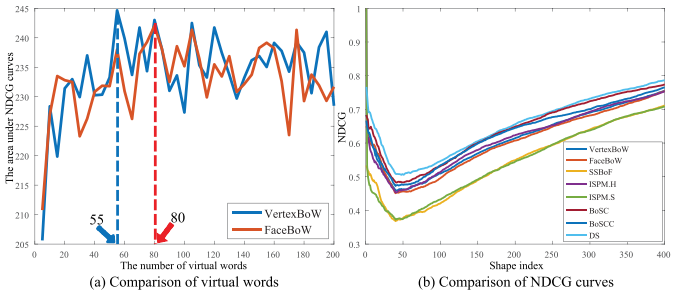

Fig. 10. (a) The comparison of area under NDCG for choosing the optimal vertex and face virtual dictionaries. (b) The comparison of NDCG obtained by different methods.

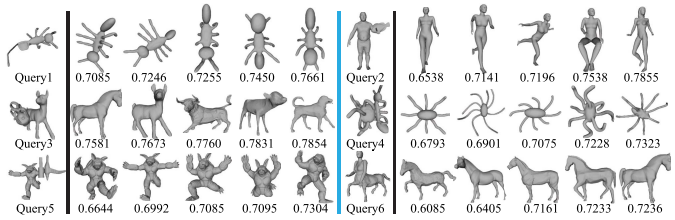

Fig. 11. The top-5 retrieved shapes of 6 queries retrieved by deep spatiality. The cosine distance between the retrieved shape and the query is also shown.

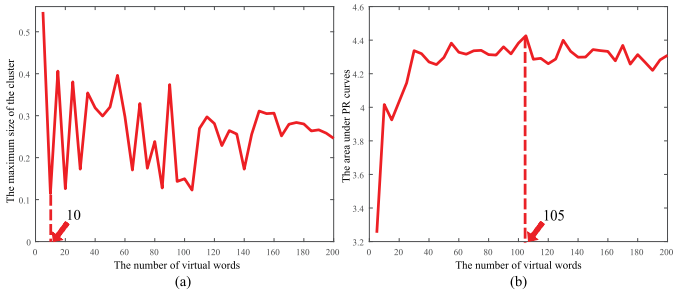

Fig. 12. The comparison for choosing the optimal face virtual dictionary for (a) SCAPE dataset and (b) Watertight07 dataset.

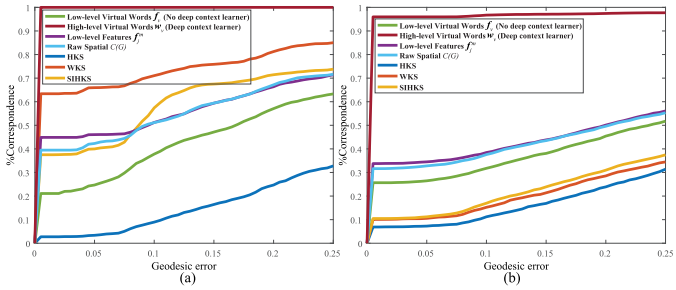

Fig. 13. The comparison of correspondence obtained by different methods under (a) SCAPE dataset and (b) Watertight07 dataset.

D. Partial Shape Retrieval

DS is evaluated on partial shape retrieval under the SHREC2007 partial retrieval dataset [27], and the quantitative evaluation is measured by the normalized discounted cumulated gain vector (NDCG) [32].

Similar to global shape retrieval, the optimal vertex or face virtual dictionary is selected from some candidate ones in terms of the performance of VertexBoW or FaceBoW under partial shape retrieval, respectively. In Fig. 10 (a), the area under NDCG curves with $\mathcal{V} \in \{5, 10, 15, \ldots, 200\}$ is compared. Then, these two optimal virtual dictionaries for VertexBoW and FaceBoW are selected as the one with $\mathcal{V} = 55$ and the one with $\mathcal{V} = 80$, respectively.

Based on the optimal virtual dictionaries, DS is compared with VertexBoW, FaceBoW, SSBoF, ISPM.H, ISPM.S, BoSC and BoSCC. As shown in Fig. 10 (b), DS outperforms all other methods. The high performance is further demonstrated by the top-5 retrieved shapes of 6 queries, as shown in Fig. 11, where the number under each retrieved shape is the cosine distance between the retrieved shape and the query. Note that the top-5 retrieved shapes are all with high relevance to the query.

E. Shape Correspondence

DS is further evaluated on shape correspondence to demonstrate the discriminability of learned local features. This experiment is conducted under two datasets with well-defined ground truth, such as the SCAPE [33] and the Watertight07 dataset [34].

The vertices sampled from one shape $\boldsymbol{M}^1$ are matched to the ones sampled from another shape $\boldsymbol{M}^2$ in the same class, where $\boldsymbol{M}^1$ and $\boldsymbol{M}^2$ with the subsequent index form a matching pair. Given a matching pair, the error measure $d(f, f_{true})$ [34] evaluates the accuracy of a predicted shape correspondence $f : \boldsymbol{M}^1 \rightarrow \boldsymbol{M}^2$ with respect to the ground truth f_{true}. Specifically, $d(f, f_{true})$ is all the geodesic distances between the matched vertices and the ground truth. Then, the statistics of errors of all matching pairs over the benchmark is presented under several geodesic error thresholds. Moreover, the simple Hungarian algorithm is used to match sampled vertices, aiming to explore the raw discriminability of learned local features.

For fair comparison, the local feature of each sampled vertex is formed by averaging the features of faces in the first neighboring ring of the sampled vertex. The feature of face can be the low-level feature $\boldsymbol{f}_j^m$ of the face, the low-level feature $\boldsymbol{f}_v$ of the type of virtual word that the face belongs to, the raw spatial representation $C(G)$, and the high-level feature $\boldsymbol{w}_v$ of the type of virtual word that the face belongs to. In addition, the state-of-the-art local descriptors for vertex are also employed to compute the features of the sampled vertices to compare with DS, such as HKS [21], WKS [35] and SIHKS [36].

In SCAPE, there is only one shape class. Therefore, the optimal face virtual dictionary is selected with the minimum scatter degree, as indicated by $\mathcal{V} = 10$ in Fig. 12 (a). In Watertight07, we use the same manner for global shape retrieval to select the optimal face virtual dictionary, as shown in Fig. 12 (b), where $\mathcal{V} = 105$.

As shown in Fig. 13, DS achieves the best performance among all features including the state-of-the-art local features in both datasets. In addition, the comparison also implies that $C(G)$ is able to capture the spatial relationship among virtual words in local region, since the results of $C(G)$ outperforms the results of low-level features of virtual words $\boldsymbol{f}_v$ in both datasets. Subsequently, DS further learns $\boldsymbol{w}_v$ from

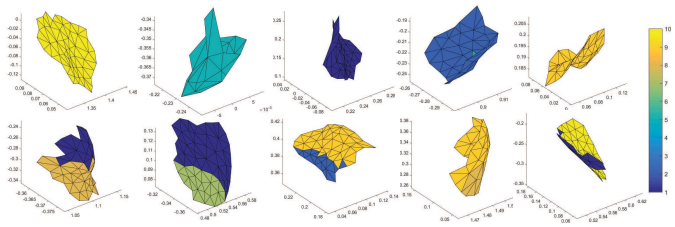

Fig. 14. Some learned spatial context patterns in the first hidden layer of DS under SCAPE dataset. The size of optimal face virtual words is 10.

$C(G)$ and $\boldsymbol{f}_v$, which is able to obtain higher performance than the one merely obtained by $C(G)$ in both datasets. Moreover, $C(G)$ outperforms HKS and is comparable to SIHKS in SCAPE, and also outperforms HKS, SIHKS and WKS in Watertight07.

The comparison in Fig. 13 also shows the powerful context learning ability of deep context learner. In both datasets, the low-level features of virtual words $\boldsymbol{f}_v$ always perform worse than the low-level features $\boldsymbol{f}_j^m$ of faces. Although the statistics of $\boldsymbol{f}_v$ enables $\boldsymbol{f}_v$ to represent global shapes through BoW, $\boldsymbol{f}_v$ are merely the centers of $\boldsymbol{f}_j^m$. Therefore, the discriminative ability of $\boldsymbol{f}_v$ is degenerated in representing local regions of $\boldsymbol{f}_j^m$. By encoding the spatial context around central virtual word, raw spatial representation $C(G)$ is able to remedy this ability degeneration of $\boldsymbol{f}_v$. However, $C(G)$ performs merely comparable to $\boldsymbol{f}_j^m$ in both datasets. On the other hand, deep context learner can simultaneously learn the spatial context between the central virtual word and the virtual words in neighboring face rings, and also the context between local regions and global shapes. In addition, deep context learner further enhances each other context in the joint learning procedure. This context learning ability enables deep context learner to significantly increase the discriminative ability of representing local regions of virtual words (as shown by the outperforming performances of high-level features $\boldsymbol{w}_v$ of virtual words in both datasets). Correspondingly, the discriminability of learned global features is also increased (as shown by the performances of DS in the former experiments).

The spatial context patterns learned in the first forward layer of DS under SCAPE dataset are briefly visualized. As shown in Fig. 14, each local region is the one with the maximum activation to a specific weight vector which corresponds to a spatial context pattern.

VII. Limitation, Future Work and Conclusion

A. Limitation and Future Work

Although DS achieves high performance on various shape analysis applications, there are still four limitations. First, DS requires to analyse the neighboring relationship between faces so that it is only suitable for 3D manifold meshes rather than other 3D forms, such as point sets and volumetric shapes. Second, DS is a framework based on the virtual dictionary. Therefore, the performance of DS is affected by the discriminability of different types of virtual words. Third, DS has to learn spatiality in an unsupervised way since it starts from BoW which is an unsupervised scenario. Thus, DS can only achieve comparable results compared to the supervised deep learning models for 3D shapes [37]–[39]. For example, the NN (0.9314) and DCG (0.8907) of DS are comparable to the NN (0.988) and DCG (0.955) of DeepShape [39] under McGill benchmark in shape retrieval. Fourth, the overfitting issue of training deep context learner is also eager to be resolved, which could further increase the discriminability of the learned global and local features.

In the future, we plan to resolve the overfitting issue by some state-of-the-art techniques, such as adaptive learning rate, dropout and L2 regularization of weights.

B. Conclusion

In this work, DS is proposed to simultaneously learn spatially-enhanced global and local 3D features in an unsupervised manner via encoding the spatial relationship among virtual words. DS effectively encodes the relative positions between pairwise virtual words along a consistent circular direction to stride the obstacles of 3D shapes including arbitrary mesh resolutions, irregular vertex topology, and orientation ambiguity on 3D surface. Based on a novel spatial context formalization, DS is proposed with two novel elements: spatial context extractor, and deep context learner. Spatial context extractor extracts the spatial information in a local region under newly proposed directed circular graphs, where the extracted spatial information forms a raw spatial representation. Over the raw spatial representations, the spatially-enhanced global and local features can be learned by deep context learner, in which a coupled softmax layer effectively guides the learning in an unsupervised manner. Finally, deep context learner encodes the discriminative information among different types of virtual words and the one among shapes. Experimental results demonstrate that DS significantly outperforms other compared state-of-the-art methods for various applications in 3D shape analysis under different shape benchmarks. The results therefore show that DS successfully encodes the rigid and non-rigid invariant spatial information of 3D shapes.

References

[1] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2006, pp. 2169–2178.

[3] G. Lavoué, "Bag of words and local spectral descriptor for 3D partial shape retrieval," in *Proc. 4th Eurographics Conf. 3D Object Retr.*, 2011, pp. 41–48.

[4] X. Bai, C. Rao, and X. Wang, "Shape vocabulary: A robust and efficient shape representation for shape matching," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3935–3949, Sep. 2014.

[5] H. Tabia, H. Laga, D. Picard, and P.-H. Gosselin, "Covariance descriptors for 3D shape matching and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4185–4192.

[6] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-bag-of-features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3352–3359.

[7] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1994–2008, May 2013.

[8] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, and X. Li, "Unsupervised 3D local feature learning by circle convolutional restricted Boltzmann machine," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5331–5344, Nov. 2016.
[9] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, and C. L. P. Chen, "Mesh convolutional restricted Boltzmann machines for unsupervised learning of features with structure preservation on 3-D meshes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2268–2281, Oct. 2017.
[10] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, and C. L. P. Chen, "Unsupervised learning of 3-D local features from raw voxels based on a novel permutation voxelization strategy," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2017.2778764.
[11] X. Li, A. Godil, and A. Wagan, "Spatially enhanced bags of words for 3D shape retrieval," in *Proc. Int. Symp. Vis. Comput.*, 2008, pp. 349–358.
[12] X. Li and A. Godil, "Exploring the bag-of-words method for 3D shape retrieval," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 437–440.
[13] C. Li and A. B. Hamza, "Intrinsic spatial pyramid matching for deformable 3D shape retrieval," *Int. J. Multimedia Inf. Retr.*, vol. 2, no. 4, pp. 261–271, 2013.
[14] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov, "Shape Google: Geometric words and expressions for invariant shape retrieval," *ACM Trans. Graph.*, vol. 30, no. 1, 2011, Art. no. 1.
[15] R. Wessel and R. Klein, "Learning the compositional structure of man-made objects for 3D shape retrieval," in *Proc. Eurographics Workshop 3D Object Retr.*, 2010, pp. 39–46.
[16] Z. Han *et al.*, "BoSCC: Bag of spatial context correlations for spatially enhanced 3D shape representation," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3707–3720, Aug. 2017.
[17] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.
[18] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.
[19] Y. Liu, H. Zha, and H. Qin, "Shape topics: A compact representation and new algorithms for 3D partial shape retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2025–2032.
[20] Z. Lian, A. Godil, and X. Sun, "Visual similarity based 3D shape retrieval using bag-of-features," in *Proc. Shape Model. Int. Conf.*, Jun. 2010, pp. 25–36.
[21] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.
[22] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Proc. Eurographics Symp. Geometry Process.*, 2003, pp. 156–164.
[23] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3D mesh segmentation and labeling," *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art. no. 102.
[24] G. Peyré and L. D. Cohen, "Geodesic remeshing using front propagation," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 145–156, 2006.
[25] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson, "Retrieving articulated 3-D models using medial surfaces," *Mach. Vis. Appl.*, vol. 19, no. 4, pp. 261–275, 2008.
[26] A. Godil, Z. Lian, H. Dutagaci, R. Fang, T. P. Vanamali, and C. P. Cheung, "Benchmarks, performance evaluation and contests for 3D shape retrieval," in *Proc. 10th Perform. Metrics Intell. Syst. Workshop*, 2010, pp. 42–47.
[27] D. Giorgi, S. Biasotti, and L. Paraboschi, "Shape retrieval contest 2007: Watertight models track," in *Proc. SHREC Competition*, 2007, pp. 1–11.
[28] Z. Lian *et al.*, "SHREC'10 track: Non-rigid 3D shape retrieval," in *Proc. Eurographics Workshop 3D Object Retr.*, 2010, pp. 101–108.
[29] G. Lavoué, "Combination of bag-of-words descriptors for robust partial shape retrieval," *Vis. Comput.*, vol. 28, no. 9, pp. 931–942, Sep. 2012.
[30] H. Tabia, M. Daoudi, J.-P. Vandeborre, and O. Colot, "A new 3D-matching method of nonrigid and partially similar models using curve analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 852–858, Apr. 2011.
[31] R. Toldo, U. Castellani, and A. Fusiello, "Visual vocabulary signature for 3D object retrieval and partial matching," in *Proc. Eurographics Conf. Workshop 3D Object Retr.*, 2009, pp. 21–28.
[32] S. Marini, L. Paraboschi, and S. Biasotti, "Shape retrieval contest 2007: Partial matching track," in *Proc. SHREC Conjunct IEEE Shape Model. Int.*, Jun. 2007, pp. 13–16.
[33] D. Anguelov, P. Srinivasan, H.-C. Pang, D. Koller, S. Thrun, and J. Davis, "The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces," in *Proc. Neural Inf. Process. Syst.*, 2004, pp. 33–40.
[34] V. G. Kim, Y. Lipman, and T. Funkhouser, "Blended intrinsic maps," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 79:1–79:12, 2011.
[35] M. Aubry, U. Schlickewei, and D. Cremers, "The wave kernel signature: A quantum mechanical approach to shape analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 1626–1633.
[36] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1704–1711.
[37] Y. Fang *et al.*, "3D deep shape descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2319–2328.
[38] J. Xie, Y. Fang, F. Zhu, and E. Wong, "Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1275–1283.
[39] J. Xie, G. Dai, F. Zhu, E. K. Wong, and Y. Fang, "Deepshape: Deep-learned shape descriptor for 3D shape retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1335–1345, Jul. 2017.

Zhizhong Han is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with Northwestern Polytechnical University, China. His current research interests include machine learning, pattern recognition, feature learning, and digital geometry processing.

Zhenbao Liu (M'11) received the Ph.D. degree from the College of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan, in 2009. In 2012, he was a Visiting Scholar with Simon Fraser University, Canada. He is currently a Professor with Northwestern Polytechnical University, China. He has published approximately 50 papers in major international journals and conferences. His current research interests include pattern recognition, computer vision, and shape analysis.

Chi-Man Vong (M'09–SM'14) received the M.S. and Ph.D. degrees in software engineering from the University of Macau in 2000 and 2005, respectively. He is currently an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau. His current research interests include machine learning methods and intelligent systems.

Yu-Shen Liu (M'18) received the B.S. degree in mathematics from Jilin University, China, in 2000, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 2006. From 2006 to 2009, he was a Post-Doctoral Researcher with Purdue University. He is currently an Associate Professor with the School of Software, Tsinghua University, Beijing, China. His current research interests include shape analysis, pattern recognition, machine learning, and semantic search.

Shuhui Bu (M'09) received the master's and Ph.D. degrees from the College of Systems and Information Engineering, University of Tsukuba, Japan, in 2006 and 2009, respectively. From 2009 to 2011, he was an Assistant Professor with Kyoto University, Japan. He is currently a Professor with Northwestern Polytechnical University, China. He has published approximately 40 papers in major international journals and conferences. His current research interests are concentrated on computer vision and robotics.

Junwei Han (M'12–SM'15) received the Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, in 2003. He is a currently a Professor with Northwestern Polytechnical University, Xi'an, China. His current research interests include multimedia processing and brain imaging analysis. He is an Associate Editor of the IEEE TRANSACTION ON HUMAN-MACHINE SYSTEMS, *Neurocomputing*, and *Multidimensional Systems and Signal Processing*.

C. L. Philip Chen (S'88–M'88–SM'94–F'07) received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. He was a Tenured Professor, the Department Head, and the Associate Dean with two different U.S. universities for 23 years. He is currently the Dean of the Faculty of Science and Technology with the University of Macau, Macau, China, where he is also a Chair Professor with the Department of Computer and Information Science.