International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01592-x

Unsupervised Multi-View CNN for Salient View Selection and 3D Interest Point Detection

Ran Song[1,2] · Wei Zhang[1,2] · Yitian Zhao[3] · Yonghuai Liu[4]

Received: 9 February 2021 / Accepted: 1 February 2022

Abstract
We present an unsupervised 3D deep learning framework based on a ubiquitously true proposition named by us view-object consistency as it states that a 3D object and its projected 2D views always belong to the same object class. To validate its effectiveness, we design a multi-view CNN instantiating it for salient view selection and interest point detection of 3D objects, which quintessentially cannot be handled by supervised learning due to the difficulty of collecting sufficient and consistent training data. Our unsupervised multi-view CNN, namely UMVCNN, branches off two channels which encode the knowledge within each 2D view and the 3D object respectively and also exploits both intra-view and inter-view knowledge of the object. It ends with a new loss layer which formulates the view-object consistency by impelling the two channels to generate consistent classification outcomes. The UMVCNN is then integrated with a global distinction adjustment scheme to incorporate global cues into salient view selection. We evaluate our method for salient view section both qualitatively and quantitatively, demonstrating its superiority over several state-of-the-art methods. In addition, we showcase that our method can be used to select salient views of 3D scenes containing multiple objects. We also develop a method based on the UMVCNN for 3D interest point detection and conduct comparative evaluations on a publicly available benchmark, which shows that the UMVCNN is amenable to different 3D shape understanding tasks.

Keywords Unsupervised 3D deep learning · Multi-view CNN · View-object consistency · View selection · 3D interest point detection

1 Introduction

Unsupervised deep learning has demonstrated its great value and impact in many tasks. One important reason is that the manual collection and annotation of a large dataset for training a deep neural network in a supervised manner is usually laborious. This is particularly the case for 3D tasks where data collection and annotation are generally more challenging than those in 2D tasks. For instance, the ground truth generation of 3D interest point detection on 3D objects is more time-consuming than that of 2D interest point detection on 2D images since human subjects need to rotate a mesh to mark the points of interest (Dutagaci et al. 2012) and may also have to rotate it forward or backward again after the

Communicated by A. Hilton.

This work was supported by the National Natural Science Foundation of China under Grants 62076148, 61991411 and U1913204, the Young Taishan Scholars Program of Shandong Province No.tsqn201909029 and the Qilu Young Scholars Program of Shandong University.

✉ Wei Zhang
davidzhang@sdu.edu.cn

Ran Song
ransong@sdu.edu.cn

Yitian Zhao
yitian.zhao@nimte.ac.cn

Yonghuai Liu
liuyo@edgehill.ac.uk

[1] School of Control Science and Engineering, Shandong University, Jinan, China

[2] Institute of Brain and Brain-Inspired Science, Shandong University, Jinan, China

[3] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China

[4] Department of Computer Science, Edge Hill University, Ormskirk, UK

Published online: 16 March 2022

first round of marking to check if some interest points are not marked due to occlusion. Therefore, a widely applicable unsupervised 3D deep learning framework is potentially of broad interest in the fields of 3D computer vision, computer graphics and 3D machine learning.

A simple but ubiquitously true proposition is that a 3D object and its projected 2D views always belong to the same object class no matter what taxonomy is applied to the classification. We name the proposition *view-object consistency* and propose a novel unsupervised 3D deep learning framework based on it. Essentially, the unsupervised framework learns a meaningful embedding of a 3D object optimised for the view-object consistency but not necessarily the most distinctive features for its classification subject to a set of explicit categorical labels.

Since it is not feasible for us to solidly and thoroughly explore the utility of the framework through various 3D tasks in one paper, we intentionally pick salient view selection and interest point detection of 3D objects to demonstrate its effectiveness for three reasons. First, both tasks are challenging as they do not only rely on low-level geometric features but also involve complex high-level semantic understandings of the objects. Thus a data-driven method is naturally sound. Second, however, they are the particular tasks where collecting a large amount of accurately and consistently annotated data is notoriously difficult. We found that all existing datasets for the two tasks are very small (e.g. 68 objects in Dutagaci et al. (2010) and 16 objects in Secord et al. (2011) for salient view selection and 43 objects in Dutagaci et al. (2012) for 3D interest point detection) no matter whether the annotations were collected directly (e.g. by asking human subjects to mark a viewpoint on a view sphere surrounding the object (Dutagaci et al. 2010) or a group of interest points on the surface of the object (Dutagaci et al. 2012) or indirectly (e.g. by paired comparisons where subjects were asked to select the preferred view from two views for multiple times (Secord et al. 2011). Third, we shall further show the advantage of an unsupervised method by extending salient view selection to 3D scenes. Salient view selection of 3D scenes can hardly be addressed by a weakly supervised method relying on such annotation as a single class label because a scene often contains objects belonging to different classes.

The problem of salient view selection of 3D objects is arguably well defined. Besides the chunk of related literatures in computer vision and graphics that will be discussed in Sect. 2, researchers in psychology (Cutzu and Edelman 1994; Blanz et al. 1999) have revealed that for many classes of familiar objects, the preferred views are reasonably consistent among human subjects. To make it clear, the most salient view of a 3D object herein is defined as the view that a human subject likes most for whatever reason. And we shall evaluate our method using the publicly available benchmark (Dutagaci et al. 2010) where subjects were asked to rotate a 3D object to directly select the view that they preferred. We also show that our method can be directly extended to 3D scenes composed of multiple 3D objects where each object typically has its own best view when appearing independently. Salient view selection of 3D scenes has a range of applications such as virtual scene understanding, panoramic scene synthesis, ray tracing optimisation and camera path planning (Zhang and Fei 2019). Since there is no available ground truth, we conduct a user study to evaluate our method and demonstrate that it achieves a good performance on selecting salient views of various 3D scenes.

To instantiate the proposition of view-object consistency in the context of both 3D deep learning and salient view selection, we develop a multi-view convolutional neural network (CNN). It formulates the view-object consistency through a two-channel architecture and a newly designed loss function. It also integrates with an important heuristic of human's view preference via a specifically designed layer. The proposed multi-view CNN is trained end-to-end in an unsupervised manner using only a collection of 3D objects without any manual annotations and is thus named as Unsupervised Multi-View CNN (UMVCNN). It exploits both intra-view and inter-view knowledge via a multi-view representation of 3D objects and scenes for salient view selection. Such intra-view knowledge is inherently local as it is based fully on the information within a particular view. Such inter-view knowledge is hardly global as well because it is based only on pairwise distinction of views. However, salient view selection of 3D objects and scenes obviously involves the global understanding for them. Therefore, we further present a global distinction adjustment (GDA) scheme by exploiting the deep features extracted through the learned UMVCNN. The GDA essentially investigates whether a local pairwise distinction is globally important or not. It is then integrated with the UMVCNN to further boost the performance of salient view selection.

To show that the proposed UMVCNN has wide applicability to 3D shape understanding tasks, we further explore its effectiveness on another downstream task, 3D interest point detection, which also has many applications, such as shape registration, mesh segmentation, mesh simplification, and object matching and retrieval (Dutagaci et al. 2012). For instance, using interest points for 3D object matching has the advantage of providing local features of both semantic significance and invariance to noise, rotation, deformation and articulation. In this work, we provide a simple method based on the UMVCNN to derive 3D interest points of various objects. It is noteworthy that we are concerned with the detection of the 3D points that most human subjects are interested in for whatever reasons. Thus it is with regard to the subjective and perceptual judgements of humans about 3D interest points and will be evaluated using human-generated ground truth data.

The contribution of our work is hence fourfold:

(1) We propose a novel unsupervised framework of 3D deep learning where the core idea is valid ubiquitously and thus potentially has a wide range of applications.
(2) We propose a new multi-view CNN, namely UMVCNN, in accordance with this unsupervised framework and integrate it with a GDA scheme for salient view selection of 3D objects.
(3) We extend salient view selection from individual 3D objects to scenes containing multiple 3D objects and demonstrate that the UMVCNN can select good views for various scenes via a user study.
(4) We show that the proposed UMVCNN can also be used for detecting the points that humans are perceptually interested in from a 3D object.

The rest of the paper is organised as follows. After briefly reviewing the related work in Sect. 2, we first introduce the details of the proposed UMVCNN in Sect. 3. Then we elaborate the GDA scheme for salient view selection based on the UMVCNN in Sect. 4. In Section 5, we exhibit the experimental results of salient view selection for both 3D objects and scenes containing multiple objects. We further demonstrate the effectiveness of the UMVCNN through its application for 3D interest point detection in Sect. 6. Finally, we draw the conclusions in Sect. 7.

A preliminary version of this work was published as a poster presentation in European Conference on Computer Vision (ECCV'20) (Song et al. 2020b)[1], which has been used as a baseline (i.e. UMVCNN-VGG in Tables 1 and 2) for comparisons in this paper.

2 Related Work

In this section, we review the literatures for salient view section and 3D interest point detection, respectively. Generally, each of them can be categorised into two groups: handcrafted and learning-based methods.

2.1 Salient View Selection

A number of methods for salient view selection are based only on the handcrafted attributes of 3D objects. Polonsky et al. (2005) explored general frameworks for view selection by analysing several handcrafted attributes associated with geometrical or statistical properties of a 3D object or its projected 2D views. Lee et al. (2005) selected salient views using the attribute of mesh saliency computed via Gaussian-weighted mean curvatures. Yamauchi et al. (2006) employed mesh saliency as the intra-view cue for finding salient views while taking into account such an inter-view cue as the similarity of projected views. Han et al. (2014) selected good views of 3D objects by first computing the saliency-based mesh segmentation and then ranking the viewpoints based on the segmentation results. Lienhard et al. (2014) used not only geometrical attributes but also aesthetic and semantic cues to find the good views for procedural 3D models. Leifman et al. (2016) computed a saliency measure based on both local geometrical and global topological attributes for salient view selection. However, most methods based on handcrafted attributes do not generalise well due mainly to the limited expressive capabilities of the attributes extracted by some fixed schemes for objects of different classes.

For the learning-based methods, we found that some of them are essentially shallow learning of a certain model to combine multiple attributes while all attributes are not learned but still handcrafted. Vieira et al. (2009) learned good views via an SVM classifier where the candidate views were represented by a collection of handcrafted attributes. To investigate human view preference, Secord et al. (2011) collected a small dataset to learn a regression model combining a list of handcrafted attributes. Mezuman and Weiss (2012) leveraged Internet images to learn the view from which we most often see the object, where the handcrafted GIST descriptor was employed to measure view similarity. Zhao et al. (2015) learned best views from hand-drawn sketches by asking participants to align a 3D model according to a given sketch. He et al. (2018) proposed a multi-view learning framework exploiting both 2D and 3D handcrafted attributes to assess and recommend viewpoints for photographing architectures.

Apart from the psychological work (Tarr and Pinker 1989; Cutzu and Edelman 1994; Hayward 1998), in computer vision, there is also evidence (Wu et al. 2015; Su et al. 2015; Novotny et al. 2017) of the relation between view selection and object recognition where view-dependent attributes were extracted via deep neural networks for 3D object recognition. Kim et al. (2017) and Song et al. (2020a) leveraged deep CNNs for salient view selection of objects instead of improving recognition accuracy. Our work is inspired by both of them but fundamentally different for two reasons: 1) both Kim et al. (2017) and Song et al. (2020a) require annotated data for training while our work is unsupervised with no need for data annotation; 2) both of them cannot be trained end-to-end where the former trains two CNNs and a Random Forest classifier separately and the latter trains a CNN and a Markov Random Field individually while our UMVCNN is trained fully end-to-end.

[1] Data, codes and the pretrained model are publicly available at https://github.com/rsong/UMVCNN.

2.2 3D Interest Point Detection

Early work on 3D interest point detection mostly relied on handcrafted attributes. Shilane and Funkhouser (2006) selected points that contribute to improving retrieval performance by assigning a predicted distinctiveness value to each selected point using a training model. Castellani et al. (2008) first defined a saliency measure by applying a Gaussian at the vertices. Then a scale space was constructed and vertices highly displaced after the filtering were marked as candidate points of interest. Zaharescu et al. (2009) assumed that the vertices of a 3D object have associated information such as curvature or photometric properties, and applied a difference of Gaussian on the function defined by the associated information for the detection. Mian et al. (2010) extracted scale-invariant key points and ranked them using a measure directly related to their repeatability and the distinctiveness of the underlying local descriptor. Song et al. (2013) detected 3D points of interest by the spectral irregularity diffusion which captures not only the geometric information about local neighbourhood of a given point in a multi-scale manner, but also cues related to the global structure of an object.

Recently, some deep learning-based methods have been proposed for 3D keypoint detection. However, we notice that most of these methods actually focused on the detection of the 3D points helpful for some particular tasks such as matching and registration. For instance, Zeng et al. (2017) presented a Siamese network that learned local geometric descriptors of keypoints for establishing correspondences between partial 3D data. Deng et al. (2018) proposed a deep neural network for 3D point matching by learning both local geometric features and global context-aware cues of 3D objects and scenes. Yew and Lee (2018) holistically learned a 3D feature detector and descriptor through a deep network for point cloud registration. Li and Lee (2019) presented a method that can detect highly repeatable and accurately localised keypoints from 3D point clouds in an unsupervised manner. Bai et al. (2020) proposed a convolutional network for the dense detection and description of 3D local features to achieve accurate and fast point cloud alignment.

We found that these methods based on deep learning just sought to improve some low-level criteria such as "repeatability" but did not aim to detect the points perceptually interesting to human subjects. As we mentioned above, collecting a large dataset with human-generated ground truth for 3D interest point detection is time-consuming as the subjects often have to rotate a 3D object forward and backward to find the interest points. Due to the lack of training data, such supervised methods are hardly applicable for detecting 3D points of interest in accordance with human perception.

3 Unsupervised Multi-View CNN

In this section, we first describe each component of our method in a piecewise manner. We then elaborate the implementation as a whole in both training and deployment modes where each component is situated in the context of the complete pipeline.

3.1 Multi-view Representation of 3D Objects

Multi-view CNNs have been widely used to adapt image-based deep networks to 3D objects where an object is represented as a selection of its projected views. Compared with other methods which generalise deep learning to non-Euclidean domains, multi-view CNNs showed state-of-the-art performance in various 3D shape understanding tasks (Su et al. 2015; Qi et al. 2016; Kalogerakis et al. 2017; Huang et al. 2018). One consensus among these tasks is that we should avoid using the very 'bad' views usually defined as the ones that cause misrecognition or misunderstanding of the objects. We propose a scheme considering two low-level attributes to ensure that the selected 2D views for representing a 3D object are at least 'not very bad'.

We start with an icosahedron to uniformly sample a view sphere surrounding the input 3D object. Then we iteratively subdivide the icosahedron to produce more vertices (i.e., viewpoints) on the view sphere. We end with a polyhedron with 162 vertices. Next, we rank the views taken from these viewpoints based on the attributes of *view area* and *silhouette length*. View area is calculated as the area of the projection of the object as seen from a particular viewpoint. Silhouette length is the length of the outer contour of the silhouette of the object as seen from a particular viewpoint. We collect the top N ($N = 20$ in this work) views with the highest ranks on average based on the two attributes as the multi-view representation of the 3D object.

3.2 UMVCNN Architecture

Overview Figure 1 illustrates the architecture of the proposed UMVCNN. It starts with the classic ResNet50 model (He et al. 2016) as the backbone and then branches off the view and the object channels after the global average pooling layers. Through the view distinction (VD) layer, it generates an inter-view heuristic using the deep features extracted from the 2D views. A weighted sum pooling (WSP) layer is then employed to incorporate this heuristic and multiple intra-view features derived from each individual view into a single tensor encoding the information corresponding to the entire 3D object. These two layers and the newly added fully connected layer Fc2 followed by a Softmax normalisation form the object channel. It outputs to the loss layer a vector composed of the probabilities of the 3D object belonging to

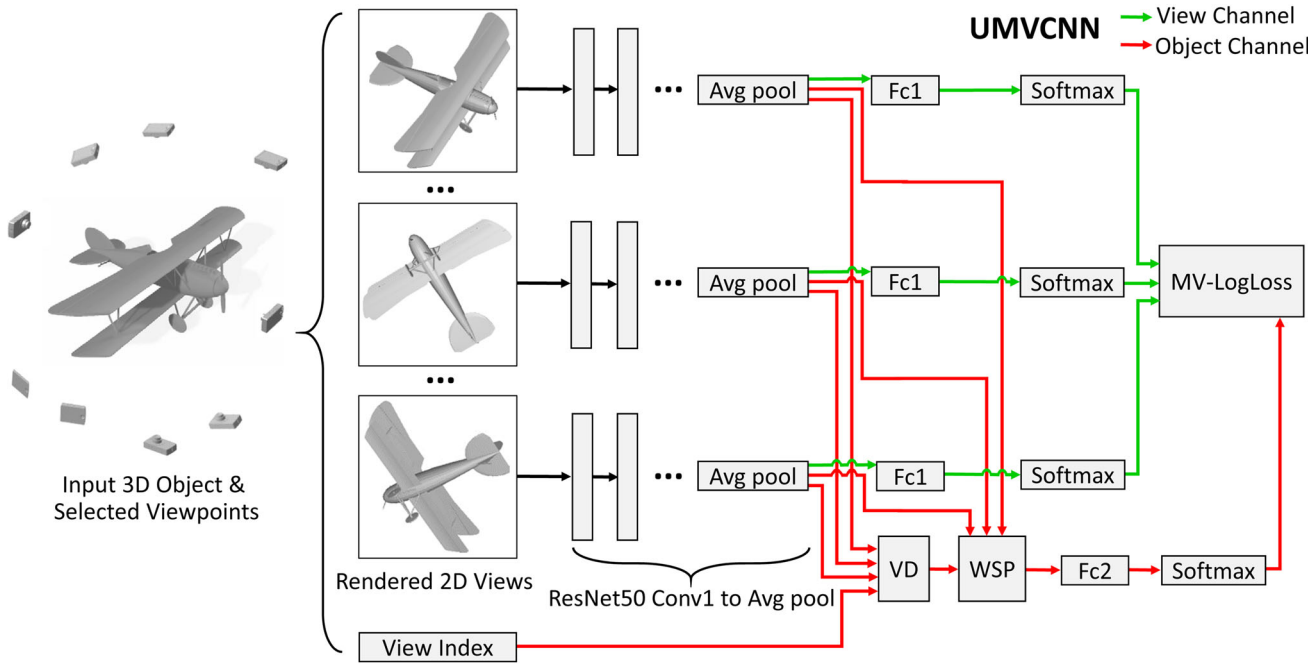


Fig. 1 Overview of the proposed UMVCNN containing two channels. The green and the red arrows denote the view channel and the object channel respectively. "VD" and "WSP" denote the view distinction and the weighted sum pooling layers respectively (Color figure online)

a certain class. On the other hand, we also add a fully connected layer Fc1 in the view channel that generates a vector for each view predicting which class the view belongs to. Every trainable ResNet50 layer from Conv1 to the average pooling layer in the UMVCNN shares the same weights for all views. Finally, the outputs of the view and the object channels converge at the newly designed Multi-View Logistic Loss (MV-LogLoss) layer which formulates the view-object consistency principle to enable an unsupervised learning.

View Distinction (VD) Layer Existing work (Yamauchi et al. 2006; Secord et al. 2011; Zhao and Ooi 2016) showed that human subjects find a good view by not only scrutinising its own intra-view content, but also comparing it with other views of the same object. Note that a limitation of most previous work is the lack of the consideration of such inter-view knowledge in their algorithms. In this work, we propose a heuristic mechanism to formulate the inter-view knowledge via paired comparisons of views. Previous work (Wolfe 1994; Koch and Poggio 1999) in psychology pointed out that a basic principle of human visual system is to suppress the response to frequently occurring features, while at the same time it remains sensitive to features that deviate from the norm. We thus propose the VD layer as a heuristic method to formulate this principle where the view most different from all the other views are regarded as the most distinct one.

The VD layer takes as input the outputs of all average pooling layers. Since one 3D object is represented as N views, the input of the VD layer is a matrix of size $2048 \times N$ for a given object. Each of its columns can be regarded as a feature descriptor of one view. The VD layer outputs an N-dimensional vector to the WSP layer. Each element of the vector corresponds to the distinction of a particular view, reflecting how distinct that view is. The more distinct the view, the larger the contribution it will make in the aggregation of multi-view information implemented by the WSP layer.

Given two views V_i and V_j, their difference can be measured as the Euclidean distance between their feature descriptors F_i and F_j output by the average pooling layer of the ResNet50 backbone. However, this measure is insufficient as a view tends to have similar content with its neighbouring views. If a view is even very different from its neighbouring ones, it is likely to contain some unique content and thus can be considered confidently distinct from the others. Hence, the dissimilarity of two views should be proportional to the difference computed as the Euclidean distance between their feature descriptors and inversely proportional to the geodesic distance between their corresponding viewpoints on the view sphere. Such a heuristic also computationally holds for symmetric objects. For symmetric views, the dissimilarity is always 0 as $F_i = F_j$ and thus has nothing to do with the geodesic distance between them. Besides the N projected views, the UMVCNN also requires as input the view index $\text{VInd}_i \in \{1, 2, ..., 162\}$ generated as a byproduct when creating the multi-view representation of the object (see Section 3.1).

Let $\text{Geod}(\text{VInd}_i, \text{VInd}_j)$ be the geodesic distance between the viewpoints corresponding to V_i and V_j, the dissimilarity

between the two views is defined as:

$$D_{ij} = \frac{\left\| F_i - F_j \right\|}{1 + \alpha \cdot \text{Geod}(\text{VInd}_i, \text{VInd}_j))}, \quad \text{s.t.} \quad i, j \in \{1, 2 \ldots, N\} \text{ and } i \neq j \tag{1}$$

where $\alpha = 2$ in our implementation. The distinction of V_i is then computed as the sum of its pairwise dissimilarity to all the other views.

$$S_i = \sum_{j \neq i} D_{ij}. \tag{2}$$

Both Eqs. (1) and (2) are differentiable. So for back-propagation, given that the gradient passed to the VD layer is an N-dimensional vector $\mathcal{S}$, according to the chain rule, the gradient $\mathcal{F}$ of this layer with regard to its input can be computed as

$$\mathcal{F}_i = \mathcal{S}_i \frac{\partial S_i}{\partial F_i} \tag{3}$$

Considering Eqs. (1) and (2) and the partial derivative of the Euclidean distance function $\frac{\partial \|x\|}{\partial x_i} = \frac{x_i}{\|x\|}$, it can be computed as

$$\frac{\partial S_i}{\partial F_i} = \sum_{j \neq i} \frac{F_i - F_j}{(1 + \alpha \cdot \text{Geod}(\text{VInd}_i, \text{VInd}_j)) \cdot \left\| F_i - F_j \right\|}. \tag{4}$$

Weighted Sum Pooling (WSP) Layer To implement the view-object consistency principle through the loss layer which requires that the outputs of the view and the object channels have the same dimensions, we need to pool to aggregate the learned knowledge across all the 2D views to create a single descriptor for the 3D object. Also, we need to consider how to cast the influence of view distinction into this aggregation process where distinct views should have larger weights. Thus instead of the popular element-wise max pooling (Su et al. 2015; Kalogerakis et al. 2017) in multi-view CNNs, we carry out a WSP to incorporate view distinction as the weights into the pooling

$$P = \sum_{i=1}^{N} F_i S_i \tag{5}$$

where F_i is the column vector of the output of the average pooling layer F which denotes the feature descriptor of view V_i and S_i is its distinction output by the VD layer. It shows that the output of the WSP layer P regarded as the feature descriptor of the 3D object is estimated as the weighted sum of the feature descriptors of all the views where the weights are their distinctions. Eq. (5) can be expressed in a bilinear form as $P = FS$. Thus in the back-propagation, the gradients $\mathcal{F}$ and $\mathcal{S}$ of the WSP layer with regard to its inputs F and S respectively can be computed as

$$\mathcal{F} = \mathcal{P}S^T, \quad \mathcal{S} = F^T\mathcal{P} \tag{6}$$

where $\mathcal{P}$ denotes the gradient passed to the WSP layer.

MV-LogLoss Layer We propose the MV-LogLoss layer to formulate the proposition of view-object consistency, which enables an unsupervised learning. The basic idea herein is that no matter what the taxonomy is, the outcome of the classification based on the information of each 2D view should be consistent with that based on the entire 3D object. Note that as illustrated in Fig. 1, either of the view and the object channels alone is specifically designed to have the architecture of a classification network, which significantly facilitates the formulation of the view-object consistency. Moreover, such a design benefits salient view selection as the features vital for object classification are usually also important for the selection of a salient view. Psychological studies (Tarr and Pinker 1989; Cutzu and Edelman 1994; Hayward 1998) have validated the strong correlation between view selection and object recognition: a good view of an object can significantly help people to correctly recognise it.

The MV-LogLoss simply adapts the log loss in a multi-view scenario. This loss layer first computes the individual log loss of the softmax-normalised output of each Fc1 layer, $\mathcal{V}(i)$ with regard to that of the Fc2 layer, $\mathcal{O}$, which represent the final outputs of the view channel and the object channel respectively. The multi-view loss is then computed as the sum of all individual log losses:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{c=1}^{C} \mathcal{O}_c \cdot \log\left(\mathcal{V}_c(i)\right) \tag{7}$$

where for simplicity, we write the output of the view channel $\mathcal{V}_c(V_i)$ as $\mathcal{V}_c(i)$. Through training, Eq. (7) is minimised by impelling $\mathcal{O}$ to be consistent with $\mathcal{V}(i)$ and the view-object consistency is thus realised. It can be clearly seen that the MV-LogLoss defined as Eq. (7) does not rely on any annotations as $\mathcal{O}_c$ and $\mathcal{V}_c(i)$ are internally generated by the object channel and the view channel of the UMVCNN respectively. C in Eq. (7) is a picked integer defining the output dimension of the Fc1/Fc2 layer when building the UMVCNN. And we shall provide an experimental study on the influence of varying C in Section 5.4.

3.3 Implementation Details

The proposed method is fully unsupervised as it is trained using only a set of 3D objects without any annotations.

We first render each 3D object as 20 2D views as described in Section 3.1 using a standard OpenGL renderer with the

perspective projection mode. The strengths of the ambient light, the diffuse light and the specular reflection are set to 0.2, 0.6 and 0.1 respectively. We apply flat shading to the meshed object. Note that using different illumination models or shading coefficients does not affect our method due to the invariance of the learned convolutional filters to illumination changes, as observed in image-based CNNs. All rendered views are printed at 200 dpi, also in the OpenGL mode, and further resized to the resolution of 224×224. Then for training we feed these views into the UMVCNN wherein the convolutional layers taken from the ResNet50 backbone are initialised with the weights pretrained on ImageNet while the fully connected layers Fc1 and Fc2 are both initialised with random weights using the popular method proposed by He et al. (2015). The UMVCNN is trained end-to-end through stochastic gradient descent with the learning rate of 10^{-5}. As we observed, the training always converged within 50 epochs for all of the variants of the UMVCNN that we shall discuss in Sect. 5. When deploying the learned UMVCNN to select the salient view or detect the interest points of a given 3D object, we again render the object as 20 views with the same rendering settings and then use the schemes described in Sects. 4 and 6 below to output salient viewpoint and 3D interest points, respectively.

4 Salient View Selection with Global Distinction Adjustment

This section describes the method for salient view selection based on the UMVCNN as illustrated in Fig. 2. Given an object represented as a set of N views, we first feed the views into the learned UMVCNN and hijack the output of the Softmax layer connected with the Fc2 layer during the forward-propagation to predict its object class $\mathcal{C}$. Then, we back-propagate a C-dimensional one-hot vector where only the entry of index $\mathcal{C}$ is 1 from this Softmax layer to the input views with all the network weights fixed. This strategy leads to a per-pixel saliency map I_i for all the pixels in each view V_i based on their influence on the predicted class $\mathcal{C}$. The 2D saliency map I_i can be interpreted as a measure of pixel importance with regard to the recognition of the object. Like most methods for salient view selection (Lee et al. 2005; Secord et al. 2011; Leifman et al. 2016) and also to facilitate evaluations, we are keen to obtain the goodness of any viewpoint on a view sphere, which requires to generate a per-vertex saliency map.

To this end, we employed the 2D-to-3D saliency transfer scheme proposed in Song et al. (2020a) to derive a 3D saliency map H_i from a single 2D saliency map I_i. Then we aggregate multi-view saliency maps H_is into a single one through a linear model

$$H = \sum_{i=1}^{N} w_i H_i \tag{8}$$

where w_i denotes the contribution of a view-based 3D saliency map H_i. As a weighting parameter, it reflects the importance of a view in the aggregation. Secord et al. (2011) showed that such a linear model performed well when estimating the importance of views for various 3D objects. We propose to compute w_i as

$$w_i = S_i \Psi_i \tag{9}$$

where S_i is the output of the VD layer which represents the learned distinction of view V_i. However, according to Eqs. (1) and (2), S_i is based only on local pairwise distinction while the perceptual importance of a view should be subject to a global observation. So we further adjust the aggregation weight w_i by a factor of global distinction Ψ_i calculated via the scheme detailed in the following.

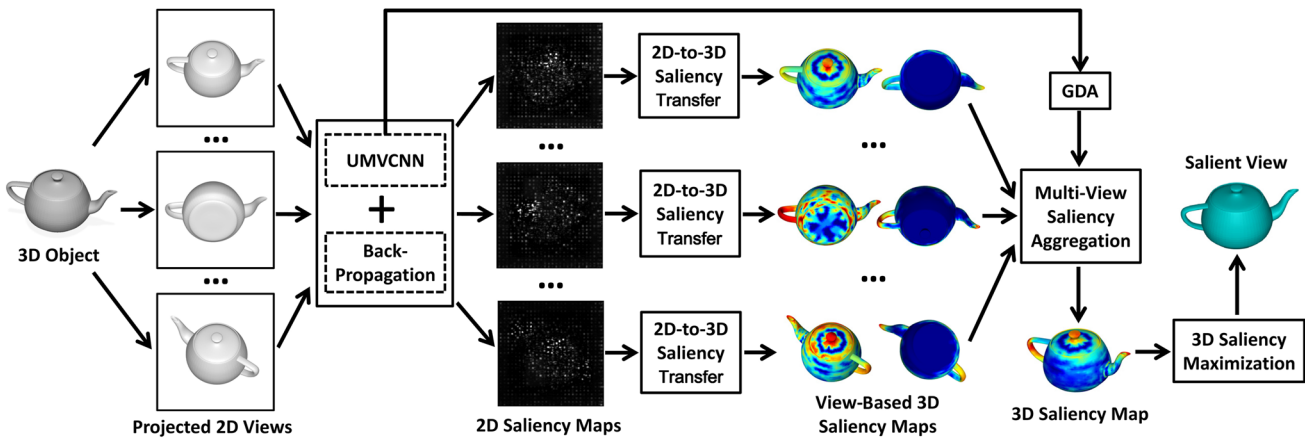


Fig. 2 The workflow of salient view selection based on the learned UMVCNN. Here we visualise two views for each view-based 3D saliency map, including the current view and the view roughly symmetric to it which mostly shows the vertices invisible to the current viewpoint

4.1 Global Distinction Computation

Directly aggregating all pairwise distinctions as in Eq. (2) cannot reliably lead to a global distinction as we do not know whether a local pairwise distinction is globally important or not. However, salient view selection of a 3D object is obviously influenced by global knowledge and thus methods integrating global cues have been proposed. For instance, Leifman et al. (2016) proposed an algorithm for detecting surface regions of interest and explored how to select viewpoints based on these salient regions. Their algorithm looks for regions that are distinct both locally and globally where the global consideration is if the object is 'limb-like' or not. We thus develop a specific scheme to calculate the global distinction Ψ_i for a view V_i.

First, we construct a matrix D where each entry D_{ij} is the pairwise distinction D_{ij} of the two views V_i and V_j computed via Eq. (1) using the deep features F_i and F_j extracted by the UMVCNN. D can be viewed as the weighted adjacency matrix of a graph where every pair of views are connected and the length connecting i and j is determined by the pairwise distinctiveness D_{ij}. As such, D encodes the information about how views are distinct from each other. We then define the global distinction as the centrality, a measure of the global influence of a node in a graph. By the Perron-Frobenius theorem (Perron 1907), D has a unique largest eigenvalue and its corresponding eigenvector ψ has strictly positive components. Accoding to Newman (2008), the i-th component of ψ gives the centrality score of the viewpoint i in the graph. Hence, we formulate the global distinction Ψ as the normalisation of ψ.

The above method can also be understood from a perspective of global distinction maximisation, formulated as below

$$\begin{aligned} \arg\max \mathcal{G} &= \sum_i G(V_i) \sum_j G(V_j) D_{ij}, \\ \text{s.t.} \quad & G \in \mathbb{R}^+ \text{ and } \|G\| = 1 \end{aligned} \tag{10}$$

where G can be understood as weights assigned to the views. A large G for the viewpoint i means that the view V_i is globally distinctive and thus its pairwise distinction D_{ij} has a large weighted impact in the overall distinction $\mathcal{G}$. Note that the ideal configuration is that the likelihood of the most distinctive view is 1 and that of any other view is 0. In practice, such situation is 'soften' and g is subject to $\|G\| = 1$ where we hope that the likelihood of the most salient view is a value close to 1. Eq. (10) can be written as

$$\arg\max \mathcal{G} = G^T D G, \quad \text{s.t.} \quad G \in \mathbb{R}^+ \text{ and } \|G\| = 1. \tag{11}$$

Since D is a symmetric real matrix, it is Hermitian. Thus Eq. (11) suggests that $\mathcal{G}$ is its Rayleigh quotient. The upper bound of the Rayleigh quotient is the largest eigenvalue of D and can be reached when G is equal to its corresponding eigenvector ψ. Therefore, as the solution to Eq. (10), the view distinction $G = \psi$ where each of its N elements is the distinction score of a particular view maximises the overall distinction $\mathcal{G}$ and essentially suggests whether a local pairwise distinction is globally important or not. Thus we calculate the global distinction Ψ_i for a view V_i as the normalised ψ_i.

4.2 Viewpoint Selection

We then select the viewpoint that maximises the sum of the saliency map H for the visible regions of the 3D object as the salient viewpoint:

$$v_s = \arg\max_v \Big(\sum_{m \in B(v)} H(m) \Big) \tag{12}$$

where $B(v)$ is the set of the vertices visible from the viewpoint v and $H(m)$ computed by Eq. (8) denotes the saliency of the vertex m. $M(v) = \sum_{m \in B(v)} H(m)$ can be regarded as the saliency map of the viewpoints. Figure 3 shows the 2D representation of the unwarped viewpoint saliency map on a view sphere normalised to the interval of [0, 1]. It is generated via the Mercator projection where the x and the y axes correspond to the latitude and the longitude, respectively. Initially, the model is not up oriented in the view sphere. A viewpoint saliency map suggests which viewpoints are salient and which are not for a specific 3D object or scene. Importantly, compared to a single salient view, a viewpoint saliency map assists us to more solidly judge whether a method for salient view selection works truly properly as it visualises the goodness of all sampled viewpoints.

5 Salient View Selection: Results

This section reports the results of salient view selection. We first introduce the datasets used in the experiments and evaluate our method qualitatively. Then, we evaluate both the

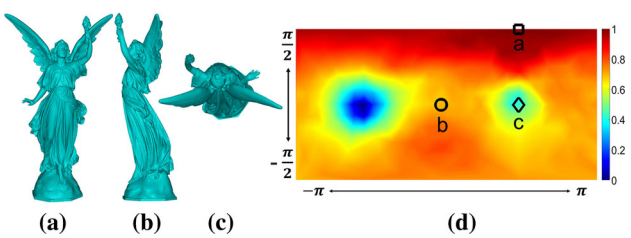


Fig. 3 Viewpoint saliency map. **a**–**c** are the projected views of the Lucy model. **d** is the viewpoint saliency map where the black square, circle and diamond mark the locations of the viewpoints corresponding to the views shown in (**a**)–(**c**) respectively (Color figure online)

proposed UMVCNN and its variants via quantitative comparisons for the demonstration of its superiority as well as a better understanding of our method. In addition, we evaluate the robustness of the UMVCNN against noise. Finally, we demonstrate via a user study that our method can be directly used to select good views for various 3D scenes to attract further interest.

5.1 Datasets

We create a new dataset containing 2747 3D objects downloaded from the Princeton ModelNet dataset (Wu et al. 2015), the Schelling dataset (Chen et al. 2012) and the Trimble 3D Warehouse (Warehouse 2020). These models are originally from 30 object categories while in this work, all categorical annotations are removed in training and validation for an unsupervised learning. We use the same data split of ModelNet40 as in Wu et al. (2015) where four fifths of the objects in each category are used for training and one fifth are used for validation.

We test our method on the Best View Selection (BVS) benchmark (Dutagaci et al. 2010). To the best of our knowledge, it is the only one publicly available benchmark suitable for quantitatively evaluating view selection methods. The BVS benchmark contains 68 3D objects of various classes including some that do not belong to any of the 30 object categories mentioned above from the perspective of human recognition. It also provides a quantitative benchmarking measure, the ground truth best viewpoints picked by 26 people and the results of 7 baseline methods.

We also used objects from the Stanford 3D Scanning Repository (Curless and Levoy 1996), the Princeton Shape Benchmark (Shilane et al. 2004) and the Watertight Models Track of SHREC'07 (Giorgi et al. 2007) for qualitative evaluations.

5.2 Qualitative Results

Figure 4 shows our results of salient view selection for a variety of 3D objects, with the ground truth best viewpoints supplied by the BVS benchmark. It is noteworthy that the ground truth best viewpoints could be more or less than 26 because 1) several human participants could select the same viewpoint and 2) the symmetry of each object is taken into account and thus the symmetric viewpoints of those picked by the participants are also included. It can be seen that the consistency of human preferred viewpoints varies over different objects. Even though, for most objects, the majority of the ground truth best viewpoints fall into the red or orange areas in the viewpoint saliency maps generated by our method, which demonstrates that it is good at predicting human's viewpoint preference over various objects. Also, for most objects, the salient viewpoint found by our method is, or at least very close to, a ground truth viewpoint picked by a human subject. It is worth mentioning that due to the default distortion of the Mercator projection, for the Ant model, the viewpoints on the bottom boundary of the viewpoint saliency map that look distant from each other are actually very close to each other on the view sphere since they are all very close to its bottom pole.

To visualise the effectiveness of the proposed GDA scheme, we show the qualitative results with and without it in Fig. 5 for comparison. It can be seen that when integrated with the UMVCNN, the GDA improves the salient view selection for various objects, producing good views that most of us would prefer. It can also be observed that for most objects, the UMVCNN without the GDA generates roughly correct viewpoint saliency maps as well. And in such cases, applying the GDA does not significantly change the viewpoint saliency maps. The mug is an exception as the method without the GDA generates an incorrect viewpoint saliency map inconsistent with the symmetry of the object. By contrast, the introduction of the GDA leads to the viewpoint saliency map consistent with the symmetry of the mug and generates a good view for it. The quantitative results with and without the GDA scheme which enable a more solid comparison can be found in Section 5.3.

We next compare our results to some produced by competing state-of-the-art methods. Since some of them require tuning of parameters and some are not open-sourced, we used our method to select salient views for the same objects used in the papers where the methods were reported. Figure 6 compared our method with Lee et al. (2005) and Yamauchi et al. (2006). It can be seen that our method is less influenced by some local geometric features such as the sharp edges at the bottom of the hand object if semantically they do not help the recognition of the object. Similarly, as shown in Fig. 7, the method proposed in Leifman et al. (2016) chose a back view of the lamp which contains many local details such as wires and screws. In comparison, for both the lamp and the jeep, our method tends to select views natural and good for recognising the objects. Figure 7 also shows that our method outperforms Song et al. (2020a) over a helicopter and a horse object while more convincing quantitative comparisons using a variety of 3D objects are provided in Section 5.3 below. Note that Song et al. (2020a) is essentially based on a weakly supervised deep learning framework where the class labels of the objects are available during training.

Please refer to the supplemental material for more qualitative results of salient view selection of 3D objects.

5.3 Quantitative Results

We tested our method on the BVS benchmark (Dutagaci et al. 2010) which contains 68 objects using a computer with an Intel i7-4790 3.6GHz CPU and 32GB RAM without any

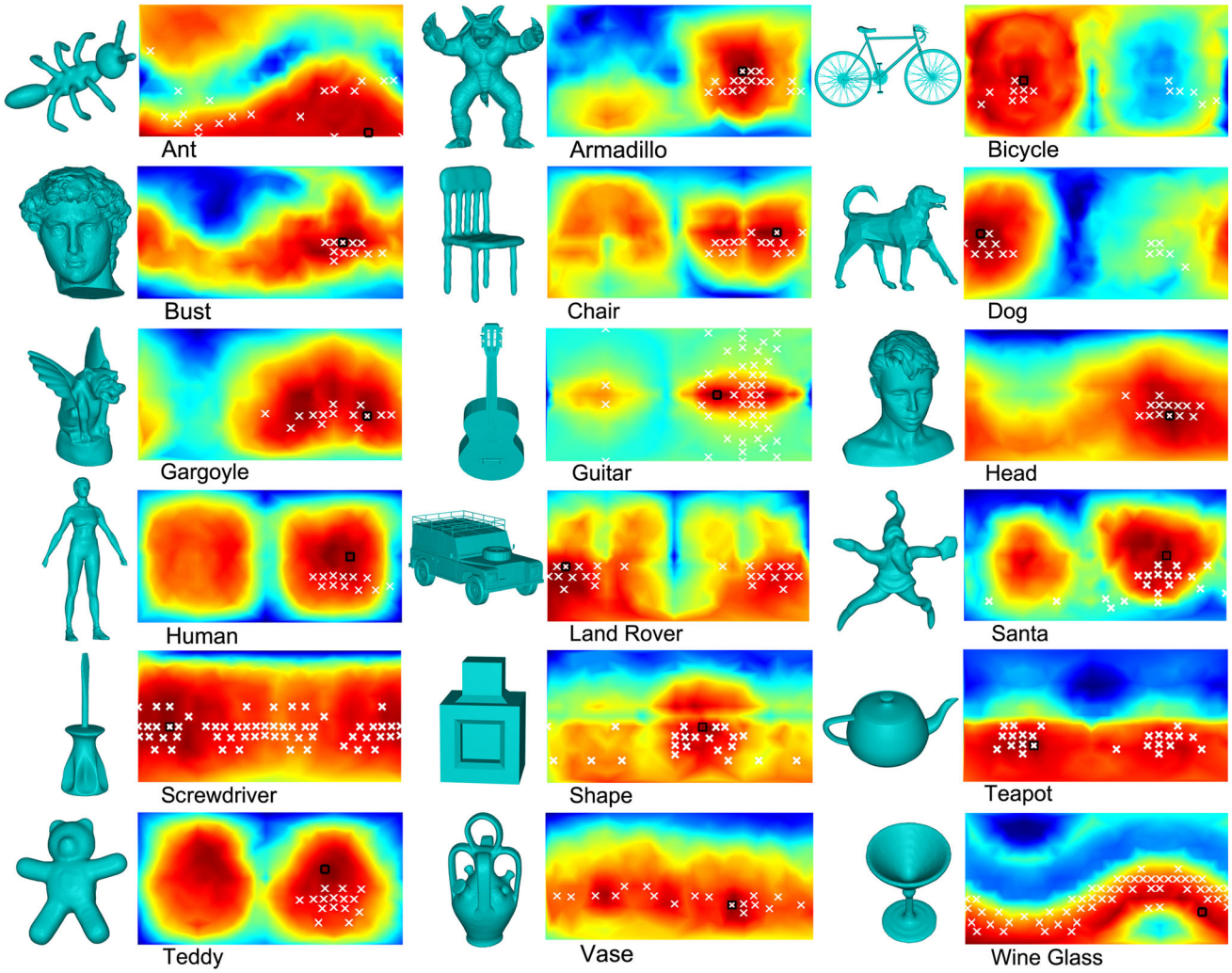


Fig. 4 Qualitative results of the salient views and the estimated viewpoint saliency maps generated by our method. In each map, the black square corresponds to the salient viewpoints selected by our method. The white "X"s correspond to the ground truth best viewpoints picked by 26 human subjects (including their symmetric viewpoints) via the user study carried out by Dutagaci et al. (2010)

GPU acceleration. The salient views of most objects can be computed within 1 minute where the vertex visibility to each viewpoint is precomputed.

Table 1 gives the statistics of the View Selection Error (VSE) of 9 automatic view selection methods over all of the 68 objects. The VSE proposed by Dutagaci et al. (2010) measures the geodesic distance between the viewpoint found by a method and the ground truth supplied by a human subject on a unit view sphere and is averaged over the choices of all subjects, with the consideration of object-specific symmetry.

According to Table 1, our method yields the best performance in terms of the mean VSE, the median VSE and the number of objects for which a method gave the lowest VSE among all the competing methods. Here we set $C = 30$ for all versions of UMVCNNs. As mentioned at the end of Section 3.2, this means that the output dimension of the Fc1 and Fc2 layers is set to 30 when we build the UMVCNN, which indicates that either of the view and the object channels categorises the objects into 30 classes. As shown in Fig. 4, due to the inconsistency of the ground truth choices of human subjects over the same object, reaching a zero mean VSE is impossible and improving the VSE is very challenging if it is already low. In most cases, a viewpoint with a mean VSE lower than 0.3 corresponds to a good view. Even though, our method (UMVCNN-ResNet-GDA) outperforms the state-of-the-art method proposed in Song et al. (2020a) by 5.0%, 4.3%, 12.1% and 32.8% in terms of the mean, the median, the standard deviation and the interquartile range of the VSE respectively. Note that their method is also based on deep learning but trained, in a weakly supervised manner, on a large dataset with the annotations of object class membership.

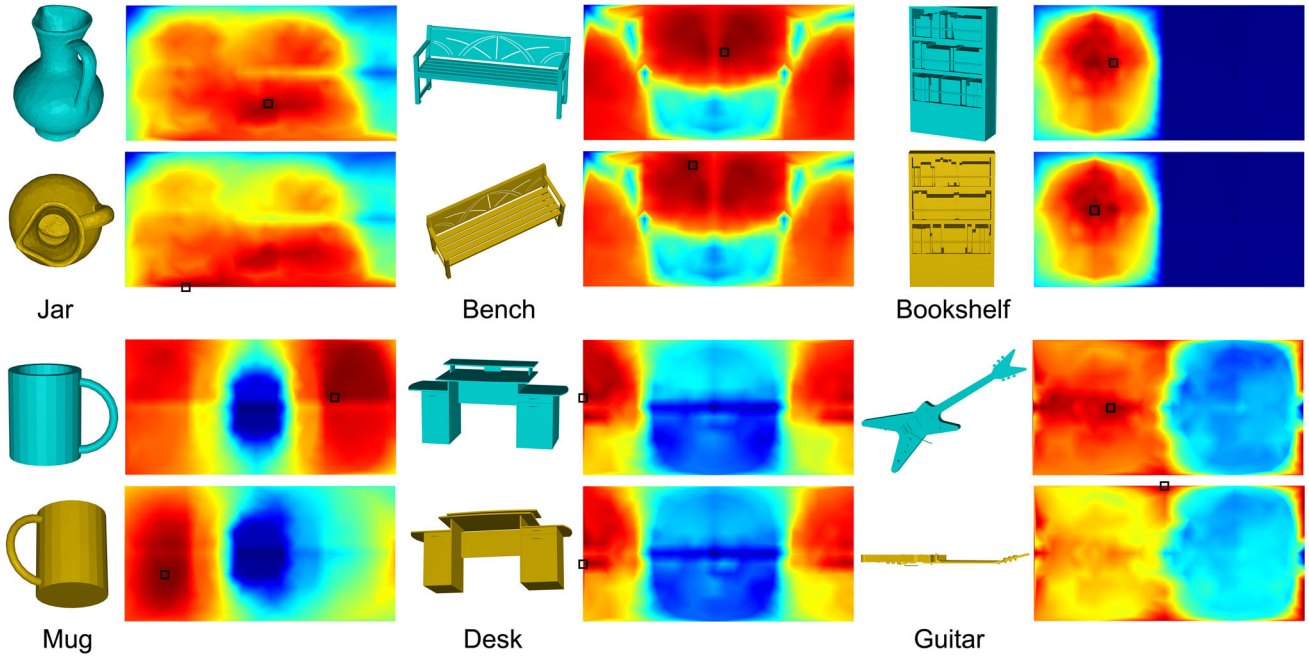


Fig. 5 Qualitative results of the methods with and without the GDA scheme. The first and the third rows show the detected salient views and the estimated viewpoint saliency maps using the method with the GDA scheme. For comparison, the second and the fourth rows show the corresponding results using the method without the GDA scheme. The black squares in the viewpoint saliency maps mark the most salient viewpoints

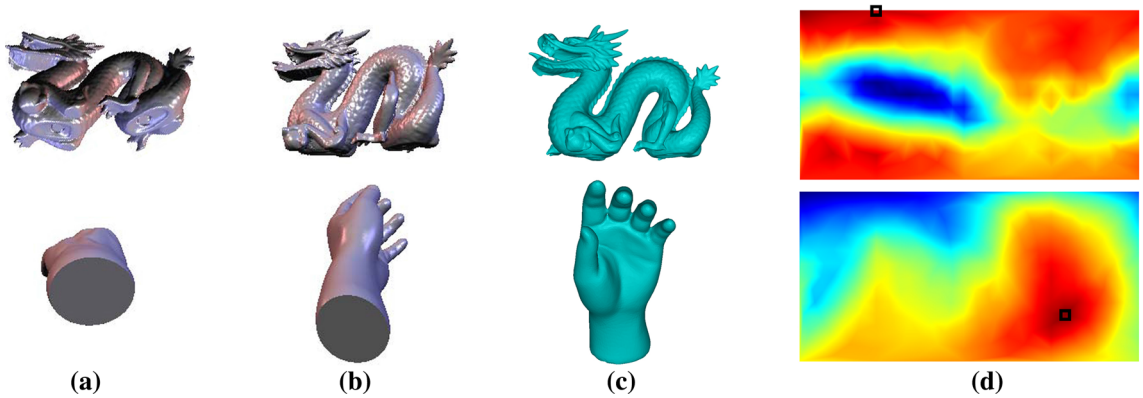


Fig. 6 Qualitative comparisons with Lee et al. (2005) and Yamauchi et al. (2006). **a** The best views selected by Lee et al. (2005) (as implemented and shown in Yamauchi et al. (2006)). **b** The best views selected by Yamauchi et al. (2006). **c** The best views selected by our method. **d** The viewpoint saliency maps generated by our method where the black squares mark the most salient viewpoints

None of the methods is consistently the best over all 68 objects although our method accomplishes the best results for 17 objects, the most over all competing methods. This is in agreement with the conclusions in Biederman (1987) and Secord et al. (2011) which argued that human's view preference is driven by a variety of attributes. But in general, the methods based on low-level attributes perform significantly worse than those based on deep neural networks which potentially learn some high-level attributes of 3D objects. It is also worth mentioning that the number of objects for which the UMVCNN-VGG gave the lowest VSE is 20 as reported in Song et al. (2020b) while it is 15 according to Table 1. This is because the updated method with GDA (i.e. UMVCNN-ResNet-GDA) outperforms UMVCNN-VGG on 5 out of the 20 objects while they achieve the same performance on most of them.

In particular, Table 1 shows that our method significantly outperforms Dutagaci et al. (2010) based on view area and Polonsky et al. (2005) based on silhouette length in terms of the VSE. This demonstrates that the improvement of the VSE

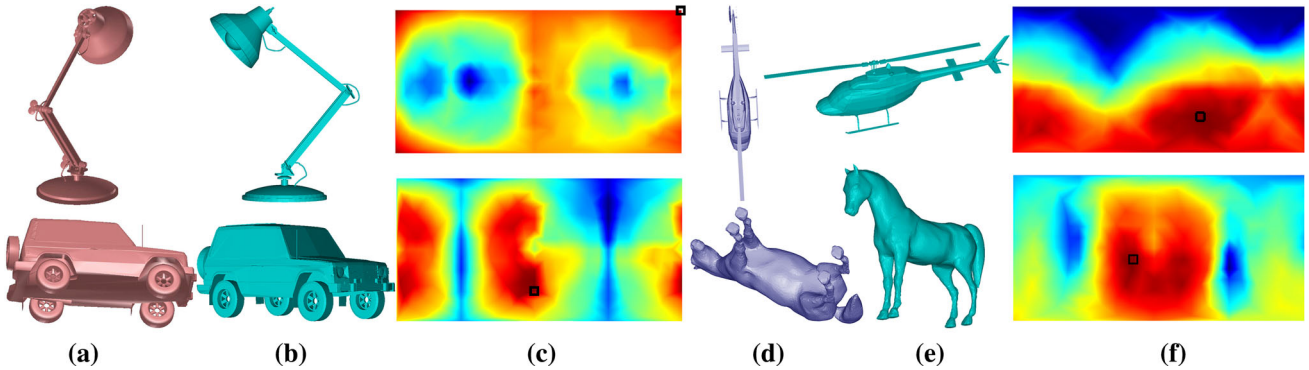


Fig. 7 Qualitative comparisons with Leifman et al. (2016) and Song et al. (2020a). **a** and **d** The best views selected by Leifman et al. (2016) and Song et al. (2020a) respectively. **b** and **e** The best views selected by our method. **c** and **f** The viewpoint saliency maps generated by our method where the black squares mark the most salient viewpoints (Color figure online)

Table 1 Statistics of the View Selection Error (VSE) of different view selection methods over the 68 objects from the BVS benchmark. SD and IQR represent the standard deviation and the interquartile range respectively. *n* gives the number of objects for which a method gave the lowest VSE (including the joint lowest) among all the competing methods. UMVCNN-VGG and UMVCNN-ResNet denote the UMVCNNs using VGG19 and ResNet50 as backbone, respectively. UMVCNN-VGG-GDA and UMVCNN-ResNet-GDA indicate that the UMVCNNs are integrated with the GDA scheme

View selection method	Mean VSE	Median VSE	SD of VSE	IQR of VSE	n
View area (Dutagaci et al. 2010)	0.517	0.539	0.186	0.306	6
Ratio of visible area (Polonsky et al. 2005)	0.473	0.473	0.196	0.338	1
Surface area entropy (Vázquez et al. 2001)	0.396	0.386	**0.144**	**0.195**	8
Silhouette length (Polonsky et al. 2005)	0.446	0.445	0.172	0.275	6
Silhouette entropy (Page et al. 2003)	0.484	0.469	0.153	0.241	4
Curvature entropy (Page et al. 2003)	0.474	0.466	0.139	0.239	7
Mesh saliency (Lee et al. 2005)	0.430	0.395	0.165	0.233	2
Deep mesh distinction (Song et al. 2020a)	0.380	0.346	0.173	0.314	11
UMVCNN-VGG (Song et al. 2020b)	0.367	0.336	0.165	0.236	15
UMVCNN-ResNet (Ours)	0.365	0.334	0.155	0.229	15
UMVCNN-VGG-GDA (Ours)	0.364	0.334	0.153	0.227	15
UMVCNN-ResNet-GDA (Ours)	**0.361**	**0.331**	0.152	0.211	17

Bold values denote the top performing methods in terms of the corresponding metrics

does come from the UMVCNN rather than the handcrafted features, i.e. view area and silhouette length that we use for the multi-view representation of a 3D object (see Section 3.1).

5.4 Evaluations over the Variants of UMVCNN

Effect of varying C Table 2 gives the mean VSE of the variants of the UMVCNN. We first redesign and test the UMVCNN with different values of the variable C introduced in Eq. (7). It can be seen that varying C from the default value 30 leads to an insignificant degradation of performance. As mentioned in Section 5.1, the 3D objects used for training are originally from 30 object categories while we removed all categorical annotations in this work for an unsupervised learning. Presumably, that $C = 30$ is indeed a good choice for designing the UMVCNN can be interpreted by the fact that salient view selection is a task highly related to 3D object classification as we observe that the objects of the same class tend to have analogous salient viewpoints while it is not the case the other way round. However, we cannot observe any obvious rule that suggests a way for deciding C. In a supervised learning, the network is forced to adopt the taxonomy of object classification consistent with human annotations while there is no guarantee that this taxonomy is optimal to the particular task such as salient view selection. Thus in different tasks, C might need to be tuned, but not necessarily fine-tuned as the UMVCNN is not very sensitive to it.

Ablation study for validating VD and WSP We are also interested in the heuristic component of the UMVCNN, i.e. the VD and WSP layers. To validate its effective-

Table 2 Mean view selection error (VSE) of the variants of the UMVCNN over 68 objects

UMVCNN variants	$C = 10$	$C = 15$	$C = 20$	$C = 25$	$C = 30$	$C = 30$, Max-pooling	$C = 30$, 30 views	$C = 35$	$C = 40$
UMVCNN-VGG (Song et al. 2020b)	0.379	0.373	0.382	0.381	0.367	0.384	0.366	0.377	0.380
UMVCNN-ResNet (Ours)	0.388	0.384	0.375	0.368	0.365	0.378	0.365	0.376	0.381
UMVCNN-VGG-GDA (Ours)	**0.375**	**0.371**	**0.370**	0.380	0.364	0.382	0.363	**0.372**	**0.374**
UMVCNN-ResNet-GDA (Ours)	0.379	0.372	**0.370**	**0.365**	**0.361**	**0.372**	**0.361**	**0.372**	0.375

Bold values denote the top performing methods in terms of the corresponding metrics

ness, we replace the VD and WSP layers with the popular element-wise max pooling which has demonstrated the state-of-the-art performance in various 3D shape understanding tasks such as classification (Su et al. 2015), retrieval (Su et al. 2015) and segmentation (Kim et al. 2017). Such variants correspond to the column of '$C = 30$, max pooling' in Table 2. To aggregate the multi-view 3D saliency maps H_i in Eq. (8), we set all weighting parameters w_i to 1 as it is not available via this variant. As shown in Table 2, the performance of the UMVCNN is significantly worse without the VD and the WSP layers. This demonstrates the effectiveness of the view distinction heuristic we introduced in Sect. 3.2. It also suggests that the unsupervised learning based on the view-object consistency principle is likely to benefit from some heuristics introduced for the specific task. It is worth mentioning that an ablation experiment with only the WSP layer is not available. This is because without the VD layer, the weights required for implementing the weighted sum pooling cannot be computed in an unsupervised setup.

Effect of the Number of Views We tested the variants corresponding to the column of '$C = 30$, 30 views' in Table 2 where a 3D object is projected into 30 (i.e. $N = 30$) instead of 20 views using the method described in Sect. 3.1. All the other variants in Table 2 used a 20-view setup. It can be seen that using 30 views usually cannot further boost the performance. Using more or different views is trivial, however, we found that a 20-view setup is already enough to achieve high performance but with an advantage of computational efficiency.

5.5 Evaluation of the Robustness against Noise

We have conducted experiments by adding Gaussian noise with $\sigma = 0.001B$, $0.002B$ and $0.004B$ respectively to all of the 68 3D objects in the BVS benchmark where B is the length of the diagonal of the bounding box of a particular object. Table 3 lists the results of salient view selection on the noisy objects using the UMVCNN-ResNet-GDA model. It demonstrates that our method is relatively robust against noise.

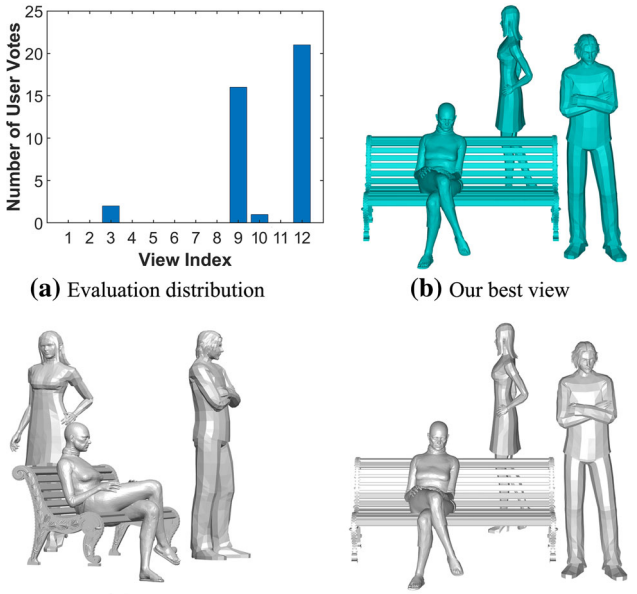
Table 3 Evaluation of the robustness of our method against noise in terms of mean VSE, median VSE and standard deviation (SD) of VSE. B denotes the length of the diagonal of the bounding box of the mesh

Noise amount	Mean VSE	Median VSE	SD of VSE
No noise	0.361	0.331	0.152
$0.001B$	0.373	0.338	0.154
$0.002B$	0.381	0.348	0.155
$0.004B$	0.387	0.356	0.159

5.6 Salient View Selection of 3D Scenes

High-quality view of 3D scenes could navigate observers to the region of interest, help them to seek the hidden relations of hierarchical structure, and improve the efficiency of virtual exploration. The selection of best views of 3D scenes thus has a range of applications including virtual reality (Freitag et al. 2018), scene synthesis (Zhang and Fei 2019), robotic manipulation (Guérin et al. 2018) and autonomous navigation (Zhu et al. 2020).

The proposed UMVCNN can be directly used for the salient view selection of 3D scenes for two reasons. First, it is a multi-view CNN where the global spatial relationship of multiple objects not connected by mesh edges is recorded in one or multiple 2D views of the scene. And such relationship is vital for selecting salient views of 3D scenes. In comparison, another popular model for 3D deep learning, graph neural network (GNN), might not be good at capturing such relationship. This is because a GNN usually learns a deep representation over the mesh treated as a non-Euclidean graph by a local operator such as Laplacian (Bruna et al. 2013; Defferrard et al. 2016) and Dirac operators (Kostrikov et al. 2018) which do not encode global spatial information among multiple objects. Second, previous deep learning-based methods for salient view selection of 3D objects (Kim et al. 2017; Song et al. 2018) relied on category labels of samples for training. Thus they cannot be directly applied to a 3D scene which usually do not associate with a unique category label as it typically contains objects of different categories. By contrast, our UMVCNN does not rely on the knowledge about object categories, which means that we can simply treat a 3D



(a) Evaluation distribution **(b)** Our best view

(c) The ground truth (View #9 and View #12)

Fig. 8 Salient views for a 3D scene. **a** Two views are picked by the evaluators as the best views for representing the 3D scene. Our computed salient view (**b**) is very close to one (View #12) of the two selected as the ground truth. **c** The two views selected by the evaluators

scene as an object and directly feed it into the UMVCNN to predict its salient view.

Since there is no ground truth currently available for evaluating the salient view selection of 3D scenes, we conducted a user study involving 25 human evaluators. The goal was to learn which views of a 3D scene are considered the most salient. For each of our 50 scenes, we produced 12 images, each taken from a different viewpoint. We decided to use 12 images as a compromise between the accuracy of the survey (which requires a large number of viewpoints) and our wish to avoid overloading the evaluators (which requires a small number of viewpoints). We asked the evaluators to mark the views that can best represent the scene. The number of the representative views for each scene that could be marked was unlimited.

Figure 8 shows a typical distribution of the evaluation. In this example, there are two views considered salient views by the evaluators. Therefore, rather than defining our ground truth to be a single view, we define it as a set of views, consisting of the highest-ranked views before the largest decrease in the histogram. To assess the results of our method, we compared the view selected by our method to the ground truth. Our result is considered correct if it is geodesically closer to a view of the ground truth than to any other view.

According to the results shown in Fig. 9, our method successfully selects good views for various 3D scenes. For 43 out of 50 scenes (86 percent), the most salient view selected by our method matched the ground truth. The viewpoint saliency maps of 3D scenes generated by our method are also informative. For instance, by observing the corresponding locations of the best and the worst views in the viewpoint saliency maps of most scenes, we find that the views with positive elevation angles are generally much more salient than those with negative ones, which is consistent with human's viewpoint preference. We also observed that the best view of a scene is not necessarily the best view of each individual object in it. For example, in the living room scene, the best view of the entire scene is not that of one of the three sofas. Similarly, in the work site scene, the best view of the scene is not that of the person in the middle and some chairs.

Please refer to the supplemental material for more qualitative results of salient view selection of 3D scenes.

6 3D Interest Point Detection

Detection of interest points on the surface of a 3D object is challenging since usually not just local attributes but some global attributes hard to compute are considered when people select them (Dutagaci et al. 2012). To extract a set of discrete interest points from a continuous saliency distribution of a 3D object, we first compute its per-vertex saliency map (i.e. H in Eq. (8)) based on the learned UMVCNN integrated with the GDA scheme. Then we remove the vertices with saliency values smaller than a global threshold (set to 70% of the maximum saliency). Finally, to collect a set of interest points, we extract any vertex that either has a saliency value larger than another global threshold (set to 90% of the maximum saliency) or corresponds to the local maximum of saliency.

6.1 Dataset and Evaluation Metrics

For a fair comparison, we test our method on the publicly available 3D Interest Point Detection (3DIPD) benchmark (Dutagaci et al. 2012) which provides 3D interest points selected by 23 human subjects as ground truth. Previous evaluation methods usually measured the repeatability rate according to varying factors, such as model deformation, scale change, different modalities, noise, and topological change. Unlike them, the 3DIPD benchmark measured how similar the detected points of interest are to those selected by human subjects. It thus proposed three metrics: false negative error (FNE), false positive error (FPE) and weighted miss error (WME). FNE and FPE are defined in the obvious way. Normally, as more 3D interest points are captured, more false positives are detected while achieving a lower FNE. If a method tends to mark fewer interest points, it results in a lower FPE, at the cost of a higher FNE. A method leads to a low WME if it manages to detect a point that is frequently voted by human subjects. Thus it measures the ability of a method to detect the most interesting points. In contrast,

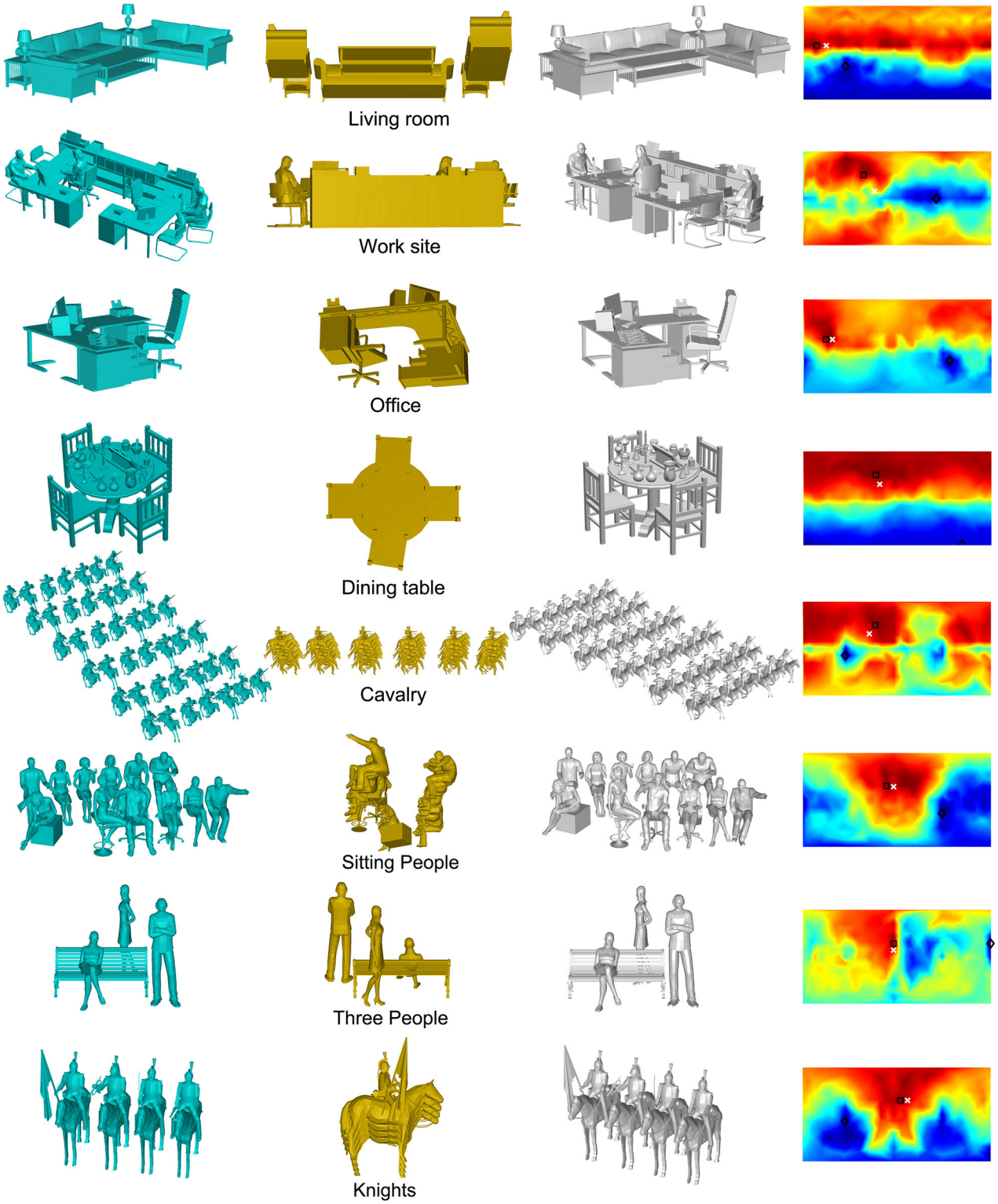


Fig. 9 Salient and non-salient views of 3D scenes (courtesy of the Trimble 3D Warehouse (Warehouse 2020)) selected by our method. From left to right: The first column shows the most salient views of the scenes selected by our method; The second column shows the least salient views of the scenes selected by our method; The third column shows the ground truth best views supplied by the user study where we only show the one closest to the salient view generated by our method if there are more than one ground truth views; The fourth column shows the viewpoint saliency maps of the scenes where the black squares mark the most salient views, the black diamonds mark the least salient views and the white "X"s mark the ground truth best views, respectively

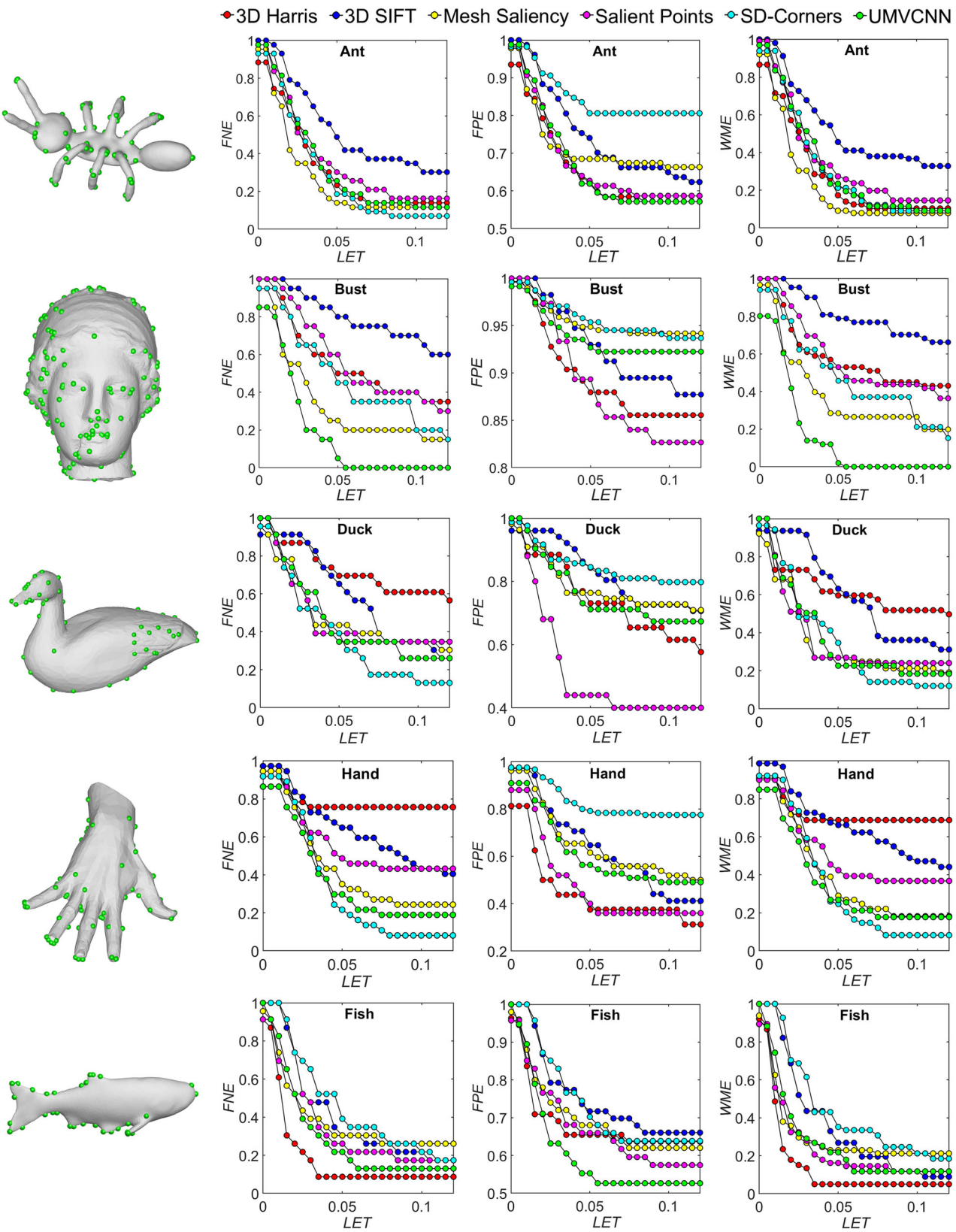


Fig. 10 Results of 3D interest point detection for various objects on the 3DIPD benchmark (Dutagaci et al. 2012). From left to right: the first column visualises the interest points detected by our method; the second to the fourth columns show the FNE, FPE and WME graphs, respectively

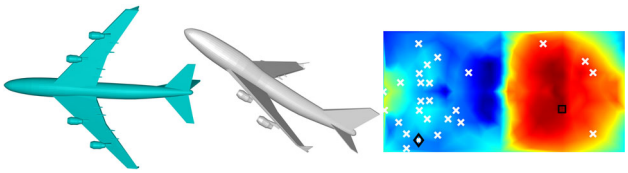


Fig. 11 Limitation. Our method sometimes tends to select views good for recognition but not necessarily "natural". Left: the view selected by a subject; Middle: the view selected by our method; Right: the viewpoint saliency map where the diamond and the square mark the views selected by the subject and our method respectively

FNE and FPE treat all ground truth points of interest equally. Hence, an ideal method should keep FNE, FPE and WME all low.

6.2 Comparative Results

We compare our method based on the UMVCNN-ResNet-GDA model for 3D interest point detection with the five baseline methods named as 3D-Harris, 3D-SIFT, Mesh Saliency, Salient Points and SD-Corners by the 3DIPD benchmark) for which the 3DIPD benchmark provides the results to facilitate comparisons. Figure 10 shows the results through FNE, FPE and WME graphs with respect to the localisation error tolerance (LET). A vertex is considered to be 'correctly detected' as a point of interest if its geodesic distance to the closest ground truth point of interest is not larger than a specific LET value. We can see that the 3D interest points detected by our UMVCNN-based method correspond to low FNE and WME, which means that quite a few detected points are of human perceptual interest. In particular, for the fish object, all of the three errors are low. Overall, our method has a good performance compared with the competing methods specifically designed for 3D interest point detection.

7 Conclusions

This work reveals that the view-object consistency principle is promising for the establishment of an unsupervised framework of 3D deep learning. We validate its effectiveness on the challenging tasks of salient view selection and 3D interest point detection through the relatively naive design of a multi-view deep architecture. While the performance of our method is impressive, it has some limitations as shown in Fig. 11 when applied to salient view selection. Our method sometimes tends to select a view good for recognising the object, such as the view that better shows some features important for recognising the airplane (e.g. the wings and the engines). However, most human subjects prefer a "natural" side view.

Future work will focus on implementing the unsupervised learning framework in more applications to demonstrate that it is amenable to a wide range of 3D shape understanding tasks. Particularly interesting applications might be some 3D scene understanding tasks hindered by the difficulty of collecting large amounts of accurately and consistently annotated data for training.

References

3D Warehouse (2020) https://3dwarehouse.sketchup.com

Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., & Tai, CL. (2020) D3feat: Joint learning of dense detection and description of 3d local features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , pp 6359–6367

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115.

Blanz, V., Tarr, M. J., & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception, 28*(5), 575–599.

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013) Spectral networks and locally connected networks on graphs. In: Proceeding of ICLR

Castellani, U., Cristani, M., Fantoni, S., & Murino, V. (2008) Sparse points matching by combining 3d mesh saliency with statistical descriptors. In: Proceeding of eurographics, pp 643–652

Chen, X., Saparov, A., Pang, B., & Funkhouser, T. (2012). Schelling points on 3d surface meshes. *ACM Transactions on Graphics (Proc SIG- 974 GRAPH), 31*(4), 29.

Curless, B., & Levoy, M. (1996) A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pp 303–312

Cutzu, F., & Edelman, S. (1994). Canonical views in object representation and recognition. *Vision Research, 34*(22), 3037–3056.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: NeurIPS, pp 3844–3852

Deng, H., Birdal, T., & Ilic, S. (2018) PPFnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 195–205

Dutagaci, H., Cheung, CP., & Godil, A. (2010) A benchmark for best view selection of 3d objects. In: Proceedings of ACM workshop on 3DOR, pp 45–50

Dutagaci, H., Cheung, C., & Godil, A. (2012). Evaluation of 3d interest point detection techniques via human-generated ground truth. *The Visual Computer, 28*, 901–917.

Freitag, S., Weyers, B., & Kuhlen, TW. (2018) Interactive exploration assistance for immersive virtual environments based on object visibility and viewpoint quality. In: 2018 IEEE Conference on virtual reality and 3D user interfaces (VR), pp 355–362

Giorgi, D., Biasotti, S., & Paraboschi, L. (2007). Shape retrieval contest 2007: Watertight models track. *SHREC Competition, 8*(7), 7.

Guérin, J., Gibaru, O., Nyiri, E., Thieryl, S., & Boots, B. (2018) Semantically meaningful view selection. In: 018 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 1061–1066

Han, H., Li, J., Wang, W., Zhao, H., & Hua, M. (2014) View selection of 3d objects based on saliency segmentation. In: IEEE conference on virtual reality and visualization, pp 214–219

Hayward, W. G. (1998). Effects of outline shape in object recognition. *Journal of experimental psychology: human perception and Performance, 24*(2), 427.

He, J., Wang, L., Zhou, W., Zhang, H., Cui, X.,& Guo, Y .(2018) Viewpoint assessment and recommendation for photographing

architectures. In: IEEE transactions on visualization and computer graphics

He, K., Zhang, X., Ren, S., & Sun, J. (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Huang, H., Kalogerakis, E., Chaudhuri, S., Ceylan, D., Kim, V. G., & Yumer, E. (2018). Learning local shape descriptors from part correspondences with multiview convolutional networks. *ACM Transactions on Graphics (TOG), 37*(1), 6.

Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S. (2017) 3D shape segmentation with projective convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 1, p 8

Kim, Sh., Tai, Y. W., Lee, J. Y., Park, J., & Kweon, I. S. (2017). Category-specific salient view selection via deep convolutional neural networks. *Computer Graphics Forum, 36*(8), 313–328.

Koch, C., & Poggio, T. (1999). Predicting the visual world: silence is golden. *Nature Neuroscience, 2*, 9–10.

Kostrikov, I., Bruna, J., Panozzo, D., & Zorin, D. (2018) Surface networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. *ACM Transition Graph (Proc SIGGRAPH), 24*(3), 659–666.

Leifman, G., Shtrom, E., & Tal, A. (2016). Surface regions of interest for viewpoint selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(12), 2544–2556.

Li, J., & Lee, GH. (2019) Usip: Unsupervised stable interest point detection from 3d point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision , pp 361–370

Lienhard, S., Specht, M., Neubert, B., Pauly, M., & Müller, P. (2014) Thumbnail galleries for procedural models. In: Computer Graphics Forum, Wiley Online Library, vol 33, pp 361–370

Mezuman, E., & Weiss, Y. (2012) Learning about canonical views from internet image collections. In: Proceedings of NeurIPS, pp 719–727

Mian, A., Bennamoun, M., & Owens, R. (2010). On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision, 89*(2), 348–361.

Newman, ME. (2008) The mathematics of networks. The new palgrave encyclopedia of economics

Novotny, D., Larlus, D., Vedaldi, A. (2017) Learning 3d object categories by looking around them. In: Proceedings of the IEEE international conference on computer vision

Page, DL., Koschan, AF., Sukumar, SR., Roui-Abidi, B., Abidi, MA. (2003) Shape analysis algorithm based on information theory. In: Proceedings 2003 international conference on image processing (Cat. No. 03CH37429), vol 1, pp I–229

Perron, O. (1907). Zur theorie der matrices. *Mathematische Zeitschrift, 64*(2), 248–263.

Polonsky, O., Patané, G., Biasotti, S., Gotsman, C., & Spagnuolo, M. (2005). What's in an image? *The Visual Computer, 21*(8–10), 840–847.

Qi, CR., Su, H., Nießner, M., Dai, A., Yan, M., & Guibas, L. (2016) Volumetric and multi-view cnns for object classification on 3d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5648–5656

Secord, A., Lu, J., Finkelstein, A., Singh, M., & Nealen, A. (2011). Perceptual models of viewpoint preference. *ACM Transactions on Graphics (TOG), 30*(5), 109.

Shilane, P., & Funkhouser, T. (2006) Selecting distinctive 3d shape descriptors for similarity retrieval. In: IEEE International conference on shape modeling and applications 2006 (SMI'06)

Shilane, P., Min, P., Kazhdan, M., Funkhouser, T. (2004) The princeton shape benchmark. In: Proceedings of shape modeling applications

Song, R., Liu, Y., Martin, R., & Rosin, P. (2013). 3d point of interest detection via spectral irregularity diffusion. *The Visual Computer, 29*(6–8), 695–705.

Song, R., Liu, Y., Martin, R., & Echavarria, K. R. (2018). Local-to-global mesh saliency. *The Visual Computer, 34*(3), 323–336.

Song, R., Liu, Y., & Rosin, P. L. (2020). Distinction of 3D objects and scenes via classification network and markov random field. *IEEE Transactions on Visualization and Computer Graphics, 26*(6), 2204–2218.

Song, R., Zhang, W., Zhao, Y., & Liu, Y. (2020). Unsupervised multi-view cnn for salient view selection of 3d objects and scenes. *European Conference on Computer Vision* (pp. 454–470). ECCV: Springer.

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, EG. (2015) Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision, pp 945–953

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*(2), 233–282.

Vázquez, P. P., Feixas, M., Sbert, M., & Heidrich, W. (2001). Viewpoint selection using viewpoint entropy. *VMV, 1*, 273–280.

Vieira, T., Bordignon, A., Peixoto, A., Tavares, G., Lopes, H., Velho, L., & Lewiner, T. (2009). Learning good views through intelligent galleries. *Computer Graphics Forum, 28*(2), 717–726.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review, 1*(2), 202–238.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015) 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1912–1920

Yamauchi, H., Saleem, W., Yoshizawa, S., Karni, Z., Belyaev, A., & Seidel, HP. (2006) Towards stable and salient multi-view representation of 3d shapes. In: IEEE International conference on shape modeling and applications 2006 (SMI'06)

Yew, ZJ., & Lee, GH. (2018) 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In: Proceedings of the European conference on computer vision (ECCV), pp 607–623

Zaharescu, A., Boyer, E., Varanasi, K., & Horaud, R. (2009) Surface feature detection and description with applications to mesh matching. In: 2009 IEEE conference on computer vision and pattern recognition, pp 373–380

Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., & Funkhouser, T. (2017) 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition , pp 1802–1811

Zhang, Y., & Fei, G. (2019). Overview of 3d scene viewpoints evaluation method. *Virtual Reality & Intelligent Hardware, 1*(4), 341–385.

Zhao, L., Liang, S., Jia, J., & Wei, Y. (2015). Learning best views of 3d shapes from sketch contour. *The Visual Comptures, 31*(6–8), 765–774.

Zhao, S., & Ooi, W. T. (2016). Modeling 3d synthetic view dissimilarity. *The Visual Comptures, 32*(4), 429–443.

Zhu, K., Chen, W., Zhang, W., Song, R.,& Li, Y. (2020) Autonomous robot navigation based on multi-camera perception. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 5879–5885