Mesh Saliency: An Independent Perceptual Measure or A Derivative of Image Saliency?

Ran Song[1,2] Wei Zhang[1,2,*] Yitian Zhao[3] Yonghuai Liu[4] Paul L. Rosin[5]

[1] School of Control Science and Engineering, Shandong University, China
[2] Institute of Brain and Brain-Inspired Science, Shandong University, China
[3] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, China
[4] Department of Computer Science, Edge Hill University, UK
[5] School of Computer Science and Informatics, Cardiff University, UK

{ransong,davidzhang}@sdu.edu.cn,yitian.zhao@nimte.ac.cn,liuyo@edgehill.ac.uk,RosinPL@cardiff.ac.uk

Abstract

While mesh saliency aims to predict regional importance of 3D surfaces in agreement with human visual perception and is well researched in computer vision and graphics, latest work with eye-tracking experiments shows that state-of-the-art mesh saliency methods remain poor at predicting human fixations. Cues emerging prominently from these experiments suggest that mesh saliency might associate with the saliency of 2D natural images. This paper proposes a novel deep neural network for learning mesh saliency using image saliency ground truth to 1) investigate whether mesh saliency is an independent perceptual measure or just a derivative of image saliency and 2) provide a weakly supervised method for more accurately predicting mesh saliency. Through extensive experiments, we not only demonstrate that our method outperforms the current state-of-the-art mesh saliency method by 116% *and* 21% *in terms of linear correlation coefficient and AUC respectively, but also reveal that mesh saliency is intrinsically related with both image saliency and object categorical information. Codes are available at* https://github.com/rsong/MIMO-GAN.

1. Introduction

Mesh saliency, first proposed by the seminal paper of Lee *et al.* [16], measures regional importance of 3D surfaces in accordance with human visual perception. While many methods [5, 23, 25, 26, 17] for mesh saliency have been presented since then, recent eye-tracking work [34, 33, 15] shows that state-of-the-art mesh saliency methods are poor at predicting human fixations. In particular, Lavoué *et al.* [15] found that even a simple centre-bias model, a prior widely used for predicting saliency of 2D natural images, generated better results for various 3D meshes than the state-of-the-art mesh saliency methods including [16, 26, 19, 17]. Apart from the centre bias, mesh saliency and image saliency also have other characteristics in common. For instance, it was found that some features such as facial areas of people or animals always attract human fixations no matter whether they are expressed by 2D images or 3D meshes.

Image saliency is mainly driven by colour and texture while the detection of mesh saliency relies largely on object geometry. But the findings above give us an impression that despite such a fundamental difference, mesh saliency might be a derivative of image saliency rather than an independent perceptual measure. To explore this proposition, we proposes to learn *mesh* saliency from ground-truth saliency of general 2D images. In addition, it has been shown that 3D objects of the same category usually have similar saliency distributions [2, 15]. One explanation is that the information vital for object classification is usually also important for saliency as it can help humans to recognise an object swiftly without the need for scrutinizing its details [27]. Therefore, considering that there already exist large-scale public datasets for image saliency (e.g. SALICON Dataset [10], MIT Saliency Benchmark [1] and DUT-OMRON Dataset [38]) and 3D object classification (e.g. ModelNet [37] and ShapeNet [21]), we present a weakly supervised deep neural network for mesh saliency trained jointly with saliency maps of 2D images and category labels of 3D objects.

Importantly, such a weakly supervised method is potentially of broad interest as gathering eye-fixation data for 3D objects is a notoriously laborious task [12, 34, 33, 15]. To the best of our knowledge, all existing fixation datasets for mesh saliency are very small (e.g. 5 objects in [12], 15 objects in [34], 16 objects in [33] and 32 objects in [15]). The

[*] Corresponding author

consequence of using such a small dataset to train a neural network that cannot be sufficiently deep (for avoiding overfitting) is that it usually failed to generalise across a diversity of objects [33]. In this paper, we shall demonstrate that with the training data of image saliency and object category labels, our weakly supervised method accurately predicts ground-truth fixations of various 3D objects. Specifically, in the view-dependent set-up, our method outperforms the state-of-the-art mesh saliency method by 116% and 21% in terms of linear correlation and AUC respectively on the currently largest fixation dataset [15] for mesh saliency.

The core of the proposed method is a Multi-Input Multi-Output Generative Adversarial Network (MIMO-GAN). It contains two input-output paths: a regression path for pixel-level saliency prediction and a classification path for object-level recognition. The two paths essentially enable transfer learning from image saliency and 3D object classification to mesh saliency. And, since projected 2D views of 3D meshes appear highly different from 2D natural scene images, we propose to use a GAN architecture so that transfer learning is compelled to minimise the gap between image saliency and mesh saliency as much as possible.

Overall, the contribution of our work is threefold:

- We propose a novel method for mesh saliency trained with image saliency and object category labels in a weakly supervised manner and thus does not need the expensive collection of human fixations for 3D objects.
- We reveal and validate that 1) image saliency helps predict mesh saliency even though 2D natural images appear highly different from projected 2D views of 3D meshes and 2) mesh saliency also associates with class membership of meshes.
- We demonstrate that our method significantly outperforms existing state-of-the-art approaches to mesh saliency on publicly available datasets in both view-dependent and independent set-ups.

2. Related work

Mesh saliency has been widely explored in computer vision and graphics. This section categorises the methods for mesh saliency into two groups depending on whether a method is based on handcrafted features or learning.

Mesh saliency via handcrafted features. Early mesh saliency methods exploited handcrafted geometric features. Lee *et al.* [16] computed mesh saliency using a centre-surround operator on Gaussian-weighted curvatures at multiple scales. Kim *et al.* [12] later demonstrated that such a mechanism has better correlation with human fixations than both random and curvature-based models. Gal and Cohen-Or [5] introduced a salient geometric feature based on curvatures characterizing a local partial shape functionally. Shilane and Funkhouser [23] developed a method for computing salient regions of a 3D surface by describing local shape geometry through a Harmonic Shape Descriptor.

Some methods also investigated global handcrafted features as saliency depends on global features of geometry according to some psychological evidence [30, 35, 13]. For example, Wu *et al.* [36] proposed an approach based on the observation that salient features are both locally prominent and globally rare. Song *et al.* [26] analysed the log-Laplacian spectrum of meshes and presented a method capturing global information in the spectral domain. Wang *et al.* [32] detected mesh saliency using low-rank and sparse analysis in a feature space which encodes global structure information of the mesh. Leifman *et al.* [17] proposed to detect surface regions of interest by looking for regions that are distinct both locally and globally where the global consideration is whether the object is ‘limb-like’ or not.

Mesh saliency via learning. Since mesh saliency reasons about human visual perception on 3D data, it is natural to consider learning saliency from data generated by human subjects. However, due to the aforementioned training data problem, existing learning-based methods rely mainly on shallow learning. For example, Chen *et al.* [2] learned a regression model from a small dataset of 400 meshes to predict the so-called Schelling distribution. It is essentially a shallow learning scheme using a selection of handcrafted features. Lau *et al.* [14] proposed the well-defined concept of tactile mesh saliency and human subjects tend to give highly consistent responses in the process of data collection. Even so, only 150 meshes were collected for both training and testing. Similar to [14] which proposed a 6-layer toy network, Wang *et al.* [33] designed a 5-layer convolutional neural network (CNN) to predict human eye fixations on 3D objects as they only collected a set of 16 objects.

It can be seen that due to the concern about overfitting, existing methods based on supervised learning cannot make good use of neural networks sufficiently deep to learn well-generalised salient features. To address this problem, Song *et al.* [27] proposed a weakly supervised method for learning mesh saliency from class membership of meshes. Li *et al.* [18] developed an unsupervised method for detecting distinctive regions on 3D meshes. The two methods avoided the training relying on vertex-level saliency annotations but were not evaluated with eye fixation ground truth.

3. Method

The pipeline of our method is illustrated in Fig. 1. In this section, we first describe each of its components in a piecewise manner. Then, we elaborate the implementation as a whole for both training and inference where each component is situated in the context of the complete pipeline.

3.1. Generation of projected 2D images

Multi-view representation of 3D objects has been widely explored to adapt CNNs to 3D data. Compared to other

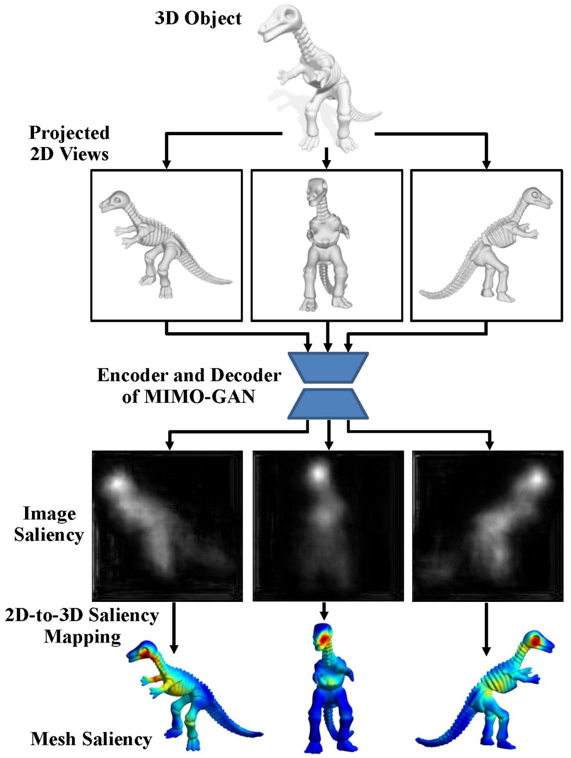

Figure 1. The pipeline of our method for generating mesh saliency.

methods for generalising deep learning to non-Euclidean domains, it shows state-of-the-art performance in various 3D object understanding tasks [28, 20, 11, 7]. In this work, we assume that each 3D object is upright oriented along the z-axis and represent it as a set of projected 2D images taken as input by the MIMO-GAN. Specifically, in the training stage, we experimented with two multi-view set-ups suggested by [28] and [27], respectively. The former created 12 rendered views for a 3D mesh with the viewpoints subject to $azimuth \in \{0, 30, \ldots, 330\}$ and $elevation = 30$, where both $azimuth$ and $elevation$ are measured in degrees. The latter produced 24 views with the same set of $azimuth$ but $elevation \in \{-30, 30\}$. The resolution of the projected images is fixed to 224×224, as required by the encoder of MIMO-GAN, no matter how many vertices the mesh contains. The projected images inherit the category labels of their corresponding mesh. In the inference stage, a given 3D mesh can be rendered either with designated viewpoints for predicting view-dependent mesh saliency, or in the way described above for generating view-independent saliency computed as the average over the saliency maps of all views.

3.2. MIMO-GAN

Fig. 2 illustrates the architecture of our MIMO-GAN. Its inputs include projected 2D images of 3D objects annotated with their category labels and 2D natural images annotated with pixel-wise saliency maps recording human fixations. As a weakly supervised network, the MIMO-GAN predicts pixel-wise saliency maps for projected 2D images based on the two types of inputs. As we mentioned above, the design of the MIMO-GAN is motivated by two observations. First, image saliency and mesh saliency have common characteristics such as centre bias and identical salient regions on some objects. Second, 3D objects of the same class usually have similar saliency distributions as the informative features important for distinguishing a 3D object from others belonging to different classes are likely to be detected as salient. Thus as shown in Fig. 2, after a shared encoder consisting of typical convolutional blocks, the MIMO-GAN branches into two paths. One is the classification path ending with the classification loss L_C which ensures that the feature extraction for saliency prediction is subject to object classification. The other is the saliency path which generates pixel-wise saliency maps via a decoder and leads to the saliency loss L_S. This path encourages the encoder and decoder to produce saliency maps of 2D natural images consistent with the corresponding fixation ground truth.

These two paths hardly impose the consistency between the saliency of natural images and that of the 2D projected views of 3D meshes to any extent, and consequently there is no guarantee that a sufficient amount of desirable characteristics of image saliency are effectively transferred into mesh saliency through the learning. Hence, a GAN architecture is further introduced to force the predicted saliency of projected 2D images of 3D meshes to be indistinguishable from that of 2D natural images. Each component of the MIMO-GAN is elaborated in the following.

Encoder. We employ the convolutional blocks of the VGG16 network [24] pre-trained on ImageNet as the encoder of MIMO-GAN. To establish the classification path, we add three fully connected (FC) layers on top of the convolutional encoder. We also bring in dropout layers next to the first and the second FC layers respectively to reduce potential overfitting as the entire network already contains a relatively large number ($\approx$ 24.9M) of trainable parameters.

Decoder/Generator. The decoder of the MIMO-GAN also acts as the generator that produces monotone saliency maps (see Fig. 1) with the same dimension as the input images. It is an expansive path including five up-convolutional blocks. Except for the first one which only contains an upsampling layer and a convolutional layer, a typical up-convolutional block consists of a 2×2 upsampling layer, a 2×2 convolutional layer that halves the number of feature channels, a concatenation with a skip-connection to a particular convolutional layer from the encoder, and one 3×3 convolution, each followed by a ReLU. Note that skip-connection has been widely used to preserve local features for image segmentation. In the MIMO-GAN, differing from most skip-connections, an extra separable convolution is used to encode the feature map output by a particular convolutional layer from the encoder and reduce its number of

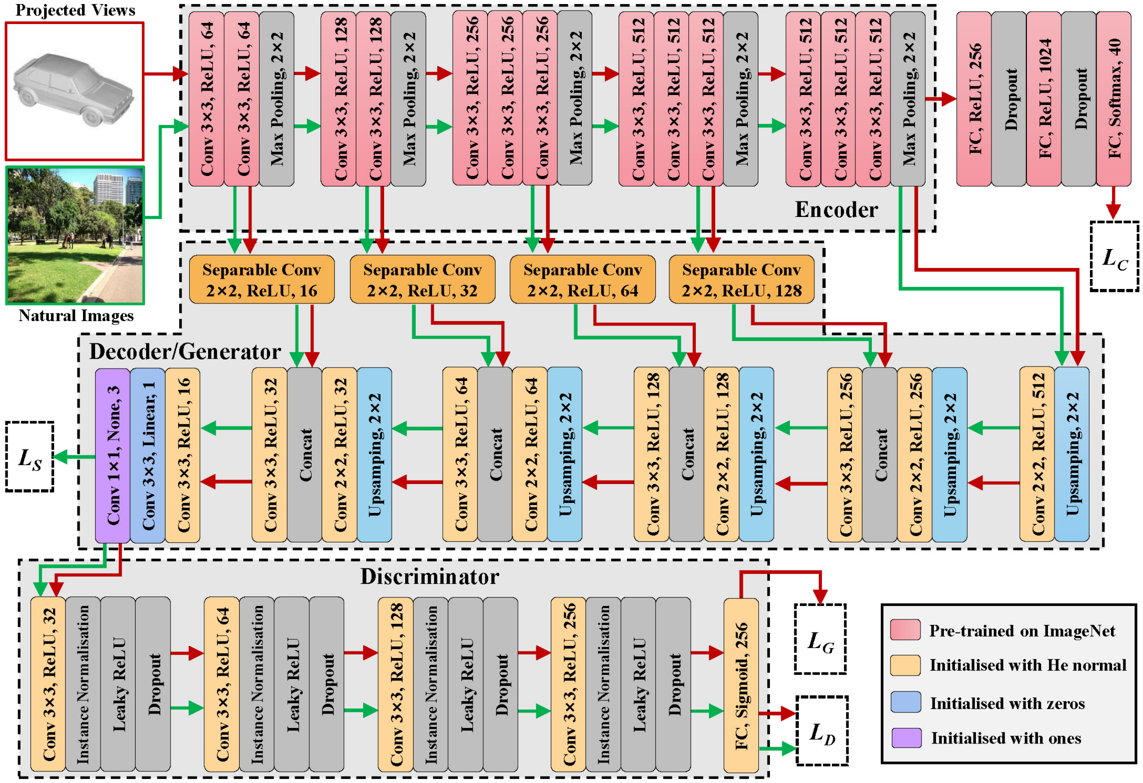

Figure 2. MIMO-GAN architecture. The MIMO-GAN takes as input projected 2D views of 3D objects and natural images, and is trained with an object classification loss L_C, an image saliency loss L_S and a GAN loss including a generator loss L_G and a discriminator loss L_D. In the inference stage, only the encoder and the decoder/generator are needed.

channels to half of the output dimension of the 2×2 convolution. This is because skip-connection applied within image segmentation focuses significantly on local details while humans can quickly attend to salient features without a slow process of scrutinising details [9]. Thus in the MIMO-GAN, the skip-connection via separable convolution ensures that features corresponding to local details just have a relatively small contribution to the concatenation.

Discriminator. For natural images with ground-truth saliency maps provided, the decoder can be trained with the saliency loss L_S, which enables an effective learning of image saliency. However, such saliency maps are not available for projected 2D views of 3D objects which appear highly different from natural images as shown in Fig. 2. This means that a specific mechanism is needed to guide the learning process of the decoder so that it can also effectively learn the saliency of projected 2D views. Considering the observation that image saliency and mesh saliency have some attributes in common, we propose a discriminator to form a GAN architecture, in order to impose consistency between the two types of saliency. In other words, although projected 2D views of 3D objects and natural images are visually different, the discriminator tends to make the generated saliency maps of projected views indistinguishable from those of natural images in the learned feature space.

As shown in Fig. 2, the discriminator consists of four convolutional blocks and one FC layer activated by the sigmoid function. In each convolutional block, a convolutional layer with ReLU activation and stride 2 for downsampling is followed by an instance normalisation (IN) layer. Experimentally, we found that IN outperforms batch normalisation. This finding is in accordance with many style transfer papers [31, 8] suggesting that IN is a good choice for a generative network as it is more adaptive to individual images.

3.3. 2D-to-3D saliency mapping

Given that MIMO-GAN generates a 2D saliency map $I(V)$ for a projected 2D view V of a 3D mesh, we employ the 2D-to-3D saliency mapping scheme proposed by Song *et al.* [27] to output a 3D saliency map. The saliency $S_m(V)$ of a 3D vertex m visible in V is computed as

$$S_m(V) = \exp(1 - Z(m)) / \exp(1 - I_i(V)) \qquad (1)$$

where $I_i(V)$ denotes the saliency of the pixel i closest to the 2D projection of m in V. $Z(m)$ is the average of the normalised distances between m and its 1-ring neighbours, which reflects the local density of vertices. The rationale of Eq. (1) is that if the local density around the vertex is low, then the 2D projection of a 3D vertex is more ambiguous and thus the 2D-to-3D correspondence is less reliable.

3.4. Implementation

Training. We first render a mesh representing a 3D object as multiple projected 2D images as described in Section 3.1 using a standard OpenGL renderer with perspective projection mode. The strengths of the ambient light, the diffuse light and the specular reflection are set to 0.3, 0.6 and 0 respectively. We apply the light uniformly across each triangular face of the mesh (i.e. flat shading). Using different illumination models or shading coefficients does not affect our method due to the invariance of the learned convolutional filters to illumination changes. All projected images are then printed at 200 dpi, also in the OpenGL mode, and further resized to the resolution of 224×224. Then we feed the projected 2D images of a collection of 3D objects and a set of natural images into the MIMO-GAN. As shown in Fig. 2, the MIMO-GAN is trained with four loss functions.

L_C denotes the loss of object classification based on a projected 2D view V, calculated as the cross-entropy loss:

$$L_C = -\sum_{c=1}^{C} \mathcal{Q}_c(V) \cdot \log\left(\mathcal{P}_c(V)\right) \tag{2}$$

where $\mathcal{Q}$ denotes the ground-truth class label of each 3D object inherited by its 2D projected views and $\mathcal{P}$ is the output of the final FC layer in the classification path of MIMO-GAN. Here $C = 40$ as we trained MIMO-GAN with ModelNet40 [37] which collected 3D objects of 40 classes.

L_S denotes the loss for predicting the saliency of a natural image I containing n pixels, calculated as the $L2$ loss:

$$L_S = \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{S}(I_i) - G(E(I_i))\right)^2 \tag{3}$$

where $\mathcal{S}$ denotes the ground truth saliency map of each natural image. G and E represent the generator and the encoder of the MIMO-GAN respectively.

The GAN loss comprises the generator loss L_G and the discriminator loss L_D, calculated as

$$\begin{aligned} L_G &= \log(1 - D(G(E(V)))) \quad \text{and} \\ L_D &= -\log(D(G(E(I))) - \log(1{-}D(G(E(V)))) \end{aligned} \tag{4}$$

where D denotes the discriminator of the MIMO-GAN.

The overall loss is a weighted sum of the four losses:

$$L_{all} = \lambda_1 L_C + \lambda_2 L_S + \lambda_3 L_G + \lambda_4 L_D \tag{5}$$

where λ_1, λ_2, λ_3 and λ_4 are set to 0.2, 1, 0.01 and 0.01 respectively through empirical observations.

We trained the MIMO-GAN with learning rate 0.001 through stochastic gradient descent and observed that it usually converged within 100 epochs.

Inference. Once the MIMO-GAN is trained, we only need its encoder and decoder for inference as shown in Fig. 1. First, we produce a set of projected images for a testing mesh with designated viewpoints using the same rendering settings as those in training. Then the projected images are fed into the MIMO-GAN to infer 2D saliency maps (output by the layer coloured purple in Fig. 2). Finally, each 2D saliency map is converted into a view-dependent mesh saliency map by the scheme described in Section 3.3.

Note that our method can also be used to produce view-independent mesh saliency while human eye fixations depend on the viewpoint. In this set-up, we render a mesh as multiple projected views as described in Section 3.1 and generate a 2D saliency map for each of them. After mapping these 2D saliency maps to 3D mesh saliency maps, we compute the view-independent mesh saliency as the average over the mesh saliency maps across all views.

4. Experimental results

All experiments were conducted on a computer with an Intel Core i9-9900K CPU, 64GB of RAM and a NVIDIA RTX 2080Ti GPU. Unless otherwise specified, we use the 24-view set-up for the MIMO-GAN. More experimental results are available in the supplementary material.

4.1. Training and testing datasets

We train the MIMO-GAN using two publicly available datasets. One is the Princeton ModelNet40 dataset [37] containing $4,000$ meshes from 40 common object categories where all meshes are upright oriented by the method proposed in either [4] or [22]. The other is the training set of the SALICON Dataset [10] comprising $10,000$ natural scene images with ground-truth saliency annotations.

We select the 3D visual attention (3DVA) dataset [15] containing 32 meshes for testing. To the best of our knowledge, it is the largest dataset (by the number of 3D objects) for evaluating mesh saliency methods with ground-truth fixation maps on 3D meshes. In the 3DVA dataset, the fixations of each mesh are gathered from three designated viewpoints and are view-dependent. It is noteworthy that Wang *et al.* [33] concluded that "salient features exhibit a tendency to be view-dependent". Nevertheless, to address the concern over the performance of our method for predicting view-independent mesh saliency, we also evaluate it with the Schelling dataset [2] which provides view-independent 3D interest points selected by human subjects for a collection of 400 meshes belonging to 20 object categories.

4.2. Evaluation with the 3DVA dataset

Fig. 3 shows the saliency maps of various 3D objects produced by our method and the corresponding human fixation maps. One observation is that these saliency maps are highly consistent with the human eye fixations. We can see that our method typically detects one or two large "blob-like" areas as salient, which accords with the ground truth. In comparison, Fig. 4 shows that other methods highlight disconnected small-scale local features such as the small rings on the wings of the gargoyle, the ears and the feet of the horse, and the fingers and the toes of the human. Another observation is that some objects of the same class have

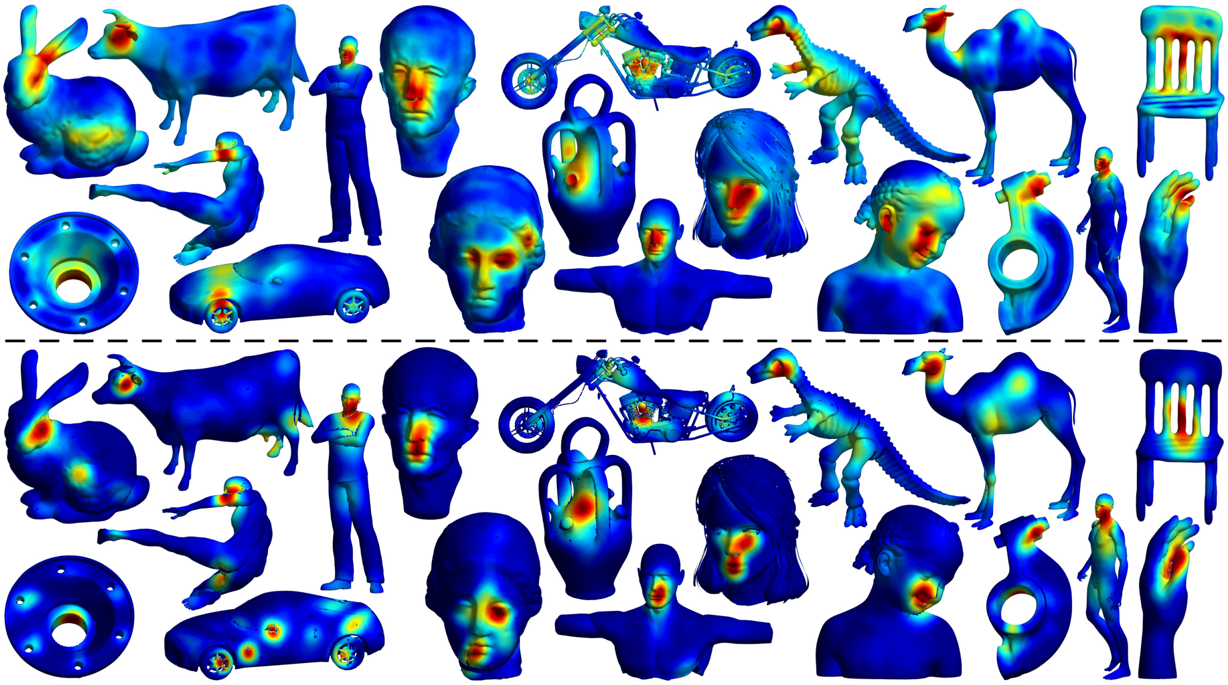

Figure 3. A gallery of mesh saliency detected by our method (top half) with the ground truth fixation maps (provided by the 3DVA dataset [15]) of the corresponding meshes (bottom half). Warmer colours show higher saliency.

analogous saliency distributions. For instance, facial areas of human and animal objects are usually detected as salient.

We use linear correlation coefficient (LCC) and area under the ROC curve (AUC) as suggested by [15] to quantitatively measure the similarity between a saliency map produced by a competing method and a ground truth fixation map. According to [15], to calculate the AUC scores, the ground truth fixation maps are thresholded to be converted into binary maps so that 20% of visible vertices are considered as fixations. The saliency map is then treated as a binary classifier of these fixations. The ROC curve represents the relationship between the probability of false positives and the probability of true positives and is obtained by varying the decision threshold on the saliency map.

Tables 1 and 2 show the overall performance of a selection of competing methods for mesh saliency and our MIMO-GAN with different ablation (see Section 4.4 for details) and multi-view (see Section 3.1) set-ups on the 3DVA dataset in terms of LCC and AUC. For LCC, 1 represents perfect positive linear relation, 0 represents no relation and -1 represents perfect negative relation. For AUC, 1 represents a perfect classification while 0.5 represents a random one. Both metrics demonstrate the overwhelming superiority of our method over all competing methods. It can be seen that the 24-view set-up outperforms the 12-view set-up. Adding further views is trivial, however, we found that our MIMO-GAN with the 24-view set-up already achieved high performance and using more views cannot further lead to a significant improvement. Specifically, it outperforms the current state-of-the-art method (i.e. CfS-CNN [27]) by 116% and 21% in terms of LCC and AUC, respectively. The quantitative results indicate that 1) mesh saliency that predicts human visual attention on 3D surfaces might be perceptually related to 2D image saliency and categorical information of 3D objects, and 2) our method that combines the two types of knowledge via a GAN framework for detecting mesh saliency is computationally effective.

We have conducted tests by adding Gaussian noise with $\sigma = 0.001B, 0.002B$ and $0.004B$ respectively to all meshes in the 3DVA dataset where B is the length of the diagonal of the bounding box of the mesh. Table. 3 lists the results of detecting saliency on the noisy meshes using our method, which demonstrates its robustness against noise.

4.3. Evaluation with the Schelling dataset

Apart from human eye fixations, human-picked 3D interest points have also been used for evaluating mesh saliency methods [3, 27]. The Schelling dataset [2] collected 3D interest points by asking people to "select points on the surface of a 3D object likely to be selected by other people". To generate a view-independent saliency map from the scattered interest points for quantitative evaluation, we employ a strategy widely used for evaluating image saliency methods [1, 10]: we project a Gaussian distribution on a mesh where each vertex is labelled by either 1 (representing interest point) or 0 (representing non-interest point) and vary the standard deviation to generate different versions of the ground truth saliency maps. When we evaluate our method

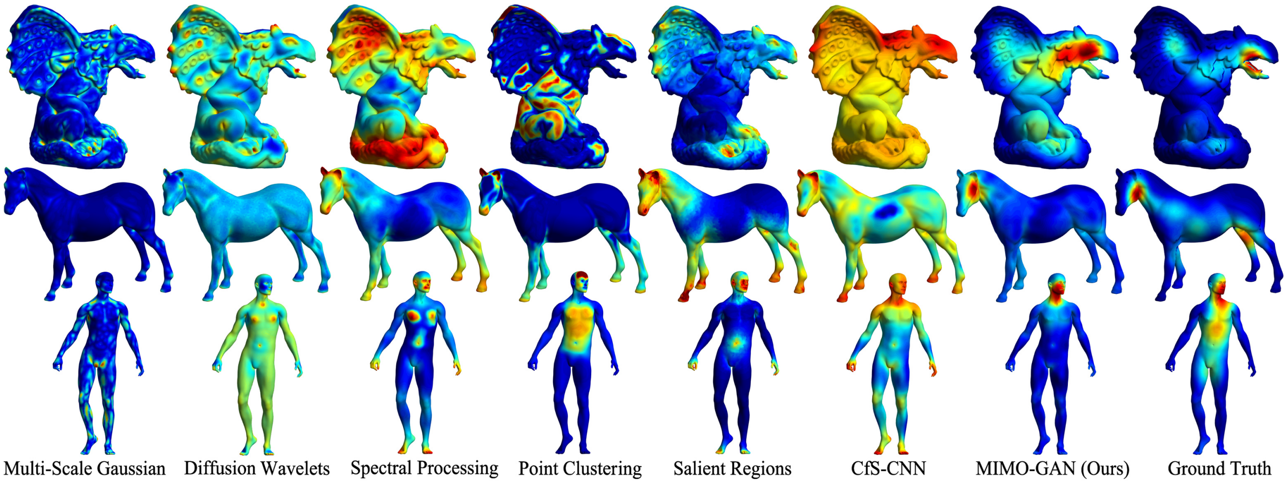

Figure 4. Comparisons of mesh saliency detected by different methods. From left to right: Multi-Scale Gaussian [16], Diffusion Wavelets [6], Spectral Processing [26], Point Clustering [19], Salient Regions [17], CfS-CNN [27], the proposed MIMO-GAN and the ground truth fixation maps provided by the 3DVA dataset [15]. Comparative results of more objects are available in the supplementary material.

Method	mean LCC	SD of LCC
Multi-Scale Gaussian [16]	0.131	0.265
Diffusion Wavelets [6]	0.088	0.222
Spectral Processing [26]	0.078	0.253
Point Clustering [19]	0.132	0.300
Salient Regions [17]	0.215	0.245
CfS-CNN [27]	0.226	0.243
MIMO-GAN-A1	0.329	0.254
MIMO-GAN-A2	0.134	**0.193**
MIMO-GAN-A3	0.477	0.221
MIMO-GAN w/ 12 views	0.451	0.226
MIMO-GAN w/ 24 views	**0.489**	0.212

Table 1. Performance of mesh saliency methods on the 3DVA dataset [15] in terms of the mean and the standard deviation (SD) of linear correlation coefficient (LCC).

Method	mean AUC	SD of AUC
Multi-Scale Gaussian [16]	0.593	0.170
Diffusion Wavelets [6]	0.558	0.143
Spectral Processing [26]	0.553	0.154
Point Clustering [19]	0.583	0.183
Salient Regions [17]	0.628	0.149
CfS-CNN [27]	0.643	0.150
MIMO-GAN-A1	0.699	0.137
MIMO-GAN-A2	0.599	0.126
MIMO-GAN-A3	0.763	0.120
MIMO-GAN w/ 12 views	0.741	0.123
MIMO-GAN w/ 24 views	**0.780**	**0.112**

Table 2. Performance of mesh saliency methods on the 3DVA dataset [15] in terms of the mean and the standard deviation (SD) of area under the ROC curve (AUC).

Noise amount	no noise	$0.001B$	$0.002B$	$0.004B$
mean LCC	0.489	0.480	0.472	0.457
mean AUC	0.780	0.768	0.768	0.761

Table 3. Evaluation of the robustness of our method against noise.

on these ground truth maps, we essentially estimate whether it can detect saliency at different scales.

Note that as demonstrated in [15], Schelling/interest points and human fixations are not correlated. Although we do not intend to argue which kind of data is more suitable for evaluating mesh saliency methods, this means that a method which performs well on the 3DVA dataset is likely to have a relatively poor performance on the Schelling dataset. However, Table 4 demonstrates that our MIMO-GAN for predicting view-independent mesh saliency is still the top performing method on the Schelling dataset. In particular, the results show that compared to other methods, the MIMO-GAN is effective at detecting saliency at relatively large scales. This finding is consistent with the qualitative results shown in Figs. 3 and 4 where our method often highlights one or two large areas. We also provide quantitative evaluation per category in the supplementary material.

Interestingly, Fig. 5 shows that apart from facial areas, our method also tends to concentrate on some long protrusions of 3D objects in a view-independent set-up. This is because our method computes view-independent mesh saliency as the average over the saliency maps across all views as mentioned in the end of Section 3.4. Since long protrusions are likely to be visible in most views, their saliency are usually high due to such a 'visibility bias', which might result in poor saliency computation for objects with many highly occluded regions.

4.4. Is mesh saliency a derivative of image saliency?

In this section, we evaluate different configurations of MIMO-GAN with the 24-view set-up to understand

Method	$\sigma = 0.1B$	$\sigma = 0.12B$	$\sigma = 0.14B$	$\sigma = 0.16B$	$\sigma = 0.18B$	$\sigma = 0.2B$
Multi-Scale Gaussian [16]	0.223	0.213	0.202	0.193	0.186	0.179
Diffusion Wavelets [6]	0.101	0.091	0.082	0.074	0.068	0.063
Spectral Processing[26]	0.324	0.322	0.313	0.301	0.293	0.284
Salient Regions [17]	0.437	0.421	0.402	0.376	0.360	0.340
CfS-CNN [27]	**0.455**	0.457	0.454	0.447	0.439	0.427
MIMO-GAN w/ 24 views	0.447	**0.462**	**0.470**	**0.472**	**0.470**	**0.463**

Table 4. Performance of saliency methods on the Schelling dataset [2] in terms of linear correlation coefficient (LCC). σ is the standard deviation of the Gaussian used to generate the pseudo ground truth. B is the length of the diagonal of the bounding box of the mesh.

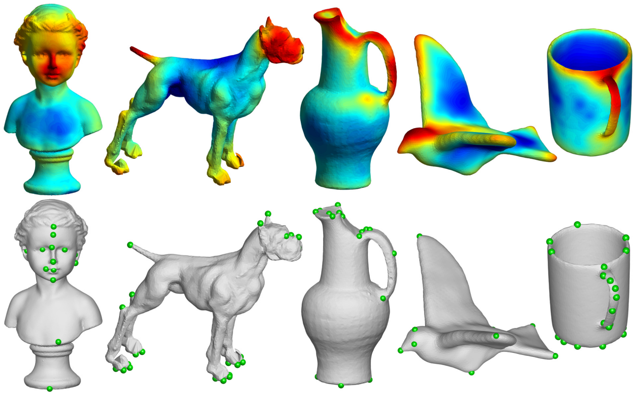

Figure 5. View-independent mesh saliency detected by our method and the human-picked interest points (Schelling points [2]).

whether and to what degree mesh saliency is a derivative of image saliency. We thus conduct three ablation studies:

(1) Remove the FC layers and the classification loss L_C from the MIMO-GAN so that its training relies only on the saliency loss and the GAN loss.
(2) Remove the saliency loss L_S so that the training relies only on the classification loss and the GAN loss.
(3) Remove the discriminator as well as the GAN loss including the generator loss L_G and the discriminator loss L_D so that the training relies only on L_C and L_S.

With a slight abuse of terminology, the three ablated versions of MIMO-GAN are named as MIMO-GAN-A1, MIMO-GAN-A2 and MIMO-GAN-A3 respectively.

According to the quantitative results listed in Tables 1 and 2, we can see that all ablated methods suffer from a degraded performance compared to the full version of MIMO-GAN. Among them, MIMO-GAN-A2 is the worst affected one although it still outperforms most of the competing methods for mesh saliency. In comparison, MIMO-GAN-A1 performs significantly better than it, which indicates that image saliency has a much greater impact than object categorical information on mesh saliency. Particularly, we can see that MIMO-GAN-A1 which essentially learns mesh saliency from image saliency already outperforms all competing methods. This suggests that mesh saliency which predicts human visual attention on 3D objects depends heavily on image saliency which predicts where human observers look in natural scene images. However, the considerable superiority of MIMO-GAN-A3 over MIMO-GAN-A2 as shown in Tables 1 and 2 demonstrates that categorical information of 3D objects also brings in a significant performance gain for mesh saliency on top of image saliency. One explanation is that the human perception system tends to capture the most informative features as salient [29] since it can help humans to recognise an object swiftly without the need for scrutinizing all of its details. Thus we argue that the informative features important for distinguishing a 3D object from others belonging to different classes are highly likely to be detected as salient.

Hence, our view is that although the prediction of mesh saliency benefits substantially from image saliency, it cannot be regarded as a derivative of image saliency as it is also influenced by other factors such as object categorical information which provides useful knowledge largely independent of image saliency for mesh saliency.

5. Conclusions

Aiming at the fact that existing methods for mesh saliency are poor at predicting human fixations on 3D objects, we propose the MIMO-GAN that combines image saliency and object category labels to effectively solve this problem. The MIMO-GAN is trained with publicly available datasets of image saliency and 3D object classification in a weakly supervised manner and thus does not require the expensive collection of fixation data for 3D objects. Therefore, it is potentially of broad interest in the community. Importantly, our work reveals and demonstrates that mesh saliency cannot be simply viewed as a derivative of image saliency although it is significantly influenced by image saliency. This is because the categorical information of 3D objects also has a great impact on it. We believe that these new insights into mesh saliency will further stimulate research on human visual perception for 3D objects and even scenes that contain multiple objects.

Acknowledgements. We gratefully acknowledge the support of the National Natural Science Foundation of China under Grants 62076148, 61991411 and U1913204, the Young Taishan Scholars Program of Shandong Province No.tsqn201909029 and Shandong Major Scientific and Technological Innovation Project 2018CXGC1503.

References

[1] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2018. 1, 6

[2] Xiaobai Chen, Abulhair Saparov, Bill Pang, and Thomas Funkhouser. Schelling points on 3d surface meshes. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):29, 2012. 1, 2, 5, 6, 8

[3] Xiaoying Ding, Weisi Lin, Zhenzhong Chen, and Xinfeng Zhang. Point cloud saliency detection by local and global feature fusion. *IEEE Trans. Image Process.*, 28(11):5379–5393, 2019. 6

[4] Hongbo Fu, Daniel Cohen-Or, Gideon Dror, and Alla Sheffer. Upright orientation of man-made objects. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 27(3):42, 2008. 5

[5] Ran Gal and Daniel Cohen-Or. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph.*, 25(1):130–150, 2006. 1, 2

[6] Tingbo Hou and Hong Qin. Admissible diffusion wavelets and their applications in space-frequency processing. *IEEE Trans. Vis. Comput. Graph.*, 19(1):3–15, 2013. 7, 8

[7] Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir G Kim, and Ersin Yumer. Learning local shape descriptors from part correspondences with multiview convolutional networks. *ACM Trans. Graph.*, 37(1):6, 2018. 3

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. ICCV*, pages 1501–1510, 2017. 4

[9] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognitive psychology*, 43(3):171–216, 2001. 4

[10] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proc. CVPR*, pages 1072–1080, 2015. 1, 5, 6

[11] Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3D shape segmentation with projective convolutional networks. In *Proc. CVPR*, volume 1, page 8, 2017. 3

[12] Youngmin Kim, Amitabh Varshney, David Jacobs, and François Guimbretière. Mesh saliency and human eye fixations. *ACM Trans. Appl. Percept.*, 7(2):12:1–12:13, 2010. 1, 2

[13] Christof Koch and Tomaso Poggio. Predicting the visual world: silence is golden. *Nat. Neurosci.*, 2:9–10, 1999. 2

[14] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. Tactile mesh saliency. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 35(4), 2016. 2

[15] Guillaume Lavoué, Frédéric Cordier, Hyewon Seo, and Mohamed-Chaker Larabi. Visual attention for rendered 3d shapes. *Comput. Graph. Forum (Proc. Eurographics)*, pages 414–421, 2018. 1, 2, 5, 6, 7

[16] Chang Ha Lee, Amitabh Varshney, and David W Jacobs. Mesh saliency. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 24(3):659–666, 2005. 1, 2, 7, 8

[17] George Leifman, Elizabeth Shtrom, and Ayellet Tal. Surface regions of interest for viewpoint selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(12):2544–2556, 2016. 1, 2, 7, 8

[18] Xianzhi Li, Lequan Yu, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Unsupervised detection of distinctive regions on 3d shapes. *ACM Trans. Graph.*, 39(5):1–14, 2020. 2

[19] Flora Ponjou Tasse, Jiri Kosinka, and Neil Dodgson. Cluster-based point set saliency. In *Proc. ICCV*, pages 163–171, 2015. 1, 7

[20] Charles Ruizhongtai Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas Guibas. Volumetric and multi-view CNNs for object classification on 3d data. In *Proc. CVPR*, pages 5648–5656, 2016. 3

[21] Manolis Savva, Angel X Chang, and Pat Hanrahan. Semantically-enriched 3D models for common-sense knowledge. In *Proc. CVPR Workshops*, pages 24–31, 2015. 1

[22] Nima Sedaghat, Mohammadreza Zolfaghari, Ehsan Amiri, and Thomas Brox. Orientation-boosted voxel nets for 3d object recognition. In *Proc. BMVC*, 2017. 5

[23] Philip Shilane and Thomas Funkhouser. Distinctive regions of 3d surfaces. *ACM Trans. Graph.*, 26(2):7, 2007. 1, 2

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 3

[25] Ran Song, Yonghuai Liu, Ralph R. Martin, and Paul L. Rosin. Saliency-guided integration of multiple scans. In *Proc. CVPR*, pages 1474–1481, 2012. 1

[26] Ran Song, Yonghuai Liu, Ralph R. Martin, and Paul L. Rosin. Mesh saliency via spectral processing. *ACM Trans. on Graph.*, 33(1), 2014. 1, 2, 7, 8

[27] Ran Song, Yonghuai Liu, and Paul L. Rosin. Mesh saliency via weakly supervised classification-for-saliency CNN. *IEEE Trans. Vis. Comput. Graph.*, 21(1):151–164, 2021. 1, 2, 3, 4, 6, 7, 8

[28] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, pages 945–953, 2015. 3

[29] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proc. SGP*, pages 1383–1392, 2009. 8

[30] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cogn. Psychol.*, 12(1):97–136, 1980. 2

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proc. CVPR*, pages 6924–6932, 2017. 4

[32] Shengfa Wang, Nannan Li, Shuai Li, Zhongxuan Luo, Zhixun Su, and Hong Qin. Multi-scale mesh saliency based on low-rank and sparse analysis in shape feature space. *Comput. Aided Geom. Des.*, 35:206–214, 2015. 2

[33] Xi Wang, Sebastian Koch, Kenneth Holmqvist, and Marc Alexa. Tracking the gaze on objects in 3d: how do people really look at the bunny? *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):1–18, 2018. 1, 2, 5

[34] Xi Wang, David Lindlbauer, Christian Lessig, Marianne Maertens, and Marc Alexa. Measuring the visual salience of 3d printed objects. *IEEE Comput. Graph. Appl.*, 36(4):46–55, 2016. 1

[35] Jeremy M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994. 2

[36] Jinliang Wu, Xiaoyong Shen, Wei Zhu, and Ligang Liu. Mesh saliency with global rarity. *Graph. Models*, 46:264–274, 2013. 2

[37] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Proc. CVPR*, pages 1912–1920, 2015. 1, 5

[38] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proc. CVPR*, pages 3166–3173, 2013. 1