3D Visual Saliency: An Independent Perceptual Measure or A Derivative of 2D Image Saliency?

Ran Song, *Senior Member, IEEE,* Wei Zhang, *Senior Member, IEEE,* Yitian Zhao, *Member, IEEE,* Yonghuai Liu, *Senior Member, IEEE,* and Paul L. Rosin

Abstract—While 3D visual saliency aims to predict regional importance of 3D surfaces in agreement with human visual perception and has been well researched in computer vision and graphics, latest work with eye-tracking experiments shows that state-of-the-art 3D visual saliency methods remain poor at predicting human fixations. Cues emerging prominently from these experiments suggest that 3D visual saliency might associate with 2D image saliency. This paper proposes a framework that combines a Generative Adversarial Network and a Conditional Random Field for learning visual saliency of both a single 3D object and a scene composed of multiple 3D objects with image saliency ground truth to 1) investigate whether 3D visual saliency is an independent perceptual measure or just a derivative of image saliency and 2) provide a weakly supervised method for more accurately predicting 3D visual saliency. Through extensive experiments, we not only demonstrate that our method significantly outperforms the state-of-the-art approaches, but also manage to answer the interesting and worthy question proposed within the title of this paper.

Index Terms—3D visual saliency, weak supervision, Generative Adversarial Network, Conditional Random Field.

1 Introduction

3D visual saliency measures regional importance of 3D surfaces in accordance with human visual perception based on their 3D information. It has a range of applications such as 3D data simplification [1], volume rendering [1], 3D printing [2], viewpoint selection [3], virtual reality [4], etc. Since the *polygon mesh* is a popular representation of 3D surfaces, Lee *et al.* first proposed the concept of *mesh saliency* in their seminal paper [3] to predict human visual attention on the surface mesh of a single 3D object. While many methods [5], [6], [7], [8], [9] for mesh saliency have been presented since then, recent eye-tracking work [10], [11], [12] shows that state-of-the-art mesh saliency methods are poor at predicting human fixations. In particular, Lavoué *et al.* [12] introduced a simple centre-bias model defined as the weighted version of a saliency model by fitting a linear model. Such a centre-bias model is a prior widely used for predicting saliency on 2D natural images and they found that it generated better results for various 3D meshes than the state-of-the-art mesh saliency methods including [3], [8], [9], [13]. Apart from centre bias, mesh saliency and image saliency have other characteristics in common. For instance, user studies [14], [15] 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. Moreover, previous work [16], [17] showed that combining 2D salient features is usually an effective tactic for predicting 3D visual saliency.

It is thus interesting to investigate the relationship between 3D visual saliency and 2D image saliency, even just from a non-cognitive perspective: we focus on a computational method that leverages image saliency to predict human eye fixations in a distal 3D environment, while researchers in psychology and cognitive science [18], [19] recently attempted to study relevant issues based on the proximal stimulation it causes. For this purpose, we propose to extend the concept of mesh saliency referring only to a single 3D object represented as a surface mesh to *3D visual saliency* referring to both a 3D object and a scene containing multiple 3D objects with varying data representations such as mesh and depth map. Compared to mesh saliency, 3D visual saliency enables a more general understanding of human visual attention, and is semantically more comprehensive and computationally more flexible when exploring such a relationship that in return helps its interpretation.

Image saliency is mainly driven by colour and texture while the detection of 3D visual saliency relies largely on 3D information such as depth and surface normals. But the findings above give us an impression that despite such a fundamental difference, 3D visual saliency might be a derivative of image saliency rather than an independent perceptual measure. To explore this proposition, we propose to learn 3D visual saliency from ground-truth saliency of general 2D images. It is noteworthy that more than a decade ago, Jansen *et al.* [20] first investigated the influence of several 2D visual features that may lead to saliency on 3D visual attention by conducting a free-viewing task on the

- *Ran Song and Wei Zhang are with the School of Control Science and Engineering, Shandong University, Jinan, China. They are also with the Institute of Brain and Brain-Inspired Science at the same university. E-mail: {ransong, davidzhang}@sdu.edu.cn.*
- *Yitian Zhao is with Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China. E-mail: yitian.zhao@nimte.ac.cn.*
- *Yonghuai Liu is with the Department of Computer Science, Edge Hill University, Ormskirk, UK. E-mail: liuyo@edgehill.ac.uk.*
- *Paul L. Rosin is with the School of Computer Science and Informatics, Cardiff University, Cardiff, UK. E-mail: RosinPL@cardiff.ac.uk.*

Manuscript revised April 10, 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants 62076148, U22A2057, and 61991411, in part by the Shandong Excellent Young Scientists Fund Program (Overseas) under Grant 2022HWYQ-042 and in part by the Young Taishan Scholars Program of Shandong Province No.tsqn201909029.
(Corresponding author: Wei Zhang)

2D and 3D versions of the same set of images. However, they merely looked into low-level features including mean luminance, luminance contrast and texture contrast while today it is well-known that visual saliency also arises from high-level features [21], [22], [23].

It has been shown that 3D objects in the same category usually have similar saliency distributions [12], [14]. One explanation is that high-level features with semantic meaning, such as faces, vital for object classification are usually also important for saliency as it can help humans to recognise an object swiftly without the need for scrutinising its details or intricacies [24]. Therefore, considering that there already exist large-scale public datasets for image saliency (e.g. SALICON dataset [15], MIT Saliency Benchmark [25] and DUT-OMRON dataset [26]) and 3D object classification (e.g. ModelNet [27] and ShapeNet [28]), we propose a weakly supervised deep neural network for 3D visual saliency trained jointly with saliency maps of 2D images and category labels of 3D objects. Such a weakly supervised method is potentially of broad interest in that gathering eye-fixation data for 3D objects is notoriously laborious [10], [11], [12], [29]. To the best of our knowledge, all existing fixation datasets for visual saliency of 3D objects are very small (e.g. 5 objects in [29], 15 objects in [10], 16 objects in [11] and 32 objects in [12]). The consequence of using such a small dataset to train a neural network that cannot be sufficiently deep (for avoiding overfitting) is that it usually fails to generalise across a diversity of objects [11]. 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 on various 3D objects and scenes.

The core of our 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 3D visual saliency. Since projected 2D views of 3D objects used as network input appear highly different from 2D natural images, we introduce a GAN architecture so that transfer learning is compelled to minimise the gap between image saliency and 3D visual saliency as much as possible. We shall show in the experiments that due to the relatively simple structure of MIMO-GAN, it can be easily decomposed to facilitate ablation studies for investigating our main research question: is 3D visual saliency an independent perceptual measure or just a derivative of 2D image saliency?

We design the MIMO-GAN through a view-based representation of 3D objects in that it can handle both non-Euclidean data (e.g. meshes) and grid-structured data (e.g. depth maps). This is motivated by the fact that 1) mesh is a popular way for representing a single 3D object in various datasets and 2) view-dependent depth data are widely used for depicting 3D scenes that contain multiple objects. Thus, our method can be directly used to compute visual saliency for both a single 3D object represented by a triangle mesh and a scene represented by a depth map which contains multiple 3D objects. With such flexibility, the proposed method potentially has a wide range of applications. Moreover, as we shall show in Section 4.6, exploring 3D visual saliency for scenes brings new insight into the above research question and actually leads to an update of the answer to it. However, the MIMO-GAN merely exploits the intra-view knowledge of the projected views of a 3D object. Therefore, we propose to combine it with a Conditional Random Field (CRF) inferred by a specifically designed simulated annealing algorithm. The CRF leverages the inter-view cues to contextualise multi-view saliency maps subject to a heuristic prior, which further boosts the performance as demonstrated by the experimental results.

Overall, the contribution of our work is threefold:

- We propose a GAN architecture for 3D visual saliency trained with 2D image saliency and category labels of 3D objects in a weakly supervised manner, which avoids the expensive collection of 3D eye-tracking data.
- We conducted a user study to create a new dataset of 3D visual saliency[1], and demonstrate that our method can handle both a single object and a scene containing multiple objects with different data representations and significantly outperforms the state-of-the-art methods on both public datasets and the newly created dataset.
- We reveal through experiments that 1) 2D image saliency helps to predict 3D visual saliency even though 2D natural images appear highly different from projected 2D views of 3D objects and 2) only 3D visual saliency for a single object associates much with object categorical information while that for a scene might not.

A preliminary version of this work has been published in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21) [30][2]. The main improvements of this extended version are elaborated as follows: First, conceptually, we extend mesh saliency to 3D visual saliency by exploring visual saliency of not just a single 3D object represented as a mesh but also a 3D scene containing multiple objects represented as either a mesh or a depth map; Second, methodologically, this work takes one step forward and presents a new method, named as MIMO-GAN-CRF, where the newly added CRF component exploits inter-view cues to further improve the prediction of 3D visual saliency; Third, experimentally, more competing methods and more datasets including a new dataset of human fixations on 3D objects established via a user study are involved in the comparative evaluations; Fourth, conclusively, we update the answer to the question in the title of the paper as analysing 3D visual saliency for scenes brings new insight that cannot be derived from mesh saliency with regard to a single 3D object.

2 Related Work

3D visual saliency has been widely explored in computer vision and graphics. This section categorises its methods into two groups depending on whether a method is based on handcrafted features or learning. It is worth mentioning that the 3D visual saliency discussed in this paper is strictly limited to saliency based only on 3D data such as 3D surface meshes and depth maps. Hence, another popular concept, RGBD saliency, and its related work is out of the scope

1. The data and code of the user study are publicly available at https://github.com/rsong/3D-ViSa.

2. Data, code and the pretrained model are publicly available at https://github.com/rsong/MIMO-GAN.

of this paper as it is usually regarded as visual saliency detection with *comprehensive information* [31] extracted using both 2D and 3D cues.

3D visual saliency via handcrafted features. Early work for 3D visual saliency heavily exploited handcrafted geometric features. Lee *et al.* [3] computed the proposed mesh saliency using a centre-surround operator on Gaussian-weighted curvatures at multiple scales. Kim *et al.* [29] later demonstrated that such a mechanism has better correlation with human fixations than both random and curvature-based models. Gal and Cohen-Or [32] introduced a salient geometric feature that characterises local partial shape based on curvatures. Shilane and Funkhouser [5] developed a method for computing salient regions of a 3D surface by describing local shape geometry through a Harmonic Shape Descriptor. Song *et al.* [6] computed local visual saliency for 3D scans and used it to guide 3D surface reconstruction.

Some methods also investigated global handcrafted features as psychological evidence [33], [34] showed that human visual attention depends on global cues. For example, Wu *et al.* [7] proposed an approach based on the observation that salient features are both locally prominent and globally rare. Shtrom *et al.* [35] detected saliency in large point sets by identifying globally distinct features in a multi-level manner. Song *et al.* [8] analysed the log-Laplacian spectrum of meshes and presented a method for capturing global information in the spectral domain. Wang *et al.* [36] detected mesh saliency using low-rank and sparse analysis in a feature space which encodes global structure information of the mesh. Leifman *et al.* [9] 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. Arvanitis *et al.* [37] exploited global information through principal component analysis for predicting 3D saliency on industrial 3D objects.

3D visual saliency via learning. Since 3D visual saliency reasons about human perception on 3D data, it is natural to consider learning it 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.* [14] learned a regression model from a small dataset of 400 3D objects to predict the so-called Schelling distribution. It is essentially a shallow learning scheme using a selection of handcrafted features. Lau *et al.* [38] proposed the well-defined concept of tactile mesh saliency and found that human subjects tend to give highly consistent responses in the process of data collection. Even so, only 150 3D objects were collected for both training and test. Similar to [38] which proposed a 6-layer toy network, Wang *et al.* [11] 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.* [24] proposed a weakly supervised method for learning mesh saliency from class membership of meshes. Li *et al.* [39] developed an unsupervised method for detecting distinctive regions on 3D shapes. The two methods avoided the training relying on vertex-level saliency annotations but were not evaluated with eye fixation ground truth. Nousias *et al.* [40] trained a CNN to detect mesh saliency using pseudo ground truth generated by the handcrafted approach proposed in [37], which does not perform well at predicting real human fixations according to our experimental results.

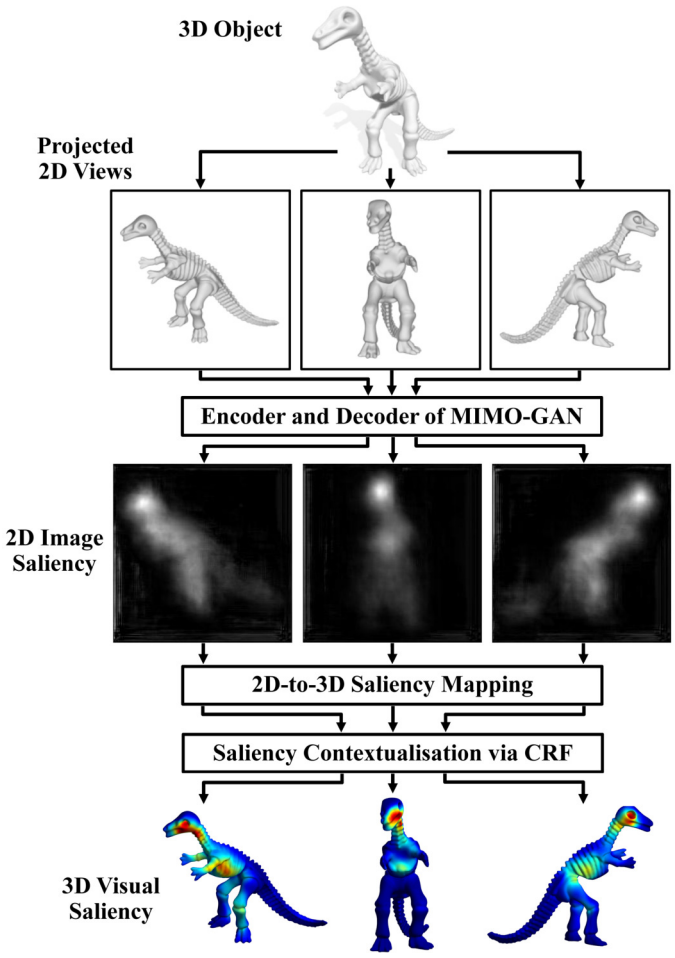


Fig. 1. The pipeline of our method based mainly on MIMO-GAN and CRF for generating 3D visual saliency.

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

View-based representation of 3D objects has been widely explored to adapt CNNs to 3D data. Compared to other methods for generalising deep learning to non-Euclidean domains, it arguably shows state-of-the-art performance in various 3D object understanding tasks [41], [42], [43], [44]. 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 [41] and [24], respectively. The

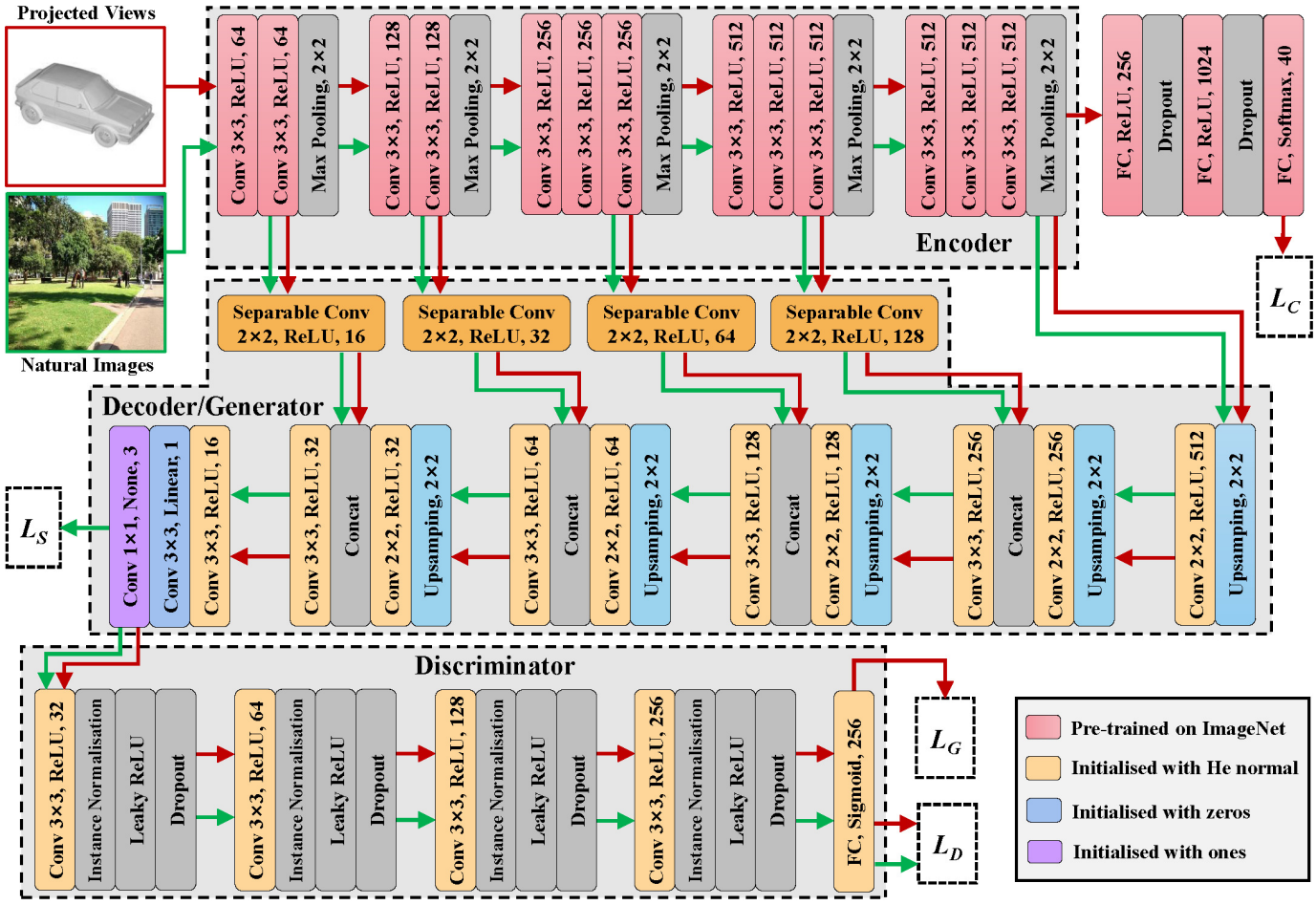


Fig. 2. MIMO-GAN architecture. The MIMO-GAN takes as input projected 2D views of 3D objects and natural images, and is trained with an object classification loss L_C, an image saliency loss L_S and a GAN loss including a generator loss L_G and a discriminator loss L_D. In the inference stage, only the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

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 3D visual saliency, or in the way described above for generating view-independent saliency computed as the average over the saliency maps of all the views.

3.2 MIMO-GAN

Fig. 2 illustrates the architecture of the 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, 2D image saliency and visual saliency of 3D objects 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 in that the informative features important for distinguishing one 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 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 objects to any extent, and consequently there is no guarantee that a sufficient amount of desirable characteristics of image saliency are effectively transferred into 3D visual saliency through the learning. Hence, a GAN architecture is further introduced to force the predicted saliency of projected 2D images of 3D objects to be indistinguishable from that of 2D natural images. Each component of the MIMO-GAN is elaborated as follows.

Encoder. We employ the convolutional blocks of the VGG16 network [45] 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 2D 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-connections have been widely used to preserve local features for image segmentation [46], [47]. 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 channels to half of the output dimension of the 2×2 convolution. This is because skip-connections applied within image segmentation focus significantly on local details while humans can quickly attend to salient features without a slow process of scrutinising details [48]. Thus in the MIMO-GAN, the skip-connections via separable convolution ensure 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 3D visual 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 transforms the generated saliency maps of projected views so that they are 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). Experimentally, we found that IN outperforms batch normalisation. This finding is in line with many style transfer works [49], [50] which suggested 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.* [24] to output a view-dependent 3D saliency map. The saliency $S_p(V)$ of a 3D vertex p visible in V is computed as

$$S_p(V) = \exp(1 - Z(p)) / \exp(1 - I_i(V)) \tag{1}$$

where $I_i(V)$ denotes the saliency of the pixel i closest to the 2D projection of p in V. $Z(p)$ is the average of the normalised distances between p 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 Saliency contextualisation via CRF

The MIMO-GAN essentially exploits the intra-view knowledge for predicting 3D visual saliency in that its mechanism does not involve any synergy between multiple views of the same 3D object. However, the distribution of human eye fixations [12] indicates that the saliency of the same 3D point in different views are moderately consistent except that the distances between its projections to the view centre vary significantly due to viewpoint change. Such a discrepancy is mainly caused by the centre bias, a heuristic widely used for predicting 2D image saliency. Hence, we propose a Conditional Random Field (CRF) to adjust view-dependent 3D saliency maps by contextualising them with such inter-view information.

The CRF is established on the graph represented by the triangle mesh of a 3D object. Given a 3D vertex p visible in the K views $\mathbf{V} = \{V_1, V_2, \ldots, V_K\}$, we propose a CRF to infer the latent view-dependent 3D visual saliency $\tilde{\mathbf{S}}(\mathbf{V})$ based on the observed one $\mathbf{S}(\mathbf{V}) = \{S(V_1), S(V_2), \ldots, S(V_K)\}$ derived from Eq. (1):

$$\begin{aligned} E(\tilde{\mathbf{S}}(\mathbf{V})|\mathbf{S}(\mathbf{V})) &= \sum_p E(\tilde{\mathbf{S}}_p(\mathbf{V})|\mathbf{S}_p(\mathbf{V})) \\ &= \sum_p U_o(\mathbf{S}_p(\mathbf{V})|\tilde{\mathbf{S}}_p(\mathbf{V})) \\ &\quad + \alpha \sum_p \sum_{q \in \mathcal{N}(p)} U_c(\tilde{\mathbf{S}}_p(\mathbf{V}), \tilde{\mathbf{S}}_q(\mathbf{V})) \\ &\quad + \beta \sum_p U_h(\tilde{\mathbf{S}}_p(\mathbf{V}), \mathbf{V}), \end{aligned} \tag{2}$$

where $\mathcal{N}(p)$ denotes the 1-ring neighbourhood of the 3D vertex p. α and β are the parameters which weight the contributions of the observation term U_o, the compatibility term U_c and the heuristic term U_h to the CRF energy E. We empirically set $\alpha = 0.5$ and $\beta = 300$ in this work. We formulate the observation term as the sum of the squared differences between $\tilde{S}_p(V_k)$ and $S_p(V_k)$ where $k = 1, 2, \ldots, K$:

$$U_o(\mathbf{S}_p(\mathbf{V})|\tilde{\mathbf{S}}_p(\mathbf{V})) = \sum_k (\tilde{S}_p(V_k) - S_p(V_k))^2. \tag{3}$$

We formulate the compatibility term U_c as neighbourhood consistency which encourages adjacent vertices to be assigned with similar saliency values:

$$U_c(\tilde{\mathbf{S}}_p(\mathbf{V}), \tilde{\mathbf{S}}_q(\mathbf{V})) = \sum_k (\tilde{S}_p(V_k) - \tilde{S}_q(V_k))^2, \quad (4)$$

where p and q are neighbouring vertices.

The heuristic term U_h aims to ensure that the saliency values of the same vertex in different views where it is visible are not significantly different unless the distances between its locations and the view centres vary significantly in different views. Therefore, for a vertex p, the corresponding heuristic term U_h should be proportional to the difference between the maximum and the minimum of its view-dependent saliency values over all the views. In addition, U_h should be inversely proportional to the difference between the maximum and the minimum distances from the locations of the vertex in different views to the view centre. The rationale is that if the location of the vertex is close to the view centre in one view but far from it in another, a large inconsistency of its saliency values in different views should be tolerated due in part to the impact of the centre bias. Thus, we formulate U_h as:

$$U_h(\tilde{\mathbf{S}}_p(\mathbf{V}), \mathbf{V}) = \frac{(\max_k(\tilde{S}_p(V_k)) - \min_k(\tilde{S}_p(V_k)))^2}{(\max_k D(p, V_k) - \min_k D(p, V_k))^2 + \epsilon} \quad (5)$$

where ϵ is a constant to stabilise the division and set to 1 in this work. $D(p, V_k)$ is calculated as the distance between the 2D projection of p in the view V_k and the centre of V_k.

The energy calculated in Eq. (2) is actually the negative logarithm of the posterior probability of the CRF. Maximum a posteriori is the most popular principle to infer the CRF in computer vision and graphics, which is equivalent to the minimisation of the energy function. Popular methods for minimising the CRF energy function such as graph cut [51] and belief propagation [52] require that the view-dependent saliency values $\tilde{S}_p(V_k)$ can only have finite discrete states. However, here $\tilde{S}_p(V_k)$ are continuous variables indicating per-vertex 3D saliency values. Note that a CRF with continuous states is in general NP hard [53]. Therefore, to infer the proposed CRF, we develop a practical method based on simulated annealing explicitly shown in Algorithm 1 where the acceptance probability function is defined and implemented through the **if-elseif-else** structure. In the annealing optimisation, we empirically set the maximum number of iterations $MaxIter = 100$ and the parameter $\delta = 0.06$ which sets the upper and lower bounds of the searching space to make a good balance between the performance and the efficiency of the algorithm.

Algorithm 1: CRF inference via simulated annealing

Data: A 3D object containing a set of vertices $P = \{p\}$ and the observed saliency $S_p(V)$ subject to a view V for each vertex
Result: The updated saliency $\tilde{S}_p(V)$ for each vertex
begin
- Initialise $\tilde{S}_p^{(0)}(V)$ as the observed saliency $S_p(V)$;
- Initialise the CRF energy $E^{(0)}$ to a very large value, e.g. 10^4;
- Initialise the temperature $T^{(0)}$ to 1;
- **for** $j \leftarrow 1$ **to** $MaxIter$ **do**
  - **for** $p \in P$ **do**
    - **if** p is invisible in V **then** continue;
    - Propose a new saliency value for p:
      $$\hat{S}_p(V) = \delta r_1 \tilde{S}_p^{(j-1)}(V) + (1 - 0.5\delta)\tilde{S}_p^{(j-1)}(V)$$
      where δ limits the searching space and r_1 is a random value in the interval $[0, 1]$;
    - Compute U_o, U_c and U_h via Eqs. (3)-(5) respectively with the new saliency $\hat{S}_p(V)$ and then the new energy $\hat{E}$ via Eq. (2);
    - **if** $\hat{E} < E^{(j-1)}$ **then**
      - Update $\tilde{S}_p^{(j)}(V) = \hat{S}_p(V)$;
      - Update $E^{(j)} = \hat{E}$;
    - **else if** $r_2 > 1 - T^{(j-1)}$ where r_2 is a random value in the interval $[0, 1]$ **then**
      - Update $\tilde{S}_p^{(j)}(V) = \hat{S}_p(V)$;
      - Update $E^{(j)} = \hat{E}$;
    - **else**
      - No update for $\tilde{S}_p(V)$ and E;
  - Update the temperature T for convergence:
    $$T^{(j)} = \frac{T^{(j-1)} \log(MaxIter)}{\log(MaxIter + j)};$$

3.5 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 the 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, it is trained with four loss terms.

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

$$L_C = -\sum_{c=1}^{C} \mathcal{Q}_c(V) \cdot \log\left(\mathcal{P}_c(V)\right) \quad (6)$$

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 the MIMO-GAN. Here $C = 40$ as we trained the MIMO-GAN with ModelNet40 [27] which collected 3D objects in 40 classes.

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

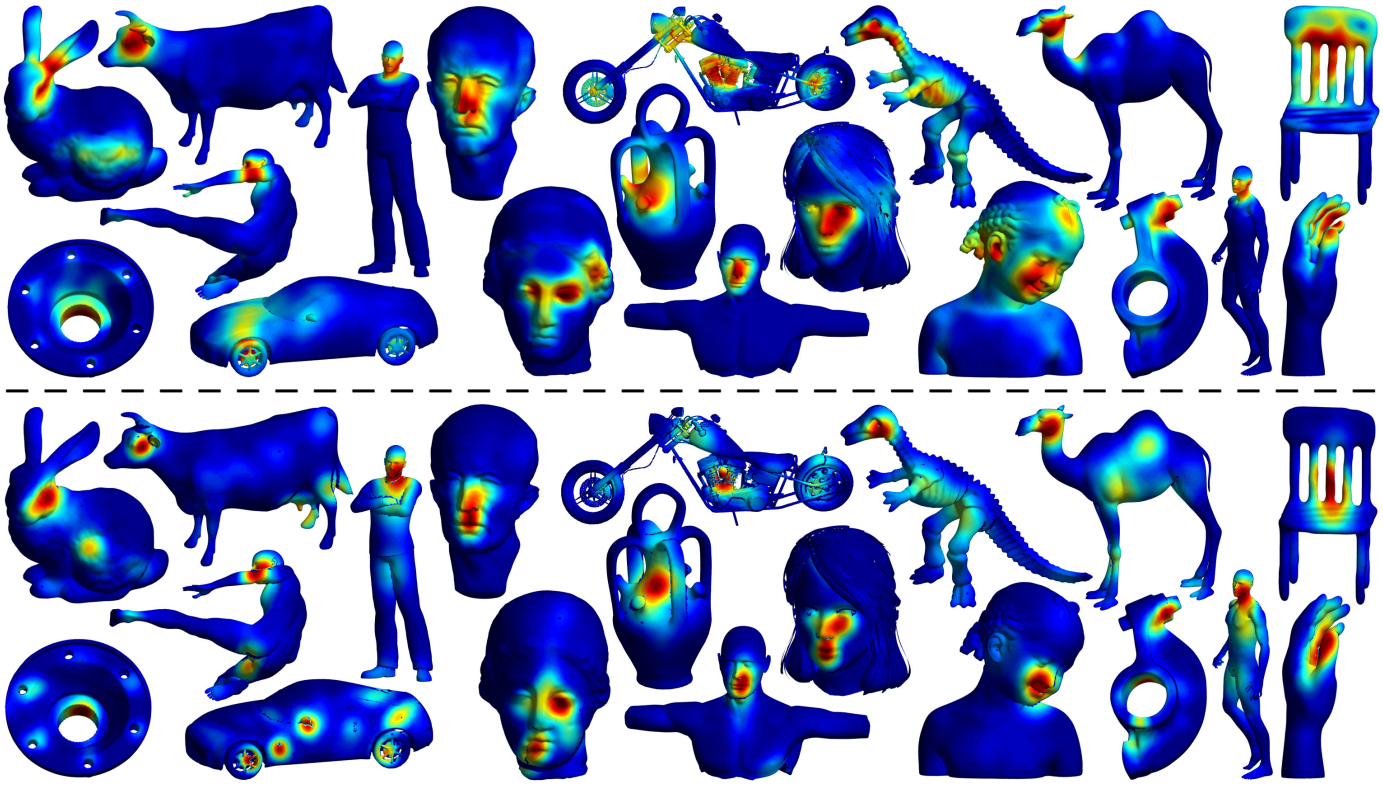


Fig. 3. A gallery of 3D visual saliency detected by our method (top half) with the ground-truth fixation maps (provided by the 3DVA dataset [12]) of the corresponding meshes (bottom half). Warmer colours show higher saliency.

loss:

$$L_S = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{S}(I_i) - G(E(I_i))\right)^2 \tag{7}$$

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

$$\begin{aligned} L_G &= \log(1 - D(G(E(V)))) \quad \text{and} \\ L_D &= -\log(D(G(E(I))) - \log(1-D(G(E(V)))) \end{aligned} \tag{8}$$

where D denotes the discriminator of the MIMO-GAN.

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

$$L_{all} = \lambda_1 L_C + \lambda_2 L_S + \lambda_3 L_G + \lambda_4 L_D \tag{9}$$

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). Next, each 2D saliency map is converted into a view-dependent 3D saliency map by the scheme described in Section 3.3. Finally, the view-dependent 3D saliency maps are further adjusted through the CRF-based saliency contextualisation described in Section 3.4.

Note that our method can also be used to produce view-independent 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 saliency maps, we compute the view-independent 3D visual saliency as the average over the mesh saliency maps across all the views.

4 Experimental Results

All the 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 [27] containing 4,000 meshes from 40 common object categories where all meshes are upright oriented by the method proposed in either [54] or [55]. Unless otherwise specified, the other dataset is the training set of SALICON [15] comprising 10,000 natural scene images with ground-truth saliency annotations. It is worth mentioning that to achieve a thorough evaluation, we also replace the SALICON dataset with other 2D image saliency datasets including CAT2000 [56] and

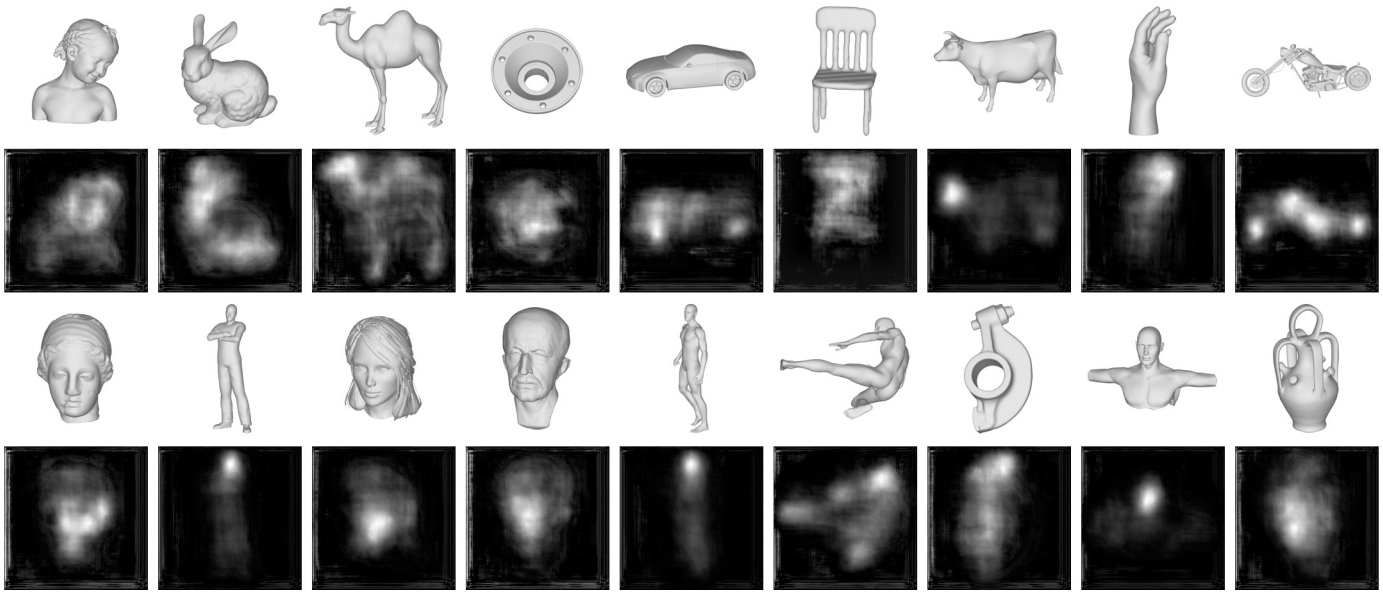


Fig. 4. Projected 2D images and their corresponding 2D saliency maps of the 3D meshes appearing in Fig. 3 except for the 'dinosaur' already shown in Fig. 1. The first and the third rows show the projected 2D images, and the second and the fourth rows show the 2D saliency maps.

FIGRIM [57] for training and evaluate the MIMO-GAN models trained with different image saliency datasets.

We select the 3D visual attention (3DVA) dataset [12] containing 32 meshes for testing. To the best of our knowledge, it is the largest dataset (by the number of 3D objects) for evaluating 3D visual saliency methods with ground-truth fixations on 3D meshes. In the 3DVA dataset, the fixations on each mesh are gathered from three designated viewpoints and are view-dependent (see the rightmost columns in Figs. 1-11 of the supplementary material). It is noteworthy that Wang *et al.* [11] 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 3D visual saliency, we also evaluate it with the Schelling dataset [14] which provides view-independent 3D interest points selected by human subjects for a collection of 400 meshes belonging to 20 object categories. In addition, we conduct a user study using an eye-tracking device to create a new dataset of 3D visual saliency for testing. It is not only larger than the 3DVA dataset but also enables statistical analysis of 3D visual saliency based on the first-hand data it provides.

4.2 Evaluation on the 3DVA dataset

Fig. 3 shows the saliency maps of various 3D objects produced by our method and the corresponding human fixation maps provided by the 3DVA dataset. Fig. 4 shows the view-based 2D saliency maps corresponding to these 3D objects except for the 'dinosaur' which we have shown in Fig. 1. 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. 5 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 analogous saliency distributions. For instance, facial areas of humans and animals are usually detected as salient.

We use linear correlation coefficient (LCC) and area under the ROC curve (AUC) as suggested in [12] to quantitatively measure the similarity between a saliency map produced by a competing method and a ground-truth fixation map. According to [12], to calculate the AUC scores, the ground-truth fixation maps are thresholded 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. 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.

Tables 1 and 2 show the overall performance of a selection of competing methods for 3D visual saliency and our method based on MIMO-GAN with different training sets (see the next paragraph), ablation configurations (see Section 4.5), and multi-view set-ups (see Section 3.1) on the 3DVA dataset in terms of LCC and AUC. 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 the 24-view set-up already achieved high performance and using more views cannot further lead to a significant improvement. Specifically, MIMO-GAN-CRF outperforms the current state-of-the-art method (i.e. CfS-CNN [24]) by 126% and 23% in terms of LCC and AUC, respectively. The quantitative results indicate that 1) 3D visual 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

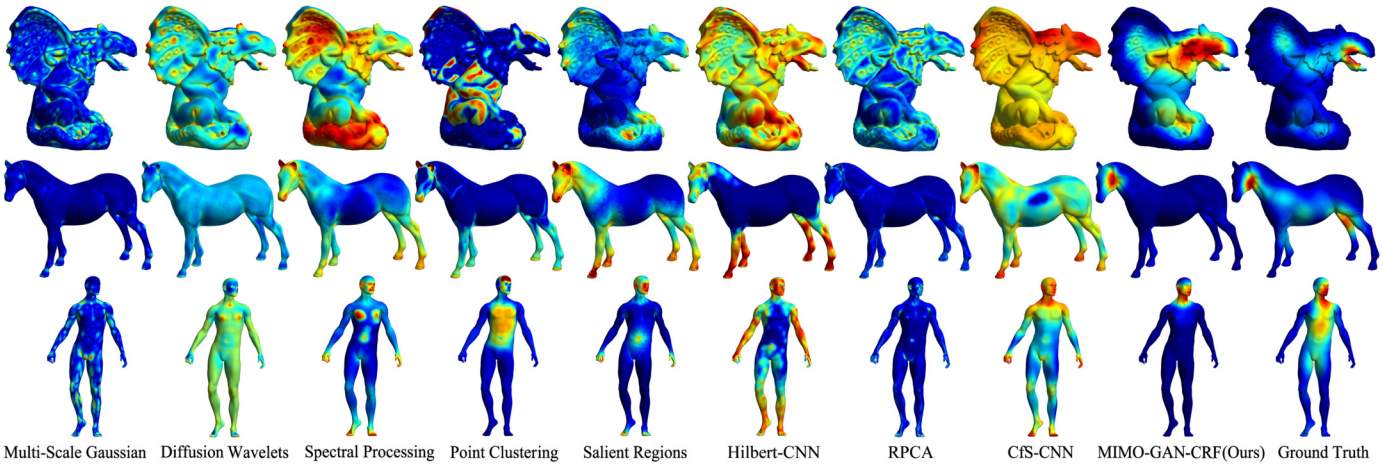


Fig. 5. Comparisons of 3D visual saliency detected by different methods. From left to right: Multi-Scale Gaussian [3], Diffusion Wavelets [58], Spectral Processing [8], Point Clustering [13], Salient Regions [9], Hilbert-CNN [40], RPCA [37], CfS-CNN [24], the proposed MIMO-GAN-CRF and the ground-truth fixation maps provided by the 3DVA dataset [12]. Comparative results of more objects are available in the supplementary material.

TABLE 1
Performance of 3D visual saliency methods on the 3DVA dataset [12] in terms of the mean and the standard deviation of LCC.

Method	mean LCC ↑	Std. Dev. of LCC ↓
Multi-Scale Gaussian [3]	0.131	0.265
Diffusion Wavelets [58]	0.088	0.222
Spectral Processing [8]	0.078	0.253
Point Clustering [13]	0.132	0.300
Salient Regions [9]	0.215	0.245
Hilbert-CNN [40]	0.113	0.267
RPCA [37]	0.199	0.251
CfS-CNN [24]	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
MIMO-GAN-CRF-T1 w/ 24 views	0.412	0.243
MIMO-GAN-CRF-T2 w/ 24 views	0.316	0.244
MIMO-GAN-CRF w/ 24 views	**0.510**	0.203

TABLE 2
Performance of 3D visual saliency methods on the 3DVA dataset [12] in terms of the mean and the standard deviation of AUC.

Method	mean AUC ↑	Std. Dev. of AUC ↓
Multi-Scale Gaussian [3]	0.593	0.170
Diffusion Wavelets [58]	0.558	0.143
Spectral Processing [8]	0.553	0.154
Point Clustering [13]	0.583	0.183
Salient Regions [9]	0.628	0.149
Hilbert-CNN [40]	0.573	0.176
RPCA [37]	0.622	0.154
CfS-CNN [24]	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
MIMO-GAN-CRF-T1 w/ 24 views	0.736	0.140
MIMO-GAN-CRF-T2 w/ 24 views	0.689	0.143
MIMO-GAN-CRF w/ 24 views	**0.790**	**0.108**

of knowledge via a GAN framework for detecting 3D visual saliency is computationally effective.

In the above evaluations, the proposed MIMO-GAN is trained with the SALICON dataset. To further explore the impact of 2D image saliency on the 3D visual saliency produced by MIMO-GAN, we train it with other datasets of 2D image saliency and evaluate the performance on the 3DVA dataset again. We first replace SALICON with CAT2000 [56], a popular dataset which provides ground-truth human fixations of 2,000 images from 24 observers for training. We also use the FIGRIM dataset [57] which provides eye fixation data of 2,157 images for training, where each image is observed by 15 subjects on average. In Tables 1 and 2, the MIMO-GAN models trained with CAT2000 and FIGRIM are denoted as 'MIMO-GAN-CRF-T1 w/ 24 views' and 'MIMO-GAN-CRF-T2 w/ 24 views', respectively. It can be seen that the MIMO-GAN models trained with CAT2000 and FIGRIM are outperformed by the one trained with SALICON. This is expected as these two datasets are much smaller than SALICON which contains 10,000 training images. Even so, we can observe that the models trained with CAT2000 and FIGRIM perform significantly better than other competing methods. Such results further demonstrate the idea of leveraging 2D image saliency to learn 3D visual saliency.

TABLE 3
Evaluation of the robustness of our method against different levels of Gaussian noise.

Gaussian noise	MIMO-GAN		MIMO-GAN-CRF	
	LCC ↑	AUC ↑	LCC ↑	AUC ↑
no noise	0.489	0.780	0.510	0.790
$\sigma = 0.001B$	0.480	0.768	0.488	0.776
$\sigma = 0.002B$	0.472	0.768	0.486	0.775
$\sigma = 0.004B$	0.457	0.761	0.481	0.773

In addition, we have conducted tests by adding Gaussian noise with $\sigma = 0.001B$, $0.002B$ and $0.004B$ respectively to all the 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 (with and without the CRF-based saliency contextualisation), which demonstrates its robustness against

TABLE 4
Performance of 3D visual saliency methods on the Schelling dataset [14] 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.

Method	$\sigma = 0.1B$	$\sigma = 0.12B$	$\sigma = 0.14B$	$\sigma = 0.16B$	$\sigma = 0.18B$	$\sigma = 0.2B$
Multi-Scale Gaussian [3]	0.223	0.213	0.202	0.193	0.186	0.179
Diffusion Wavelets [58]	0.101	0.091	0.082	0.074	0.068	0.063
Spectral Processing [8]	0.324	0.322	0.313	0.301	0.293	0.284
Salient Regions [9]	0.437	0.421	0.402	0.376	0.360	0.340
CfS-CNN [24]	0.455	0.457	0.454	0.447	0.439	0.427
Hilbert-CNN [40]	0.127	0.129	0.127	0.124	0.124	0.123
RPCA [37]	0.336	0.323	0.309	0.295	0.286	0.277
MIMO-GAN w/ 24 views	0.447	0.462	0.470	0.472	0.470	0.463
MIMO-GAN-CRF w/ 24 views	**0.476**	**0.498**	**0.510**	**0.515**	**0.516**	**0.510**

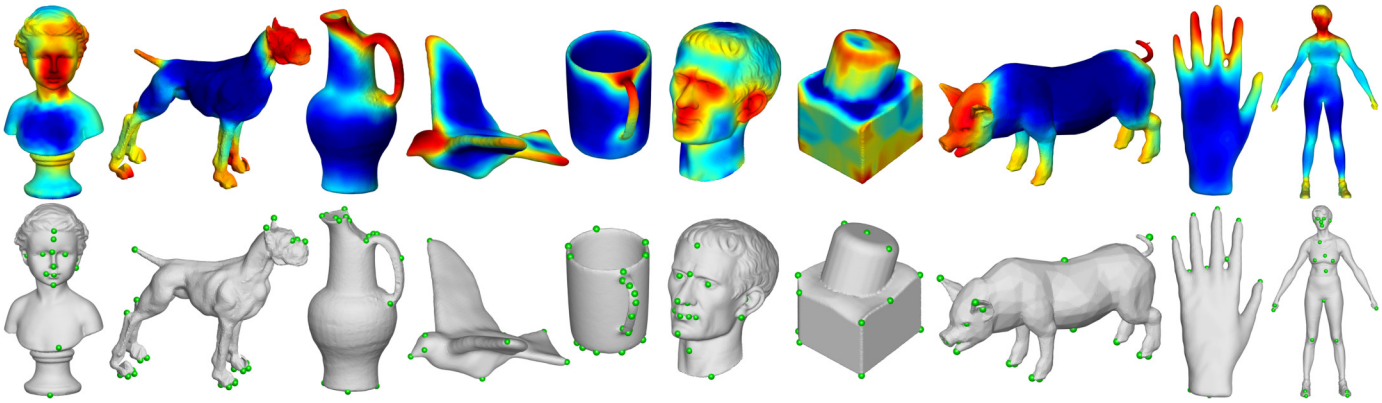


Fig. 6. View-independent 3D visual saliency detected by our method and the human-picked interest points (Schelling points [14]).

different levels of noise.

4.3 Evaluation on the Schelling dataset

Apart from human eye fixations, human-picked 3D interest points have also been used for evaluating 3D visual saliency methods [24], [59]. The Schelling dataset [14] 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 [15], [25]: we project a Gaussian distribution on a mesh where each vertex is labelled by either 1 (interest point) or 0 (non-interest point) and vary the standard deviation to generate different versions of ground-truth saliency maps. When we evaluate our method on such pseudo ground truth, we essentially estimate whether it can detect saliency at different scales.

Note that as demonstrated in [12], 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 3D visual 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-CRF for predicting view-independent 3D visual saliency is still the top performing method. In particular, we find that only MIMO-GAN and MIMO-GAN-CRF perform better with $\sigma = \{0.18B, 0.2B\}$ than $\sigma = \{0.1B, 0.12B\}$, which shows that our method is effective at detecting saliency at relatively large scales. This finding is consistent with the qualitative results shown in Figs. 3 and 5 where our method often highlights one or two large areas. We also provide quantitative evaluation per object category in the supplementary material.

Interestingly, Fig. 6 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 3D visual saliency as the average over the saliency maps across all the views as mentioned at the end of Section 3.5. 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 prediction for objects with many highly occluded regions.

4.4 Evaluation via a user study

On the one hand, the 3DVA dataset [12] is of a small scale while it directly records human eye fixations. On the other hand, the Schelling dataset [14] is much larger while it is not specifically designed for saliency evaluation and does not directly provide eye fixation data. Thus, to enable a thorough evaluation in a more orthodox manner, we conducted a user study to collect first-hand human eye fixation data at a significantly larger scale. We largely adopted the pipeline of constructing the 3DVA dataset to conduct the user study. In detail, we selected 3 viewpoints for each 3D mesh and generated 3 rendered 2D views for it. We developed a GUI, shown in Fig. 7, to exhibit the 2D views to a total of 16 subjects composed of 8 females and 8 males

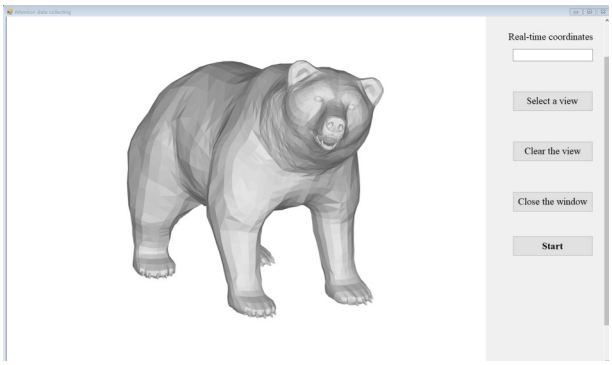


Fig. 7. The GUI for collecting human fixation data from 3D meshes.

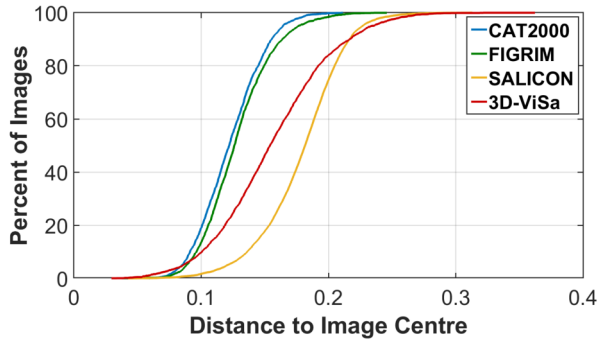


Fig. 8. Cumulative distributions of mean distances from the fixations to the image centre. Distances are normalised with the length of image diagonal.

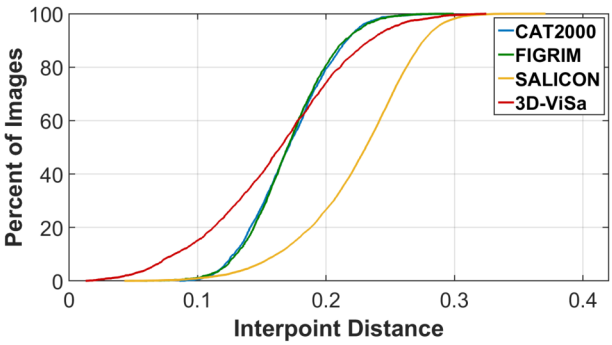


Fig. 9. Cumulative distributions of mean interpoint distances between the fixations. Distances are normalised with the length of image diagonal.

in the user study where we used an eye tracker to collect their eye fixations. The eye tracker fixed at the bottom edge of a laptop monitor was set to record human eye fixations within the initial glance of 6 seconds. During the process of data collection, a subject sat in front of the monitor, and the distance between the eyes and the monitor is about 45cm. Then, we employed the postprocessing schemes suggested in [12] to generate the ground-truth saliency maps based on the collected raw eye fixation data. We name our dataset as 3D-ViSa[3], short for 3D Visual Saliency. The 3D-ViSa dataset includes 540 view-dependent saliency maps for 180 meshes while in contrast, the 3DVA dataset contains only 96 view-dependent saliency maps for 32 meshes. Note that the 3DVA dataset does not classify the 3D meshes. In this work, we attempt to explore the per-category performance of the proposed method. Therefore, to build our dataset, we selected 6 meshes from each of the 30 categories including 'airplane', 'bear', 'bed', 'bench', 'bike', etc.

Centre bias is a spatial prior that regions near the image centre tend to attract more fixations, which exists in almost all eye-tracking datasets for 2D images. In our user study, we recorded the fixations of each subject to validate through statistical analysis whether and to what degree the centre bias prior exists in 3D visual saliency. The cumulative distribution of the mean distances from the fixations to the image centre is shown in Fig. 8 where we normalised the distance to the image centre by the length of the image diagonal. It can be seen that similar to 2D image saliency, 3D visual saliency is also subject to centre bias. However, the distribution corresponding to the 3D-ViSa dataset is significantly different from those corresponding to other 2D image saliency datasets: it has both more fixations (with distance smaller than around 0.08) close to image centre and more fixations (with distance larger than around 0.22) distant from it.

Fig. 9 shows the cumulative distribution of the mean interpoints distance between the fixations where the distance is also normalised by the length of the image diagonal. It can be seen that the 3D-ViSa dataset has significantly more fixations close to each other than the 2D eye-tracking datasets. This is because a large area in a projected 2D image of a 3D mesh is filled with the white background and thus the fixations only locate within the small foreground area in the middle of the image. By observing both Fig. 8 and Fig. 9, we can draw a conclusion that compared to 2D image saliency, 3D visual saliency tends to be less affected by the centre bias.

Fig. 10 shows the saliency maps of various 3D objects produced by our method and the corresponding human fixation maps provided by the 3D-ViSa dataset. We can observe that these saliency maps are visually consistent with the ground-truth human fixations. Tables 5 and 6 list the quantitative performance of the competing methods for 3D visual saliency on the newly created 3D-ViSa dataset. Similar to the evaluation implemented on the 3DVA dataset, we compute the LCC and AUC scores between the saliency maps produced by each competing method and the ground-truth fixation maps provided by the 3D-ViSa dataset. It is noteworthy that many 3D meshes in the 3D-ViSa dataset contain reconstruction noise such as holes, self-intersecting faces, non-manifold edges, T-vertices, etc. Consequently, we found that Point Clustering [13] and Salient Regions [9] failed to generate saliency maps from a number of 3D meshes in the 3D-ViSa dataset. Therefore, they are excluded in the tables. The results shown in Tables 5 and 6 further demonstrate the superiority of the proposed method based on MIMO-GAN. In particular, we can see that it performs slightly better on 3D-ViSa than on 3DVA. Presumably, this is because 3D-ViSa does not include any object categories unseen in SALICON while 3DVA contains some 3D objects categorically unseen in SALICON, such as 'dragon', 'gargoyle', 'octopus', 'protein', etc.

3. The data and code of 3D-ViSa are publicly available at https://github.com/rsong/3D-ViSa.

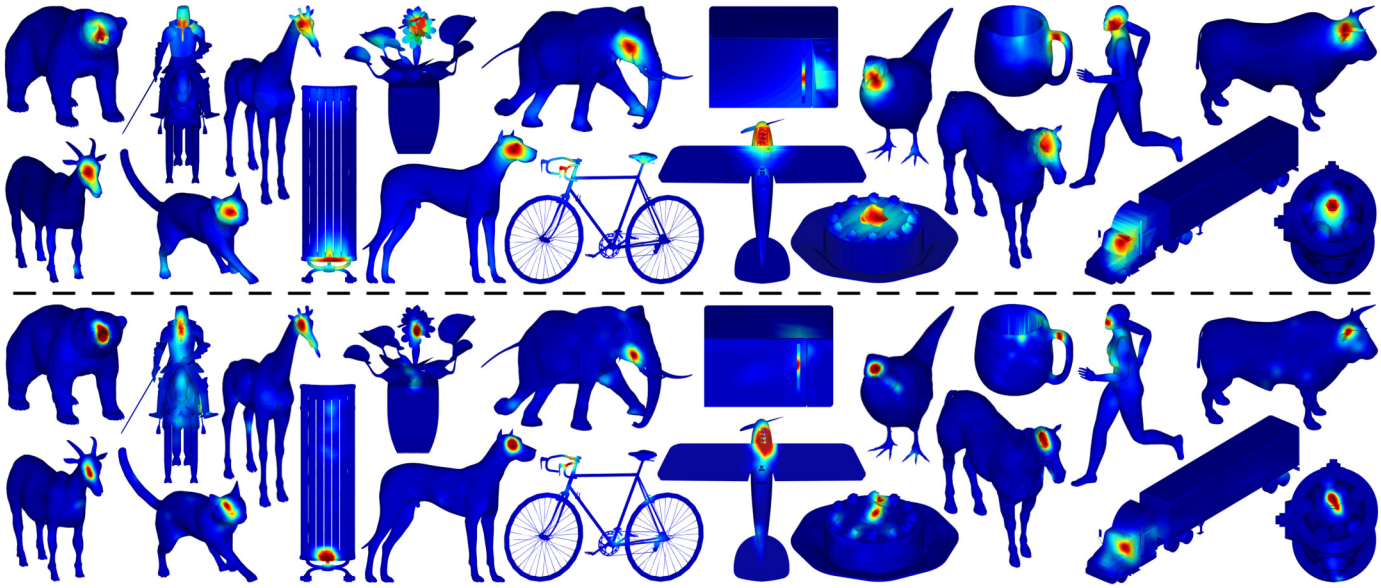


Fig. 10. A gallery of 3D visual saliency detected by our method (top half) with the ground-truth fixation maps (provided by our 3D-ViSa dataset) of the corresponding meshes (bottom half).

Differing from the 3DVA dataset, the 3D-ViSa dataset explicitly classifies the 3D objects, which facilitates the investigation of the categorical difference of the performance. Figs. 11 and 12 show the performance per category of our method in terms of LCC and AUC. It can be observed that the LCC and the AUC plots are highly consistent with each other. For example, both LCC and AUC scores indicate that 'bench', 'keyboard', and 'monitor' are the most challenging categories for our method. A further study shows that the ground-truth fixations of the 3D objects of the 3 categories made by different subjects are highly inconsistent and dispersed, which suggests that such objects hardly contain any stable salient features.

4.5 Ablation studies

In this section, we evaluate different configurations of MIMO-GAN with the 24-view set-up to understand whether and to what degree 3D visual saliency for a single object is a derivative of 2D image saliency. We thus conduct three ablation studies on the 3DVA dataset:

(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 MIMO-GANs suffer from a degraded performance compared to its full version. Among them, MIMO-GAN-A2 is the worst affected one although it still outperforms most of the competing methods. In comparison, MIMO-GAN-A1 performs significantly better than MIMO-GAN-A2, which indicates that image saliency has a much greater impact than object categorical information on 3D visual saliency. Particularly, we can see that MIMO-GAN-A1 which essentially learns 3D visual saliency from image saliency already outperforms all competing methods. This suggests that 3D visual saliency for predicting human visual attention on a 3D object depends heavily on image saliency which predicts where human observers look in natural scene images. However, the considerable superiority of the full version of MIMO-GAN over MIMO-GAN-A1 as shown in Tables 1 and 2 demonstrates that categorical information of 3D objects also brings in a significant performance gain for 3D visual saliency of a single object on top of image saliency. One explanation is that the human perception system tends to capture the most informative features as salient [60] 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.

4.6 Evaluation on 3D scenes

All of the above evaluations are subject to a single object. In the following, we conduct evaluations on 3D scenes. The visual saliency of a 3D scene aims to create a per-point saliency map that indicates the perceptual importance of each 3D point in the scene, but it is not the simple combination of the saliency of all objects appearing in the scene. This is because the saliency of an object in a scene is a relative concept compared with the others in the context of the scene. In other words, the 3D visual saliency of a scene depends on not only the objects it contains, but also the way they coexist [24]. Consequently, existing saliency methods only concerning a single 3D object [8], [9], [11], [59] cannot cope with a scene in that they are unable to capture the global spatial and semantic relationship between the objects not connected by mesh edges.

TABLE 5
Performance of 3D visual saliency methods on our 3D-ViSa dataset in terms of the mean and the standard deviation of LCC.

Method	mean LCC↑	Std. Dev. of LCC↓
Multi-Scale Gaussian [3]	0.055	0.244
Diffusion Wavelets [58]	0.003	0.210
Spectral Processing [8]	0.126	0.263
Hilbert-CNN [40]	0.084	0.220
RPCA [37]	0.089	0.232
CfS-CNN [24]	0.176	0.223
MIMO-GAN w/ 24 views	0.509	**0.201**
MIMO-GAN-CRF w/ 24 views	**0.526**	0.215

TABLE 6
Performance of 3D visual saliency methods on our 3D-ViSa dataset in terms of the mean and the standard deviation of AUC.

Method	mean AUC↑	Std. Dev. of AUC↓
Multi-Scale Gaussian [3]	0.446	0.238
Diffusion Wavelets [58]	0.512	0.147
Spectral Processing [8]	0.581	0.176
Hilbert-CNN [40]	0.536	0.180
RPCA [37]	0.556	0.168
CfS-CNN [24]	0.611	0.159
MIMO-GAN w/ 24 views	0.818	0.133
MIMO-GAN-CRF w/ 24 views	**0.823**	**0.132**

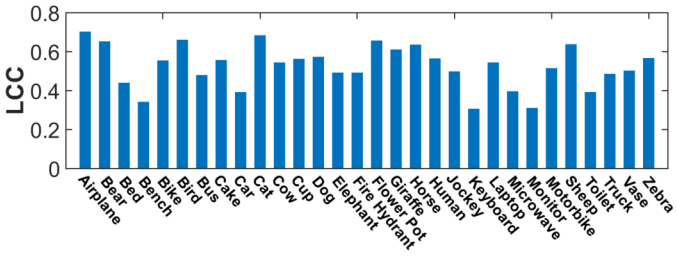


Fig. 11. Per-class performance of our method on the 3D-ViSa dataset in terms of the mean LCC.

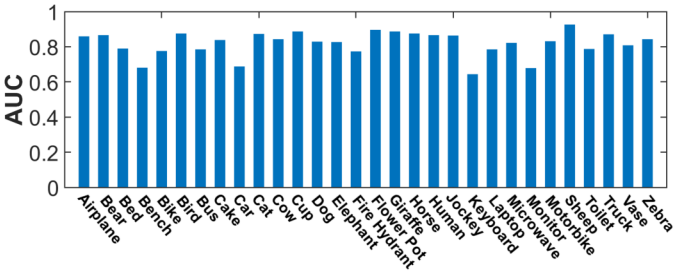


Fig. 12. Per-class performance of our method on the 3D-ViSa dataset in terms of the mean AUC.

In sharp contrast, our method can be directly used to predict the visual saliency of a scene that contains multiple 3D objects and is represented by either a mesh or a depth map, which significantly expands the range of its applications. Specifically, if a scene is represented as a mesh, the pipeline illustrated in Fig. 1 for predicting 3D visual saliency of a single object can be applied without any change. This is because our pipeline is based on a multi-view set-up where the information related to the global positional relationship of multiple objects is encoded in multiple projected 2D views of the scene. If a scene is represented as a depth map, MIMO-GAN will directly take it as input and outputs a saliency map, where the depth map is processed as a 2D intensity image. To make a solid evaluation for our method and demonstrate its wide applicability, we conduct qualitative and quantitative experiments on various 3D scenes represented by either meshes or depth maps.

Table 7 shows the comparative results on the NUS3D-Saliency dataset [61] which provides 600 depth maps of various scenes with the corresponding eye fixation ground truth and might be the largest 3D eye-tracking dataset according to [62]. It is noteworthy that all the competing methods listed in Table 7, including DSM (Depth Saliency Mapping) [17], the proposed MIMO-GAN and its variants rely only on the 3D stimuli, i.e. the depth information, for saliency estimation. All the 600 depth maps in the NUS3D-Saliency dataset were used for testing where the MIMO-GAN and its variants were pre-trained on ImageNet and SALICON without any further fine-tuning. This benefits by avoiding potential overfitting and more importantly, allows us to explore the effect of 2D image saliency on 3D scene saliency as the training phase is not intertwined with the 3D eye-tracking data. Since a scene is represented as a depth map, its corresponding saliency map is only valid for a particular view. Thus the CRF-based saliency contextualisation is not

TABLE 7
Evaluation of 3D scene saliency on the NUS3D-Saliency dataset [61] in terms of the mean and the standard deviation of LCC and AUC.

Method	mean LCC↑	Std. Dev. of LCC↓	mean AUC↑	Std. Dev. of AUC↓
DSM [17]	0.222	0.130	0.726	0.125
MIMO-GAN-A1	**0.290**	0.128	**0.781**	**0.103**
MIMO-GAN-A2	0.057	**0.115**	0.584	0.109
MIMO-GAN-A3	0.259	0.140	0.753	0.123
MIMO-GAN	0.267	0.132	0.761	0.116

needed. Based on the observations from Table 7 where the naming of the ablated versions of MIMO-GAN is exactly the same as Section 4.5, we have the following findings:

- Despite a performance drop compared to the saliency prediction of a single 3D object (see Tables 1 and 2), MIMO-GAN and its variants except MIMO-GAN-A2 are effective for predicting 3D scene saliency. Such a drop is expected, partly because depicting a 3D scene via a single depth map is usually ambiguous due to occlusion, while by contrast a 3D mesh represents a 3D object more comprehensively and precisely.
- There still exists a gap between 2D image saliency and 3D visual saliency for a scene. And that the full version of MIMO-GAN consistently outperforms MIMO-GAN-A3 demonstrates that the proposed GAN architecture is reliable for minimising such a gap in the prediction of 3D visual saliency for both a single object and a scene.
- The knowledge vital for recognising a single object is barely useful and probably distracting for scene saliency. We can see that MIMO-GAN-A2 has a poor performance and MIMO-GAN-A1 outperforms MIMO-GAN. This indicates that the classification path which imposes a knowledge transfer from 3D object classification to 3D visual saliency for a scene actually hurts the performance on the latter task, a phenomenon referred

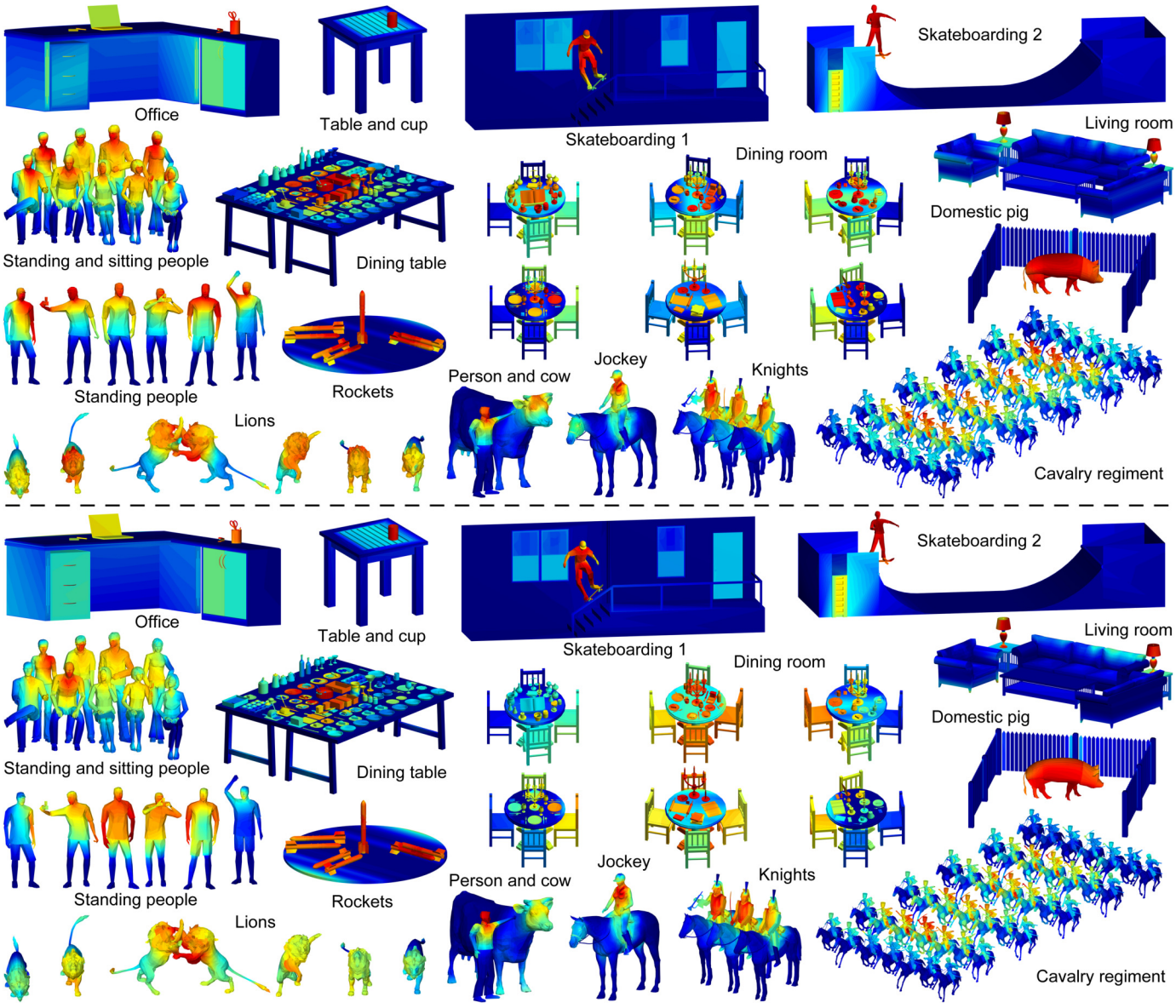


Fig. 13. 3D scene saliency produced by MIMO-GAN-CRF (top half) and MIMO-GAN-CRF-A1 (bottom half) which removes the classification path.

to as *negative transfer* in transfer learning. In sharp contrast, the information flow between the two tasks of 3D object classification and 3D visual saliency for a single object, enabled by exactly the same subnetwork, yields *positive transfer* as discussed in Section 4.5.

Therefore, we next apply MIMO-GAN-CRF and its ablated version MIMO-GAN-CRF-A1 respectively to 3D visual saliency for scenes represented as meshes and show the results in Fig. 13. We can see that although the classification path is removed in MIMO-GAN-CRF-A1, its results for most scenes are consistent with those of the full version. In line with our finding based on the quantitative results in Table 7, this qualitatively indicates that the categorical information of a single 3D object seems not important for 3D visual saliency of most scenes, particularly those containing objects that belong to different categories. Instead, we found that the coexisting relationship between multiple objects probably matters. For example, compared to the 'horse' in Fig. 5, in the scenes 'jockey', 'knights' and 'cavalry regiment', the saliency of the horses is consistently suppressed due to the coexistence of the humans. Be that as it may, we observe some small differences despite the consistent centre bias in several scenes, including 'standing and sitting people', 'standing people', 'lions', 'person and cow' and 'dining room'. In these scenes, due to the involvement of object categorical information, MIMO-GAN-CRF highlights some local areas important for object recognition, such as the facial areas of the people and animals and the tableware on the tables, at the left and right ends of the scenes, even if they are distant from scene centres. In comparison, MIMO-GAN-CRF-A1 concentrates more on global semantics, e.g. a more prominent centre bias that largely suppresses the above local areas in these scenes.

Fig. 14 shows the view-based 2D saliency maps corresponding to all of the 24 views of the 'skateboarding 1' scene appearing in Fig. 13. It can be observed that the person in the scene always attracts human fixations as long as he is not occluded. In the views where the person is not visible

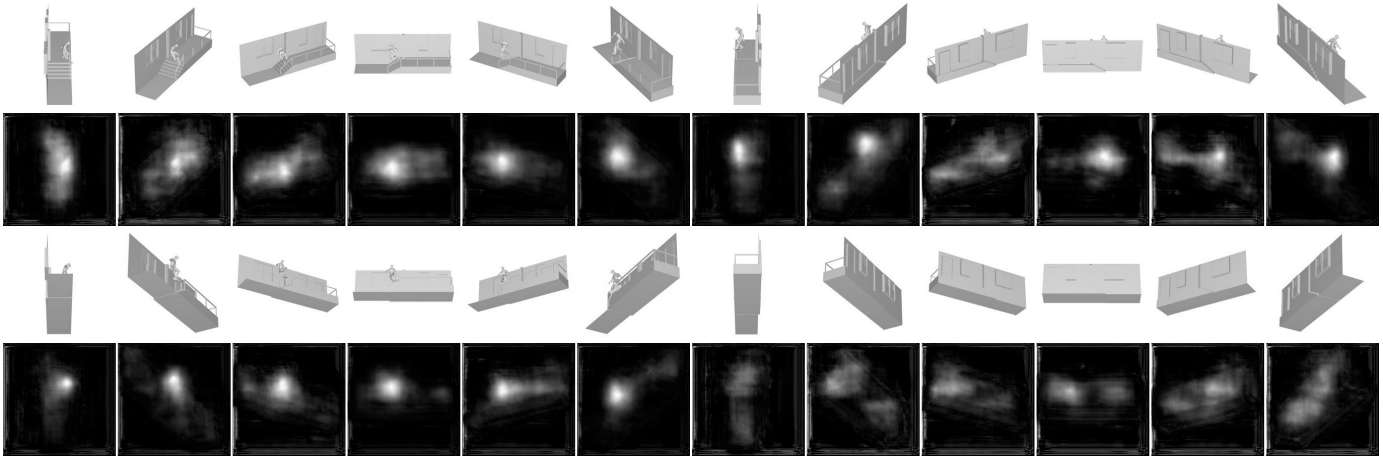


Fig. 14. Projected 2D images for all 24 views and their corresponding 2D saliency maps of the 'skateboarding 1' scene appearing in Fig. 13. The first and the third rows show the projected 2D images, and the second and the fourth rows show the 2D saliency maps.

(e.g. the last 6 views in the third row of Fig. 14), there is no area of high saliency for representing the theme of the scene about "skateboarding".

Overall, the proposed method based on MIMO-GAN can be directly used for 3D scene saliency and produces reasonable results. However, as a limitation of our method, its accuracy of predicting saliency for a 3D scene remains worse than that for a single 3D object in most cases. On the one hand, a scene containing multiple 3D objects is generally more complex than a single object and thus human fixations on a 3D scene are more likely to have inconsistent distributions. On the other hand, the categorical information of a single 3D object cannot benefit significantly the prediction of 3D scene saliency. Therefore, other types of weak supervision need to be introduced into MIMO-GAN to improve its performance on 3D scenes.

5 Conclusions

Is 3D visual saliency an independent perceptual measure or a derivative of 2D image saliency? Our answer is that although the prediction of 3D visual saliency for both a single object and a scene benefits substantially from image saliency, it cannot be regarded as a derivative of image saliency for two reasons. First, there exists a gap between 2D image saliency and 3D visual saliency for both cases although the transfer learning compelled by the GAN loss of our MIMO-GAN can effectively alleviate it. Second, we quantitatively demonstrate that 3D visual saliency for a single object is also influenced by other factors such as object categorical information which provides useful knowledge largely independent of image saliency. We nevertheless demonstrate that 3D visual saliency for most scenes containing various objects does not associate much with object categorical information but might be related to the way that the multiple objects in the scene coexist. Therefore, one future work is to investigate the factors that can improve 3D visual saliency for scenes on top of image saliency and establish a computational pipeline based on them.

Furthermore, since MIMO-GAN is trained with publicly available datasets in a weakly supervised manner with no requirement for the costly collection of 3D eye-tracking data, it is potentially of broad interest in the community. Thus another future work is to adapt the proposed approach to other tasks of 3D object and scene understanding. Specifically, the classification path of MIMO-GAN might have to be adapted for another type of weak supervision depending on the particular task of 3D scene understanding.

References

[1] X. Liu, L. Liu, W. Song, Y. Liu, and L. Ma, "Shape context based mesh saliency detection and its applications: A survey," *Comput. Graph.*, vol. 57, pp. 12–30, 2016. 1

[2] W. Wang, H. Chao, J. Tong, Z. Yang, X. Tong, H. Li, X. Liu, and L. Liu, "Saliency-preserving slicing optimization for effective 3d printing," *Comput. Graph. Forum*, vol. 34, no. 6, 2015. 1

[3] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 659–666, 2005. 1, 3, 9, 10, 13

[4] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1633–1642, 2018. 1

[5] P. Shilane and T. Funkhouser, "Distinctive regions of 3d surfaces," *ACM Trans. Graph.*, vol. 26, no. 2, p. 7, 2007. 1, 3

[6] R. Song, Y. Liu, R. R. Martin, and P. L. Rosin, "Saliency-guided integration of multiple scans," in *Proc. CVPR*, 2012, pp. 1474–1481. 1, 3

[7] J. Wu, X. Shen, W. Zhu, and L. Liu, "Mesh saliency with global rarity," *Graph. Models*, vol. 46, pp. 264–274, 2013. 1, 3

[8] R. Song, Y. Liu, R. R. Martin, and P. L. Rosin, "Mesh saliency via spectral processing," *ACM Trans. on Graph.*, vol. 33, no. 1, pp. 1–17, 2014. 1, 3, 9, 10, 12, 13

[9] G. Leifman, E. Shtrom, and A. Tal, "Surface regions of interest for viewpoint selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2544–2556, 2016. 1, 3, 9, 10, 11, 12

[10] X. Wang, D. Lindlbauer, C. Lessig, M. Maertens, and M. Alexa, "Measuring the visual salience of 3d printed objects," *IEEE Comput. Graph. Appl.*, vol. 36, no. 4, pp. 46–55, 2016. 1, 2

[11] X. Wang, S. Koch, K. Holmqvist, and M. Alexa, "Tracking the gaze on objects in 3d: how do people really look at the bunny?" *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–18, 2018. 1, 2, 3, 8, 12

[12] G. Lavoué, F. Cordier, H. Seo, and M.-C. Larabi, "Visual attention for rendered 3d shapes," *Comput. Graph. Forum (Proc. Eurographics)*, pp. 414–421, 2018. 1, 2, 5, 7, 8, 9, 10, 11

[13] F. Ponjou Tasse, J. Kosinka, and N. Dodgson, "Cluster-based point set saliency," in *Proc. ICCV*, 2015, pp. 163–171. 1, 9, 11

[14] X. Chen, A. Saparov, B. Pang, and T. Funkhouser, "Schelling points on 3d surface meshes," *ACM Trans. Graph.*, vol. 31, no. 4, p. 29, 2012. 1, 2, 3, 8, 10

[15] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proc. CVPR*, 2015, pp. 1072–1080. 1, 2, 7, 10

[16] E. Potapova, M. Zillich, and M. Vincze, "Learning what matters: combining probabilistic models of 2d and 3d saliency cues," in *Proc. ICVS*, 2011, pp. 132–142. 1

[17] J. Wang, M. P. Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3d visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, 2013. 1, 13

[18] J. Morales, A. Bax, and C. Firestone, "Sustained representation of perspectival shape," *Proc. Natl. Acad. Sci.*, vol. 117, no. 26, pp. 14 873–14 882, 2020. 1

[19] V. Nejati, "Effect of stimulus dimension on perception and cognition," *Acta Psychol.*, vol. 212, p. 103208, 2021. 1

[20] L. Jansen, S. Onat, and P. König, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *J. Vis.*, vol. 9, no. 1, pp. 29–29, 2009. 1

[21] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, no. 1, pp. 243–271, 1999. 2

[22] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 5–5, 2011. 2

[23] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends Cogn. Sci.*, vol. 17, no. 11, pp. 585–593, 2013. 2

[24] R. Song, Y. Liu, and P. L. Rosin, "Mesh saliency via weakly supervised classification-for-saliency CNN," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 1, pp. 151–164, 2021. 2, 3, 5, 8, 9, 10, 12, 13

[25] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2018. 2, 10

[26] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, 2013, pp. 3166–3173. 2

[27] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D shapenets: A deep representation for volumetric shapes," in *Proc. CVPR*, 2015, pp. 1912–1920. 2, 6, 7

[28] M. Savva, A. X. Chang, and P. Hanrahan, "Semantically-enriched 3D models for common-sense knowledge," in *Proc. CVPR Workshops*, 2015, pp. 24–31. 2

[29] Y. Kim, A. Varshney, D. Jacobs, and F. Guimbretière, "Mesh saliency and human eye fixations," *ACM Trans. Appl. Percept.*, vol. 7, no. 2, pp. 12:1–12:13, 2010. 2, 3

[30] R. Song, W. Zhang, Y. Zhao, Y. Liu, and P. L. Rosin, "Mesh saliency: An independent perceptual measure or a derivative of image saliency?" in *Proc. CVPR*, 2021, pp. 8853–8862. 2

[31] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, 2018. 3

[32] R. Gal and D. Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Trans. Graph.*, vol. 25, no. 1, pp. 130–150, 2006. 3

[33] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994. 3

[34] C. Koch and T. Poggio, "Predicting the visual world: silence is golden," *Nat. Neurosci.*, vol. 2, pp. 9–10, 1999. 3

[35] E. Shtrom, G. Leifman, and A. Tal, "Saliency detection in large point sets," in *Proc. ICCV*, 2013, pp. 3591–3598. 3

[36] S. Wang, N. Li, S. Li, Z. Luo, Z. Su, and H. Qin, "Multi-scale mesh saliency based on low-rank and sparse analysis in shape feature space," *Comput. Aided Geom. Des.*, vol. 35, pp. 206–214, 2015. 3

[37] G. Arvanitis, A. S. Lalos, and K. Moustakas, "Robust and fast 3-d saliency mapping for industrial modeling applications," *IEEE Trans. Industr. Inform.*, vol. 17, no. 2, pp. 1307–1317, 2021. 3, 9, 10, 13

[38] M. Lau, K. Dev, W. Shi, J. Dorsey, and H. Rushmeier, "Tactile mesh saliency," *ACM Trans. Graph.*, vol. 35, no. 4, 2016. 3

[39] X. Li, L. Yu, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, "Unsupervised detection of distinctive regions on 3d shapes," *ACM Trans. Graph.*, vol. 39, no. 5, pp. 1–14, 2020. 3

[40] S. Nousias, G. Arvanitis, A. S. Lalos, and K. Moustakas, "Mesh saliency detection using convolutional neural networks," in *Proc. ICME*, 2020, pp. 1–6. 3, 9, 10, 13

[41] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proc. ICCV*, 2015, pp. 945–953. 3

[42] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas, "Volumetric and multi-view CNNs for object classification on 3d data," in *Proc. CVPR*, 2016, pp. 5648–5656. 3

[43] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proc. CVPR*, vol. 1, no. 2, 2017, p. 8. 3

[44] H. Huang, E. Kalogerakis, S. Chaudhuri, D. Ceylan, V. G. Kim, and E. Yumer, "Learning local shape descriptors from part correspondences with multiview convolutional networks," *ACM Trans. Graph.*, vol. 37, no. 1, p. 6, 2018. 3

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015. 5

[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, June 2015. 5

[47] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019. 5

[48] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognitive psychology*, vol. 43, no. 3, pp. 171–216, 2001. 5

[49] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. CVPR*, 2017, pp. 6924–6932. 5

[50] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1501–1510. 5

[51] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, 2004. 6

[52] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vision*, vol. 70, no. 1, pp. 41–54, 2006. 6

[53] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous markov random fields for robust stereo estimation," in *Proc. ECCV*, 2012, pp. 45–58. 6

[54] H. Fu, D. Cohen-Or, G. Dror, and A. Sheffer, "Upright orientation of man-made objects," *ACM Trans. Graph.*, vol. 27, no. 3, p. 42, 2008. 7

[55] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, "Orientation-boosted voxel nets for 3d object recognition," in *Proc. BMVC*, 2017. 7

[56] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," in *Proc. CVPR Workshop on Future of Datasets*, 2015. 7, 9

[57] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision Res.*, vol. 116, pp. 165–178, 2015. 8, 9

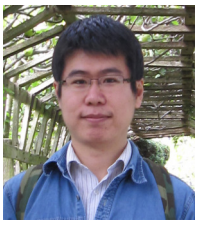
[58] T. Hou and H. Qin, "Admissible diffusion wavelets and their applications in space-frequency processing," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 1, pp. 3–15, 2013. 9, 10, 13

[59] X. Ding, W. Lin, Z. Chen, and X. Zhang, "Point cloud saliency detection by local and global feature fusion," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5379–5393, 2019. 10, 12

[60] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," in *Proc. SGP*, 2009, pp. 1383–1392. 12

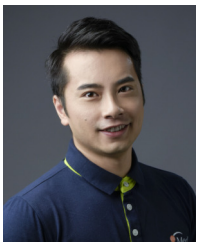
[61] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. ECCV*, 2012, pp. 101–115. 13

[62] A. Banitalebi-Dehkordi, E. Nasiopoulos, M. T. Pourazad, and P. Nasiopoulos, "Benchmark 3d eye-tracking dataset for visual saliency prediction on stereoscopic 3d video," *arXiv preprint arXiv:1803.04845*, 2018. 13



Prof Ran Song is a Professor with the School of Control Science and Engineering, Shandong University, China since 2020. Before his current post, he was a senior lecturer at the University of Brighton, UK. He received his Ph.D. degree in electronic engineering from the University of York, UK in 2009 and his first degree from Shandong University in 2005. He has published more than 60 papers in peer-reviewed international conference proceedings and journals. His research interests lie in 3D shape analysis, 3D visual perception and 3D vision for robotics.

Prof Wei Zhang is currently a Professor with the School of Control Science and Engineering, Shandong University, China. He received the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong in 2010. He has published over 70 papers in international journals and refereed conferences. His research interests include computer vision, image processing, pattern recognition, and robotics. Dr. Zhang served as a program committee member and a reviewer for various international conferences and journals in image processing, computer vision, and robotics.



Prof Yitian Zhao is currently the Director and Professor of the Lab of Intelligent Medical Imaging (iMED) at Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, and also an honorary lecturer with the Department of Eye and Vision Science, University of Liverpool, UK. He finished his Ph.D. degree in 3D image analysis at Aberystwyth University, UK in 2013. Previously, he was a lecturer at the Beijing Institute of Technology from 2015 to 2017. His research expertise is ophthalmic medical image processing, vessel structure analysis, eye and brain joint computing.

Prof Yonghuai Liu is a Professor with the Department of Computer Science, Edge Hill University, UK since 2018. Before that, he was a senior lecturer at Aberystwyth University, UK. He is currently associate editor and an editorial board member for a number of international journals, including Pattern Recognition Letters and Neurocomputing. He has published three books and more than 180 papers in international conference proceedings and journals. His primary research interests lie in 3D computer vision. He is a senior member of IEEE and Fellow of Higher Education Academy of United Kingdom.

Prof Paul L. Rosin is a Professor with the School of Computer Science and Informatics, Cardiff University, UK. Previous posts include lecturer at Brunel University London, UK, research scientist at the Institute for Remote Sensing Applications, Joint Research Centre, Ispra, Italy, and lecturer at Curtin University of Technology, Perth, Australia. His research interests include image representation, semantic segmentation, low level image processing, machine vision approaches to remote sensing, methods for evaluation of approximation algorithms, medical and biological image analysis, mesh processing, non-photorealistic rendering and analysis of shape in art and architecture.