

Learning Category-level Shape Saliency via Deep Implicit Surface Networks

Chaozheng Wu¹, Lin Sun², Xun Xu³, Kui Jia^{1*}

¹South China University of Technology, ²Samsung, USA, ³I2R, ASTAR

eeczwu@mail.scut.edu.cn, sunlin@cs.stanford.edu,

alex.xun.xu@gmail.com, kuijia@scut.edu.cn

Abstract

This paper is motivated from a fundamental curiosity on what defines a category of object shapes. For example, we may have the common knowledge that a plane has wings, and a chair has legs. Given the large shape variations among different instances of a same category, we are formally interested in developing a quantity defined for individual points on a continuous object surface; the quantity specifies how individual surface points contribute to the formation of the shape as the category. We term such a quantity as category-level shape saliency or shape saliency for short. Technically, we propose to learn saliency maps for shape instances of a same category from a deep implicit surface network; sensible saliency scores for sampled points in the implicit surface field are predicted by constraining the capacity of input latent code. We also enhance the saliency prediction with an additional loss of contrastive training. We expect such learned surface maps of shape saliency to have the properties of smoothness, symmetry, and semantic representativeness. We verify these properties by comparing our method with alternative ways of saliency computation. Notably, we show that by leveraging the learned shape saliency, we are able to reconstruct either category-salient or instance-specific parts of object surfaces; semantic representativeness of the learned saliency is also reflected in its efficacy to guide the selection of surface points for better point cloud classification.

1. Introduction

The conception of saliency in cognitive science is referred to the information of interest filtered by human visual system (HVS) [2] which is distinct from those around it. Saliency learning is important in computer vision, especially in pattern recognition, as it help us understand what contributes to the success of recognition and what features are of high importance with comparison to others [23, 38].

*Corresponding author

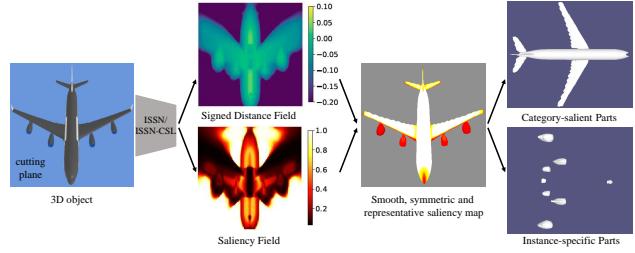


Figure 1: Abstract concept of ISSN and ISSN-CSL, that can learn continuous signed distance fields and saliency fields of objects in 3D space simultaneously. The learned category-level saliency maps have the properties of smoothness, symmetry, and semantic representativeness. By leveraging the learned shape saliency, we are able to reconstruct either category-salient or instance-specific parts of object surfaces.

One uses the knowledge of saliency to localize the target objects in the images [23], design adversarial attack [18], etc. Moreover, saliency map estimation on 3D data can benefit many applications. For example, Lee *et al.* introduce the mesh saliency and apply it to mesh simplification and good view selection. Song *et al.* [27] estimate mesh saliency via spectral processing and apply it to mesh segmentation and scan integration. Saliency map estimation also shows potentials for other applications such as shape matching and retrieval, alignment and sampling.

Compared to 2D image processing, saliency map estimation on 3D data is challenging due to following factors: (1) unlike the regular and discrete representation of 2D image, 3D data is irregular and continuous, e.g. the locations of the mesh vertices can be anywhere in 3D space, (2) 3D data is invariant to permutation, e.g. if we shuffle the order of the point cloud, its appearance does not change, (3) the degree of freedom of transformation in 3D space is more than that in 2D space, making it hard to learn transformation invariant features.

There has been numerous studies to investigate the mesh saliency which can be categorized into local contrast based and global contrast based methods. Many of the local con-

trast based methods [24, 25, 37, 8] extend the classical operators in 2D image processing to 3D space to detect the local variance, *e.g.* Harris [12], SIFT [14], *etc*. Nevertheless, the saliency maps estimated by these methods are dispersive and small-scale. On the other hand, global contrast based methods [28, 26, 30] try to estimate saliency maps in large scale. However, most of them are built on local contrast based methods. As a result, the saliency maps estimated by these methods are still lack of representativeness. Very recently, following [23], Zheng *et al.* [38] introduce the Point Cloud Saliency Map (PCSM) to quantify the contribution of each point to the semantic classification task via the backward gradient on the point. Although such saliency maps can capture the representativeness of the object to some extend, they look noisy and discontinuous which is a common issue of such gradient based methods [10].

Recently, 3D shape reconstruction has been attracting great attention in computer vision, we find that the dominant methods can reconstruct the common parts shared among objects of the same category much easier than the instance-specific details. In this paper, we rethink two fundamental problems in 3D saliency map estimation about what defines a category of object shapes and what contributes to discriminate the shapes of different categories. Naturally, we turn to 3D shape reconstruction to learn the representativeness of a category of objects. Specifically, we choose a recently proposed implicit fields learning framework DeepSDF [19] as our backbone, where the continuous signed distance function (SDF) is learned to represent the object shape, to learn a continuous saliency maps along the manifolds of the object surfaces. We term our proposed method the Implicit Surface Saliency Network (ISSN). Furthermore, motivated by Tags2Parts [17] where the region of interest specified by the tags can be detected using a binary classification network, we introduce the contrastive saliency learning to generate the saliency maps that can discriminate between the shapes of the category of interest and those of other categories. We term our implicit surface saliency network with contrastive saliency learning the ISSN-CSL. The abstract concept of ISSN and ISSN-CSL is shown in Figure 1. Unlike PCSM [38] where saliency maps are obtained by backward gradient of classification network, both our proposed ISSN and ISSN-CSL generate the saliency maps via a feedforward decoder. Experiments show that the saliency maps learned by ISSN and ISSN-CSL have excellent properties of smoothness, symmetry, and semantic representativeness. Notably, by leveraging the learned shape saliency of ISSN, we are able to reconstruct either category-salient or instance-specific parts of object surfaces. We summarize our contributions as follows:

- We are the first to introduce two implicit surface saliency network, ISSN and the one with contrastive saliency learning ISSN-CSL, to learn category-level

shape saliency via deep implicit surface networks.

- To compare the smoothness and symmetry of saliency maps of different methods quantitatively, we introduce two evaluation metrics, smoothness ratio and symmetry distance of saliency. To quantify the representativeness of the saliency maps, we design the saliency points classification task.
- We demonstrate that the saliency maps learned by our proposed ISSN and ISSN-CSL have better properties of smoothness, symmetry, and semantic representativeness than those of other methods. We also conduct experiments to show that the saliency maps of ISSN can be used to reconstruct either category-salient or instance-specific parts of object surfaces.

2. Related Works

2.1. 3D Shape Reconstruction

Traditionally, people reconstruct the complete 3D object surface by using the stereo correspondences from multi-view 2D images. 3D shape reconstruction from single image using learning-based methods have been explored in prior studies. Most of them decode the embedding of 2D image into 3D voxel [6, 7, 32]. However, the reconstructed results of these approaches are low-resolution due the sparsity of 3D voxel and limitation of GPU memory. Most relevant to our work, Chen and Zhang [5] propose the IM-Net to learn an implicit field of object in the continuous 3D space which is divided into two parts, inside and outside the object, and the implicit field learning can be recast as a binary classification task. Similarly, Park *et al.* [19] propose the DeepSDF to estimate the continuous Signed Distance Function (SDF) in 3D space. For the space inside the object, $SDF < 0$, and for the space outside the object, $SDF > 0$. The surface of object is implicitly encoded on the manifold where $SDF = 0$. The major differences between IM-Net and DeepSDF are that IM-Net adopts an auto-encoder network and learns the implicit field via a classification loss, while DeepSDF introduces a novel auto-decoder network to regress the continuous SDF values. After training the implicit field, both IM-Net [5] and DeepSDF [19] adopt the Marching Cubes algorithm to obtain the object surface and the reconstructed object can be achieved at arbitrary resolution depending on the grid size.

2.2. Saliency Map Estimation

In the field of computer science, saliency firstly emerged in 2D image analysis. With the increasing demand of geometric analysis and processing on 3D data, saliency map estimation is drawing attention in graphics community recently. Borrowing ideas from 2D image processing, Sipiran

and Bustos [24, 25] introduce the 3D version of Harris operator [12], Zaharescu *et al.* [37] develop the MeshDoG to detect the salient points, and Godil and Wagan [8] extend the SIFT operator [14] to 3D space. Song *et al.* [28] propose the conditional-random field-based (CRF) method to detect mesh saliency, which is more effective and robust. To capture semantic representativeness of point cloud data, Zheng *et al.* [38] and Gupta *et al.* [10] propose to estimate the point cloud saliency maps using the gradients of trained classification networks. In contrast, we take a rather different approach that learns the continuous shape saliency maps via deep implicit surface networks.

2.3. Contrastive Learning

With the rapid development of self-supervised learning, contrastive learning is getting more and more attention. Wu *et al.* [34] treat each image instance as a category and utilize noise-contrastive estimation (NCE) [11] to tackle challenge of computing the similarity among a large number of instances, aiming to capture the instance-level apparent similarity. Tian *et al.* [29] propose the contrastive multi-view coding to learn the features shared among different views of the same scene. Misra and Maaten [16] introduce the Pretext-Invariant Representation Learning (PIRL) that learns representations invariant to transformation by solving the jigsaw puzzles. Chen *et al.* [4] design the simplified contrastive learning of visual representations (SimCLR) that does not require memory bank or special architectures. He *et al.* [13] develop the Momentum Contrast (MoCo) where they build a dynamic dictionary with a queue and a moving-averaged encoder so as to accelerate contrastive learning. Grill *et al.* [9] propose an approach BYOL that does not require negative samples and replace the NCE-based loss with L2 loss. In 3D space, Xie *et al.* [35] propose PointContrast to learn representations on 3D real point cloud data. They find that the pretrained model can be used to boost the performances of existing methods on several benchmarks of 3D segmentation and detection. Muralikrishnan *et al.* [17] utilize a simple binary classification network to specify whether the input object contains the region specified by the tag or not. As a result, they can segment the region of interest from other regions on the object. Similarly, we propose the contrastive saliency learning based on a binary classification network to capture the representativeness of the category of interest.

3. The Proposed Method

We illustrate two methods, Implicit Surface Saliency Networks (ISSN) and Implicit Surface Saliency Networks with Contrastive Saliency Learning (ISSN-CSL), to estimate the shape saliency maps for objects from the same category via deep implicit surface networks which are initially designed for shape reconstruction. Our motivation lies

in that the intra-category objects share some common parts and it's easy for the network to estimate the implicit field of these common parts, while it may not be easy to estimate the implicit field of the instance-specific parts. Therefore, we design the network to output saliency score associated with the SDF value for every query point, and use the saliency score to weight the corresponding SDF error (Section 3.1). The saliency map can represent the category commonality of the object surface. However, this kind of saliency map may not capture the semantic representativeness of the object. In order to generate the saliency map that can capture the semantic cues for the objects from the category of interest, we introduce the contrastive saliency learning that discriminates between instances of the category of interest and those of other categories (Section 3.2).

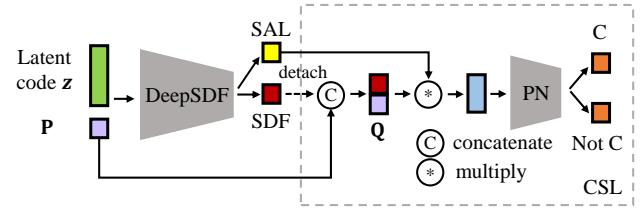


Figure 2: The Architecture of Proposed ISSN and ISSN-CSL. The input of the networks are the latent code z and a set of query points P , and the output are the SDF values and saliency scores (SAL). For contrastive saliency learning (CSL), the points in P are concatenated with their predicted SDF values (denoted as Q) and multiplied by the saliency scores, then pass through a PointNet (PN) [21] to specify whether the object denoted by Q is from the category of interest (C) or not.

3.1. Implicit Surface Saliency Network

The architecture of ISSN is illustrated in Figure 2. It is built on a Sign Distance Function (SDF) estimation network [19]. The input of the network are the latent code z that encodes the shape of object and query points P in 3D space. Different from DeepSDF, our network outputs the point-wise SDF values and saliency scores (SAL) simultaneously. To make the network predict the saliency scores in an unsupervised manner, we apply the loss function:

$$\mathcal{L}_{issn} = \frac{1}{N} \sum_{i=0}^N (\mathcal{L}_i s_i - \lambda \log(s_i)) + \beta \|z\|_2^2, \quad (1)$$

where N denotes the number of input query points, and s_i is the saliency score of input query point p_i . Similar to [19], \mathcal{L}_i is the L_1 loss that penalizes L_1 distance between the predicted and Ground-Truth (GT) SDF values,

$$\mathcal{L}_i = \left| \text{clamp}(f_i, \delta) - \text{clamp}(\hat{f}_i, \delta) \right|, \quad (2)$$

where f_i and \hat{f}_i are the predicted and GT SDF values of query point \mathbf{p}_i . Please refer to [19] for the definition of *clamp* function and the choice of δ . We introduce a regularization term $\log(s_i)$ in Equation (1) to avoid vanishing of s_i , and λ is to balance the weighted SDF loss and saliency scores. Besides, we add $L2$ -norm of the latent \mathbf{z} to constrain the expressiveness so that decoder can focus on the reconstruction of intra-category common part while omits the instance-specific details. Note that if λ in Equation (1) is set to 0, we will not blend s_i with L_i , the loss will degrade to DeepSDF loss.

To understand why the proposed method works, here, we provide more insights into it. An object instance consists of common parts \mathbb{C} shared among objects from the same category and instance-specific details \mathbb{D} . It is easier for the network to regress the SDF values of query points around \mathbb{C} than those around \mathbb{D} . Thus, for \mathbf{p}_i around \mathbb{C} , \mathcal{L}_i will have smaller value, then incur high penalty from regularization term $-\log(s_i)$ and lead to a high value of s_i . On the contrary, for \mathbf{p}_i around \mathbb{D} , \mathcal{L}_i will have larger value, then bring about low value of s_i .

An alternative way to estimate this kind of saliency map is using the Principle Component Analysis (PCA). As shown in [20, 1], PCA is usually used to estimate the curvature of the manifold, which captures the local variance of single object, while ours is able to capture how individual surface points contribute to the formation of the shape as the category.

3.2. Contrastive Saliency Learning

The design strategy in Section 3.1 enables the network to capture the common parts shared among the objects of the same category, but it may be lack of semantic representativeness which discriminates between instances of the category of interest and those of other categories. In PCSM [38], the authors proposed to estimate the point cloud saliency map based on the derivative of the loss function of a trained classification network with respect to the input point cloud. Although the saliency map of PCSM contains the semantic information to some extent, PCSM inevitably face the noise and discontinuities which are common in gradient based saliency map [10]. Inspired by the Tags2Parts [17] where the region of interest specified by the tag can be segmented via a simple binary classification networks, we introduce a contrastive classification network following the ISSN to estimate the saliency map that can capture the semantic cues of the objects from the category of interest, giving rise to ISSN-CSL.

During training, we sample the instances in C and those not in C equally in a mini batch. Besides the ISSN regression loss in Equation (2), we add a binary classification loss:

$$\mathcal{L}_{cls} = -\hat{l} \log(l) - (1 - \hat{l}) \log(1 - l), \quad (3)$$

where \hat{l} and l are the GT label and predicted score, respectively. For the shapes of the category of interest, $\hat{l} = 1$, otherwise, $\hat{l} = 0$. The total loss function for ISSN-CSL is:

$$\mathcal{L}_{csl} = \mathcal{L}_{issn} + \gamma \mathcal{L}_{cls}, \quad (4)$$

where γ is a hyperparameter to balance \mathcal{L}_{issn} and \mathcal{L}_{cls} .

At last, we review the architecture step by step, as shown in Figure 2, given a point set \mathbf{P} with N points and the latent code \mathbf{z} with D dimension, we firstly duplicate the latent code N times and concatenate it with the point coordinates in \mathbf{P} , then feed them to the DeepSDF network to predict the point-wise SDF values \mathbf{F} and saliency scores \mathbf{S} . After that, \mathbf{F} is detached from DeepSDF network and concatenated with \mathbf{P} , formed \mathbf{Q} . \mathbf{Q} can represent the shape of the object. Finally, we blend \mathbf{S} with \mathbf{Q} and feed it into a PointNet [21] to classify whether \mathbf{Q} is from the category of interest C or not. For those points around the discriminative parts of the category of interest, they will contribute more to the contrastive classification, then the corresponding saliency scores will be higher. We will show the saliency map of ISSN-CSL has the potential to boost the performance of point cloud classification in the later part.

3.3. Saliency Map Properties

To evaluate the properties of smoothness and symmetry of the saliency maps estimated quantitatively, we introduce two evaluation metrics, saliency smoothness ratio and saliency symmetry distance.

3.3.1 Saliency Smoothness

Given the saliency maps \mathbf{S} of a point cloud \mathbf{P} , to quantitatively evaluate the saliency smoothness, we define our saliency smoothness ratio (SSR) as follows,

$$SSR = \frac{1}{N} \sum_{i=1}^N |s_i - \tilde{s}_i|, \quad (5)$$

where \tilde{s}_i is the saliency score of the nearest neighbor of point \mathbf{p}_i . The smaller value of SSR indicates the smoother saliency map.

3.3.2 Saliency Symmetry

We denote the point cloud of an object by \mathbf{P} and obtain its mirror point cloud \mathbf{P}' by flipping \mathbf{P} along the plane of symmetry. The plane of symmetry for each category need to be searched manually. For each $\mathbf{p}'_i \in \mathbf{P}'$, we find its nearest neighbor \mathbf{p}_i on \mathbf{P} , as illustrated in Figure 3f. If the maximum distance between \mathbf{p}'_i and \mathbf{p}_i is less than 10cm, we consider it a symmetry object, otherwise, it is asymmetric.

To evaluate the symmetry of saliency map quantitatively, we define the symmetry distance of saliency as following:

$$d_{sym} = \frac{1}{N} \sum_{i=1}^N |s_i - s'_i|, \quad (6)$$

where s_i and s'_i are the saliency scores of \mathbf{p}_i and \mathbf{p}'_i , respectively.

4. Experiments

4.1. Datasets

To verify our proposed method, two public datasets, ModelNet40 [33] and ShapeNet [3] are used. The objects in both datasets are aligned to canonical poses.

ModelNet40 [33] - ModelNet40 contains 12,311 objects from 40 categories, 9,843 objects of which are for training and 2,468 objects are for test. The object models are normalized to unit sphere, and 250,000 query points and corresponding SDF values are sampled from the unit sphere for each object model, where the space near the mesh surface is sampled much denser than the space far away. 1,024 points are sampled from the object model using Farthest Point Sampling (FPS) for object classification.

ShapeNet [3] - We only use the objects existing in ShapeNet Part [36] for training and testing. There are 16,880 objects form 16 categories in ShapeNet Part, 14,006 objects of which are for training and 2,874 objects are for test. The training data is obtained in the same way as ModelNet40.

4.2. Implementation Details

The input of ISSN and ISSN-CSL are 16,384 points randomly sampled from 250,000 query points of each object and the latent codes initialized with normal distribution. We train ISSN and ISSN-CSL with the batch size of 48 on two 2080Ti GPU. The samples of category of interest and other categories are sampled equally in a mini-batch of ISSN-CSL. We optimize the weights in decoder and latent codes jointly using Adam optimizer with initial learning rate $5e^{-4}$ and $1e^{-3}$, respectively, for 500 epochs (divided by 2 every 200 epochs). β and δ are set to 0.1. The values of λ and γ will be specified in the experiments. We will discuss how we set the values of hyper-parameters in Appendix A.

4.3. Saliency Map Analysis

In this Section, we will present quantitative comparisons of different saliency maps with respect to the metrics of saliency smoothness ratio and saliency symmetry distance.

4.3.1 Smoothness Analysis

We evaluate the smoothness of the saliency maps generated by different methods quantitatively using SSR. The results are shown in Table 1 and Table 2. It clearly shows that the saliency maps of our proposed methods are smoother than those of PCSM [38] and PCA [1].

Methods	Mean	Min	Max
PCA	0.119	0.024	0.352
PCSM	0.311	0.257	0.347
ISSN ($\lambda = 1e - 3$)	0.089	0.012	0.315
ISSN-CSL ($\lambda = 0, \gamma = 1.0$)	0.059	0.011	0.392
ISSN-CSL ($\lambda = 1e - 3, \gamma = 0.1$)	0.042	0.009	0.178

Table 1: Statistical SSR of objects on ModelNet40 [33]. Smaller is better for all columns.

Methods	Mean	Min	Max
PCA	0.083	0.025	0.214
PCSM	0.300	0.265	0.331
ISSN ($\lambda = 1e - 3$)	0.062	0.013	0.274
ISSN-CSL ($\lambda = 0, \gamma = 1.0$)	0.033	0.007	0.228
ISSN-CSL ($\lambda = 1e - 3, \gamma = 0.1$)	0.043	0.004	0.243

Table 2: Statistical SSR of objects on ShapeNet [3]. Smaller is better for all columns.

4.3.2 Symmetry Analysis

We evaluate the symmetry of the saliency maps on the symmetric objects quantitatively using our proposed symmetry distance. The results are shown in Table 3 and Table 4. We can find that the saliency maps of our proposed methods can maintain better symmetry than those of PCSM [38] and PCA [1].

Methods	Mean	Min	Max
PCA	0.105	0.024	0.460
PCSM	0.305	0.234	0.346
ISSN ($\lambda = 1e - 3$)	0.064	0.017	0.333
ISSN-CSL ($\lambda = 0, \gamma = 1.0$)	0.161	0.026	0.492
ISSN-CSL ($\lambda = 1e - 3, \gamma = 0.1$)	0.157	0.016	0.507

Table 3: Statistical symmetry distance of saliency map on ModelNet40 [33]. Smaller is better for all columns.

4.4. Visualization

In this section, we will visualize the saliency maps on point clouds and analyze the saliency maps generated by different methods.

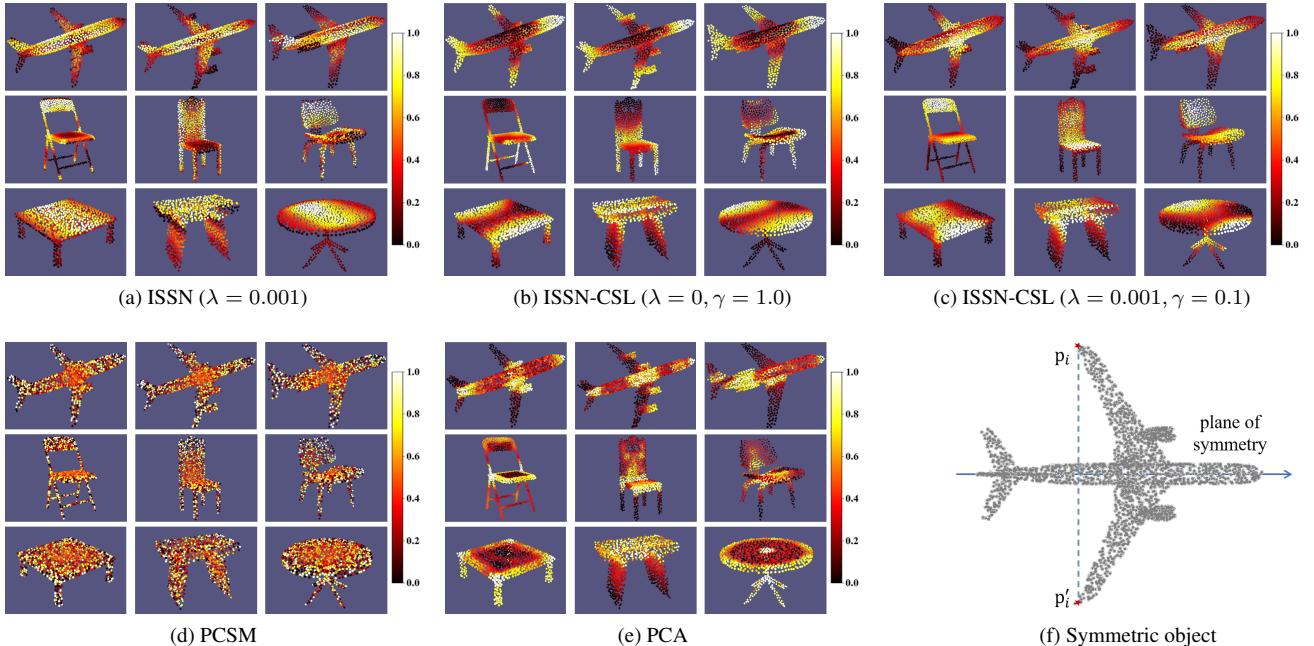


Figure 3: (a-e) Visualization of saliency maps of different methods on point cloud. These models are form ModelNet40 [33] and the saliency maps of each object are normalized to $[0, 1]$. (f) An example of symmetric object.

Methods	Mean	Min	Max
PCA	0.144	0.007	0.408
PCSM	0.293	0.079	0.341
ISSN ($\lambda = 1e - 3$)	0.067	0.016	0.295
ISSN-CSL ($\lambda = 0, \gamma = 1.0$)	0.077	0.007	0.403
ISSN-CSL ($\lambda = 1e - 3, \gamma = 0.1$)	0.073	0.001	0.500

Table 4: Statistical symmetry distance of saliency map on ShapeNet [3]. Smaller is better for all columns.

For the point cloud \mathbf{P} sampled from the mesh surface, it's easy for us to obtain its saliency map by feeding the points in \mathbf{P} and its corresponding latent code into the decoder directly. Figure 3 shows the saliency maps generated by our proposed ISSN and ISSN-CSL, PCSM [38], and PCA [1] on ModelNet40 [33]. (More visualization results of our methods are shown in Appendix B.) For fair comparisons, all of the saliency maps are normalized to $[0, 1]$. The figures demonstrate that the saliency maps of our proposed methods are much smoother than those of PCSM and PCA. Besides, the saliency maps of our proposed ISSN can maintain symmetry for those symmetric objects, which are even better than those of the analytic method, PCA. Although the saliency maps of our proposed ISSN-CSL are asymmetric on some categories and look different when λ and γ are assigned with different values, but they can capture consistent pattern across different objects from the same cate-

gory. Contrarily, saliency maps generated by PCSM look noisy and asymmetric and we can not find any regular pattern among them.

Apart from the properties of smoothness and symmetry, the saliency maps generated by our proposed ISSN indeed capture the representativeness and common parts shared among the objects of the same category. For example, all of the airplanes in Figure 3a have similar fuselages and wings but different engines and empennages. As a result, the saliency scores on the fuselages and wings are much higher than those on the engines and empennages. The regular pattern can also be found on the objects from other category, e.g. the chairs and the tables in Figure 3a. Moreover, the saliency map generated by ISSN-CSL can capture discriminations among different categories. As shown in 3b, both the chairs and tables have legs, to avoid confusing, it guides the saliency maps of chairs to focus on their legs and the saliency maps of tables to focus on their tops. The saliency map of both ISSN and ISSN-CSL can capture consistent pattern across different objects from the same category. In Section 5.2, we will show their potential on improving the performance of point cloud classification.

5. Applications

A smooth and symmetry saliency map with semantic representations will contribute to the reconstruction of category-salient and instance-specific parts and point cloud classification.

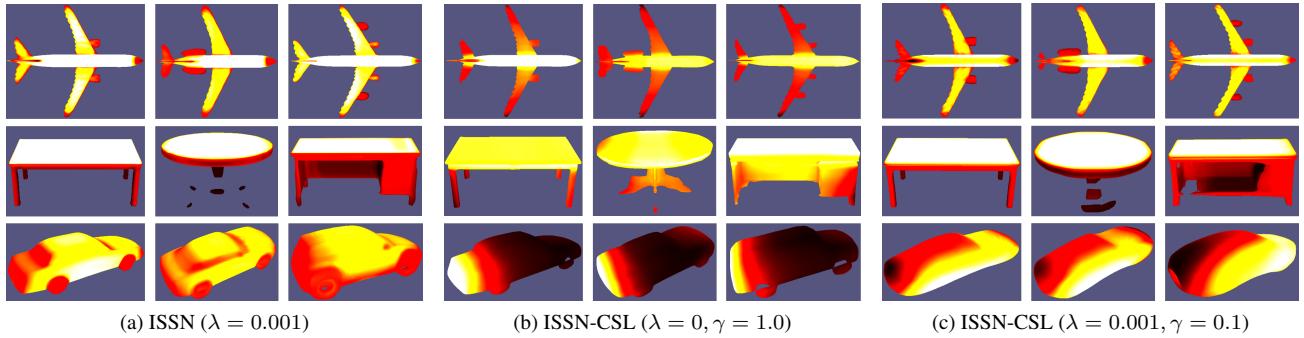


Figure 4: Visualization of saliency maps on reconstructed meshes by our proposed methods. The brighter color denotes higher saliency. These models are from ShapeNet [3].

5.1. Reconstruction

During test phase, we sample 100,000 query points near the surfaces of objects and calculate corresponding SDF values with the help of point clouds and normals provided in [36]. Then, we fix the weights of decoder in DeepSDF and use the sampled query points and SDF values to optimize the latent codes. With the trained decoder and optimized latent codes, we feed the coordinates of a normalized 3D grid with resolution of 256^3 into the network to obtain the SDF values and corresponding saliency scores. Then, we apply the Marching Cubes algorithm to reconstruct the mesh model. Note that the saliency scores are interpolated together with the SDF values when searching the vertices on the zero iso-surface in the Marching Cubes algorithm. The reconstructed meshes and saliency maps on ShapeNet [3] are shown in Figure 4. We can find that ISSN and ISSN-CSL can learn to generate continuous and smooth saliency map along the manifold of the object surface and maintain the symmetry property for those symmetric objects.

Another straightforward application of the saliency map of our proposed ISSN is to reconstruct the category-salient parts and the instance-specific parts, respectively. We remove those points whose saliency scores are below a threshold, then apply the Marching Cubes algorithm to reconstruct the common parts shared among the objects of the same category. We can also obtain the instance-specific parts by removing those points whose saliency scores are above a threshold. The results are shown in Figure 5.

5.2. Point Cloud Classification

To verify the efficacy of saliency maps generated by our proposed ISSN-CSL on point cloud classification, we use the saliency maps on training set of ModelNet [33] to modify the backward gradient of PointNet++ [22] and DGCNN [31]. We firstly train the saliency networks for each category on training set, and use the trained networks to predict the saliency map for each sampled point cloud on training

set. Since that we do not know the label of point cloud on test set, we introduce the following strategy to use the saliency map on training set to train the classification network.

We denote by $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^n$ the object point cloud, $\mathbf{S} = \{s_i\}_{i=1}^n$ its saliency map, \mathbf{f}_{mi} the feature of \mathbf{p}_i at layer m and \mathcal{L}_{cls} the classification loss. Then we replace the derivative of \mathcal{L}_{cls} with respect to \mathbf{f}_{mi} with $\mathbf{g}_{mi} = s_i \frac{\partial \mathcal{L}_{cls}}{\partial \mathbf{f}_{mi}}$ and use \mathbf{g}_{mi} to update the learnable weights at layer m . We use the PyTorch implementations of DGCNN [31] and PointNet++ [22] and keep the hyperparameters and optimizers consistent to ones in their papers.

The results are shown in Table 5, we can see that the performances on both PointNet++ and DGCNN can be improved consistently when trained with saliency maps.

Methods	Accuracy (%)	
	Overall	Avg. class
PointNet++ [22]	90.7	-
PointNet++ (our Exp.)	92.5	90.4
PointNet++ w/ saliency	93.3	91.3
DGCNN [31]	92.9	90.2
DGCNN (our Exp.)	93.3	91.0
DGCNN w/ saliency	93.7	90.4

Table 5: Classification results on ModelNet40 [33].

In order to show the representativeness and discrimination of the saliency maps estimated by different methods, we conduct the saliency points clustering experiment. Specifically, for the point cloud of each object in ModelNet40 [33], we obtain its saliency maps using our proposed ISSN and ISSN-CSL, PCSM [38], and PCA[1]. Then, we rank the points according to the saliency scores and select the top-10 points as the saliency points of the object. We utilize t-SNE [15] to clutter and visualize the saliency points of different objects. Note that we use the Hausdorff distance to calculate the distance between two sets of saliency points. The clustering results are illustrated in Figure 6. Our pro-

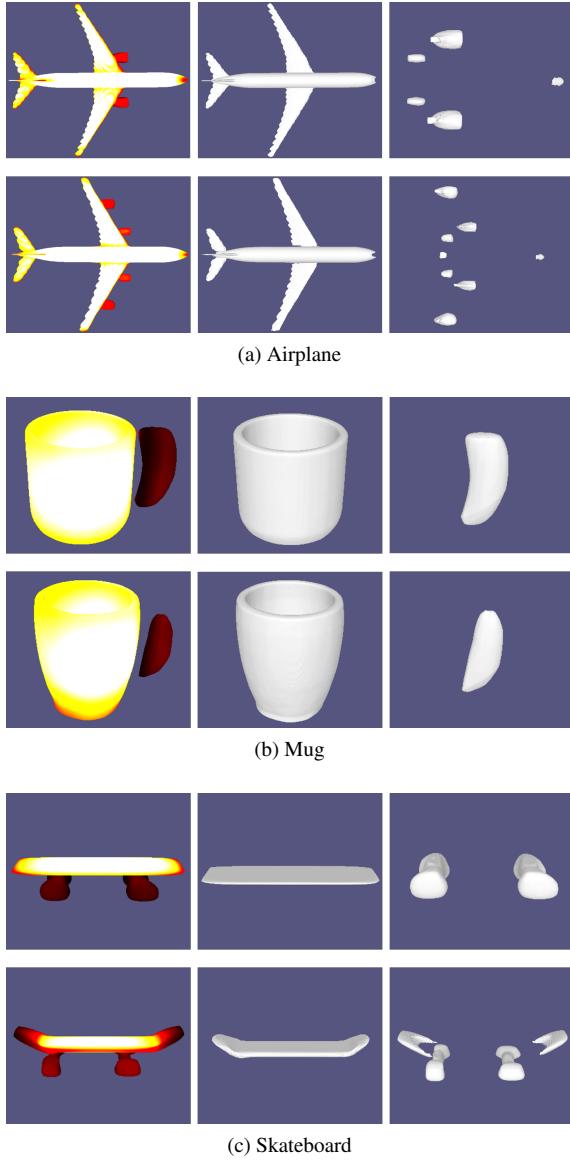


Figure 5: Reconstruction results. Saliency maps (left), category-salient parts (middle), and instance-specific parts (right).

posed ISSN-CSL presents the best clustering results, concentrating within the same category and discriminating between different category. Surprisingly, although we do not guide the ISSN to learn discrimination of instances from different categories, the clustering result is better than that of PCSM where the saliency maps are estimated via semantic classification networks.

Additionally, to compare the representativeness and discrimination of the saliency map estimated by different methods quantitatively, we design the saliency points classification task. In particular, we select the top- k salient points of each object in both training set and test set of Model-

Net40 [33]. Then, we use the salient points in training set to train DGCNN [31] and evaluate the overall accuracy on test set. The results are shown in Table 6. We can find that our proposed ISSN and ISSN-CSL outperform PCSM and PCA with a large margin, which are consistent to the clustering results. We also find that after dropping the redundant points based on the saliency maps of ISSN and ISSN-CSL, the performances of point classification can be improved a lot compared to the baseline using 1,024 points (overall accuracy is 93.3% in our experiment). The performances of PCSM and PCA on saliency points classification degrade dramatically when the number of saliency points decreases, while the performances of ISSN and ISSN-CSL are keeping robust. The training and testing curves are shown in Appendix C.

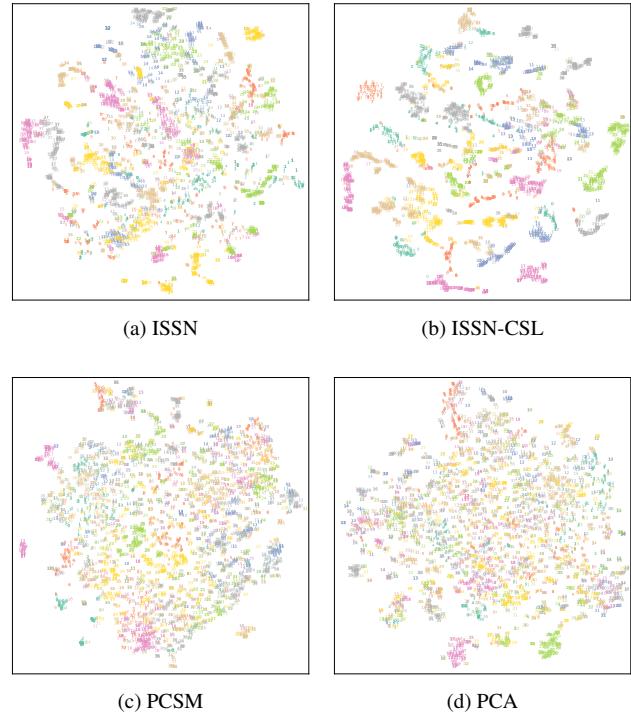


Figure 6: t-SNE clustering results of saliency points (zoom in for details, the numbers in the figures correspond to categories in ModelNet40).

# Saliency points	32	64	128	256	512
PCA	75.3	79.5	84.1	87.0	91.1
PCSM	85.6	87.9	90.7	92.3	92.9
ISSN	96.5	98.0	98.2	98.6	99.0
ISSN-CSL	98.0	98.8	99.4	99.6	99.7

Table 6: Overall accuracy (%) of saliency points classification on ModelNet40 [33] using DGCNN [31].

6. Conclusion

This paper exploits and explores a fundamental problem about what defines a category of object shapes. This is the first attempt in this direction. Given the large shape variations among different instances of a same category, our proposed ISSN and ISSN-CSL specify how individual surface points contribute to the formation of the shape as the category and the distinction between these instances and those of other categories, respectively. The learned saliency maps of our proposed methods have the properties of smoothness, symmetry, and semantic representativeness which contribute to reconstruction of either category-salient or instance-specific parts of object surfaces and point cloud classification.

References

- [1] Kwang-Ho Bae and Derek D Lichti. A method for automated registration of unorganised point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2008. [4](#), [5](#), [6](#), [7](#), [12](#)
- [2] John Berger. *About looking*. Bloomsbury Publishing, 2015. [1](#)
- [3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. [5](#), [6](#), [7](#), [11](#)
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. [3](#)
- [5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [2](#)
- [6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, 2016. [2](#)
- [7] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, 2016. [2](#)
- [8] Afzal Godil and Asim Imdad Wagan. Salient local 3d features for 3d shape retrieval. In *Three-Dimensional Imaging, Interaction, and Measurement*, 2011. [2](#), [3](#)
- [9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 2020. [3](#)
- [10] Ananya Gupta, Simon Watson, and Hujun Yin. 3d point cloud feature explanations using gradient-based methods. *arXiv preprint arXiv:2006.05548*, 2020. [2](#), [3](#), [4](#)
- [11] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010. [3](#)
- [12] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, 1988. [2](#), [3](#)
- [13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. [3](#)
- [14] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004. [2](#), [3](#)
- [15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. [7](#)
- [16] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. [3](#)
- [17] Sanjeev Muralikrishnan, Vladimir G Kim, and Siddhartha Chaudhuri. Tags2parts: Discovering semantic regions from shape tags. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [2](#), [3](#), [4](#)
- [18] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuRoS&P)*, 2016. [1](#)
- [19] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [2](#), [3](#), [4](#)
- [20] Mark Pauly, Richard Keiser, and Markus Gross. Multi-scale feature extraction on point-sampled surfaces. In *Computer graphics forum*, 2003. [4](#)
- [21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. [3](#), [4](#)
- [22] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, 2017. [7](#)
- [23] K. Simonyan, A. Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, 2014. [1](#), [2](#)
- [24] Ivan Sipiran and Benjamin Bustos. A robust 3d interest points detector based on harris operator. In *3DOR*, 2010. [2](#), [3](#)
- [25] Ivan Sipiran and Benjamin Bustos. Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer*, 2011. [2](#), [3](#)

- [26] Ivan Sipiran and Benjamin Bustos. Key-components: detection of salient regions on 3d meshes. *The Visual Computer*, 2013. 2
- [27] Ran Song, Yonghuai Liu, Ralph R Martin, and Paul L Rosin. Mesh saliency via spectral processing. *ACM Transactions on Graphics (TOG)*, 2014. 1
- [28] Ran Song, Yonghuai Liu, Yitian Zhao, Ralph R Martin, and Paul L Rosin. Conditional random field-based mesh saliency. In *2012 19th IEEE International Conference on Image Processing*, 2012. 2, 3
- [29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 3
- [30] 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. *Computer Aided Geometric Design*, 2015. 2
- [31] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 2019. 7, 8, 12, 13
- [32] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2
- [33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Ligang Zhang, Xiaou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 5, 6, 7, 8, 12, 13
- [34] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [35] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas J Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. *arXiv preprint arXiv:2007.10985*, 2020. 3
- [36] Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *SIGGRAPH Asia*, 2016. 5, 7
- [37] Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu Horaud. Surface feature detection and description with applications to mesh matching. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 3
- [38] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 3, 4, 5, 6, 7, 12

Appendix

In this appendix, we will provide more experiments to explain the hyper-parameters selection, present more saliency maps visualizations and training and testing curves on the saliency point classification.

A. Hyper-parameter Settings

A.1. λ in ISSN

We try different values of λ in ISSN, the results are shown in Figure 7. We find that when $\lambda \geq 1e-2$, the learned saliency maps all have high values over the object surfaces and can not capture differences between the category-salient parts and the instance-specific parts. When $\lambda = 1e-4$, the learned saliency maps will become noisy. When $\lambda < 1e-4$, the saliency scores will be vanished and even affect the SDF learning. Therefore, in the paper we set $\lambda = 1e-3$ to learn better saliency representations.

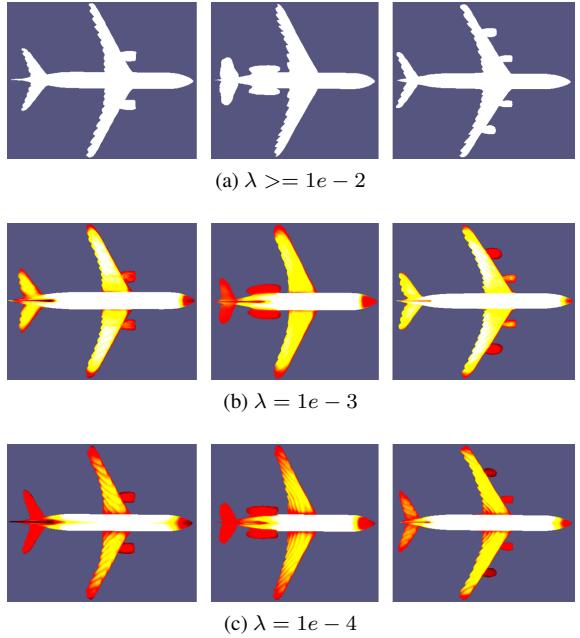


Figure 7: The influence of λ in ISSN.

A.2. λ and γ in ISSN-CSL

We further explore the influence of λ and γ in ISSN-CSL. When we set $\lambda = 0$, the influence of γ is shown in Figure 8. We find that when $\gamma = 1e-1, 1e-2, 1e-3$, the saliency maps does not show the ‘saliency’ with almost the same saliency scores on the object surfaces and look less smooth, while when $\gamma = 1.0$, the saliency maps focus on the fuselages more than other parts and look much smoother.

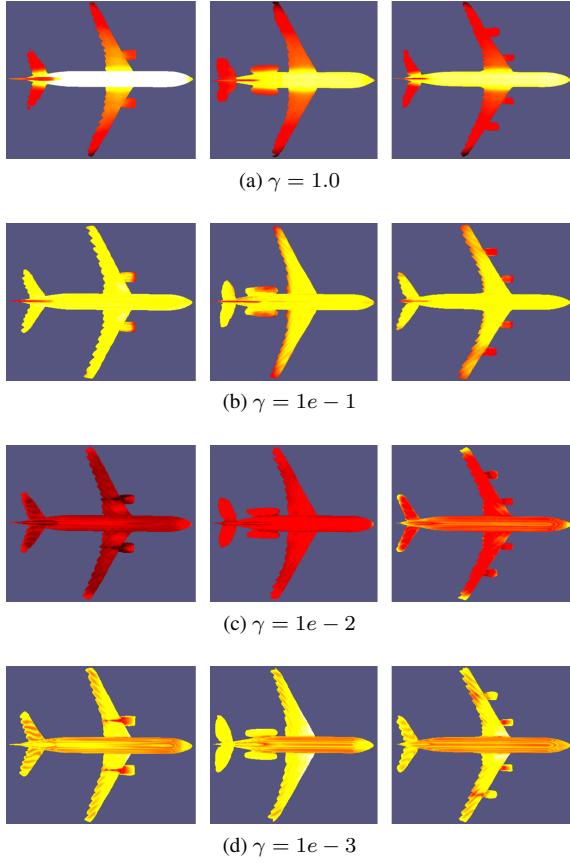


Figure 8: The influence of γ when λ is set to 0 in ISSN-CSL.

When $\lambda > 0$, we experiment that if the ratio of $\frac{\gamma}{\lambda} \geq 1000$, we can not reconstruct the object surfaces from the learned signed distance fields. In Section A.1, we are aware that when $\lambda \geq 1e-2$, the learned saliency maps have high values all over the object surfaces, so we can not set $\gamma \geq 1.0$. Therefore, in the main paper, we choose $\lambda = 1e-3$ and $\gamma = 0.1$. Here, we present more examples when λ and γ are set to other values in Figure 9. Corresponding saliency smoothness ratios and symmetry distances on ShapeNet [3] are shown in Table 7 and Table 8. Compared to our original setting in the main paper, the saliency maps of these three settings are comparably symmetric but less smooth.

ISSN-CSL	Mean	Min	Max
$\lambda = 1e-3, \gamma = 1e-2$	0.052	0.006	0.224
$\lambda = 1e-3, \gamma = 1e-3$	0.059	0.010	0.247
$\lambda = 1e-4, \gamma = 1e-3$	0.056	0.006	0.242

Table 7: Statistical SSR of objects on ShapeNet [3].

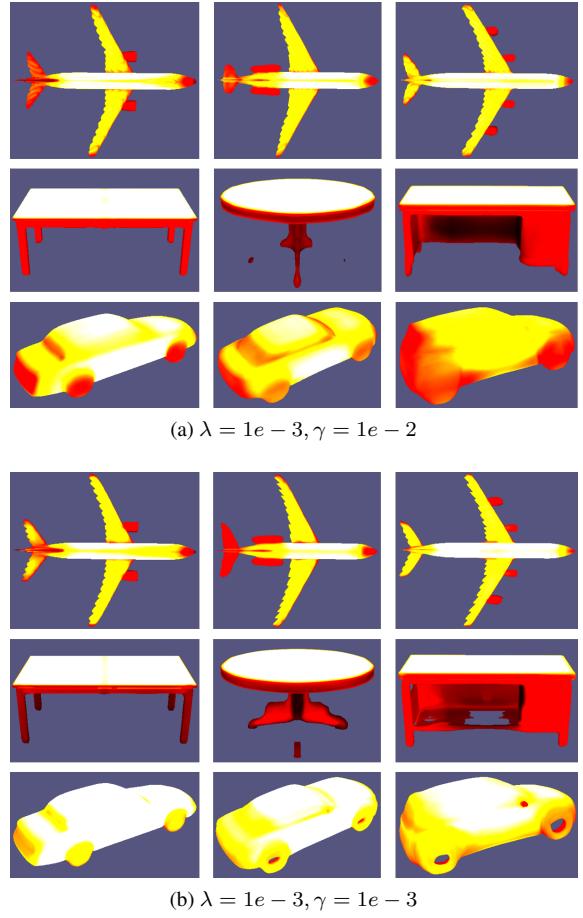


Figure 9: The influence of λ and γ . These models are from ShapeNet [3].

ISSN-CSL	Mean	Min	Max
$\lambda = 1e-3, \gamma = 1e-2$	0.073	0.008	0.493
$\lambda = 1e-3, \gamma = 1e-3$	0.070	0.018	0.492
$\lambda = 1e-4, \gamma = 1e-3$	0.085	0.009	0.496

Table 8: Statistical symmetry distance of saliency map on ShapeNet [3].

B. More Visualization of Saliency Maps

We present more examples of saliency maps on point clouds estimated by ISSN and ISSN-CSL in Figure 10. We can see ISSN and ISSN-CSL can learn to generate continuous and smooth saliency map along the manifold of the object surface and maintain the symmetry property for those symmetric objects. Our proposed ISSN indeed capture the representative and common parts shared among the objects of the same category while ISSN-CSL can capture the discriminative parts between different categories.

C. Training and Testing Curves on Saliency Points Classification

We show the training and testing accuracy curves of saliency points classification on ModelNet40 [33] using DGCNN [31] in Figure 11. We observe that ISSN and ISSN-CSL have better convergence rate compared to that of PCSM [38] and PCA [1]. We can find that after 120 epochs, all the methods become saturated on training set. However, ISSN and ISSN-CSL can achieve about 99% accuracy on test set, while PCSM and PCA only reach about 93% on test set, illustrating that the saliency points selected by ISSN and ISSN-CSL are more representative and the models trained on them generalize well from training set to test set, while the saliency points selected by PCSM and PCA are less representative and the models trained on them are more prone to overfitting.

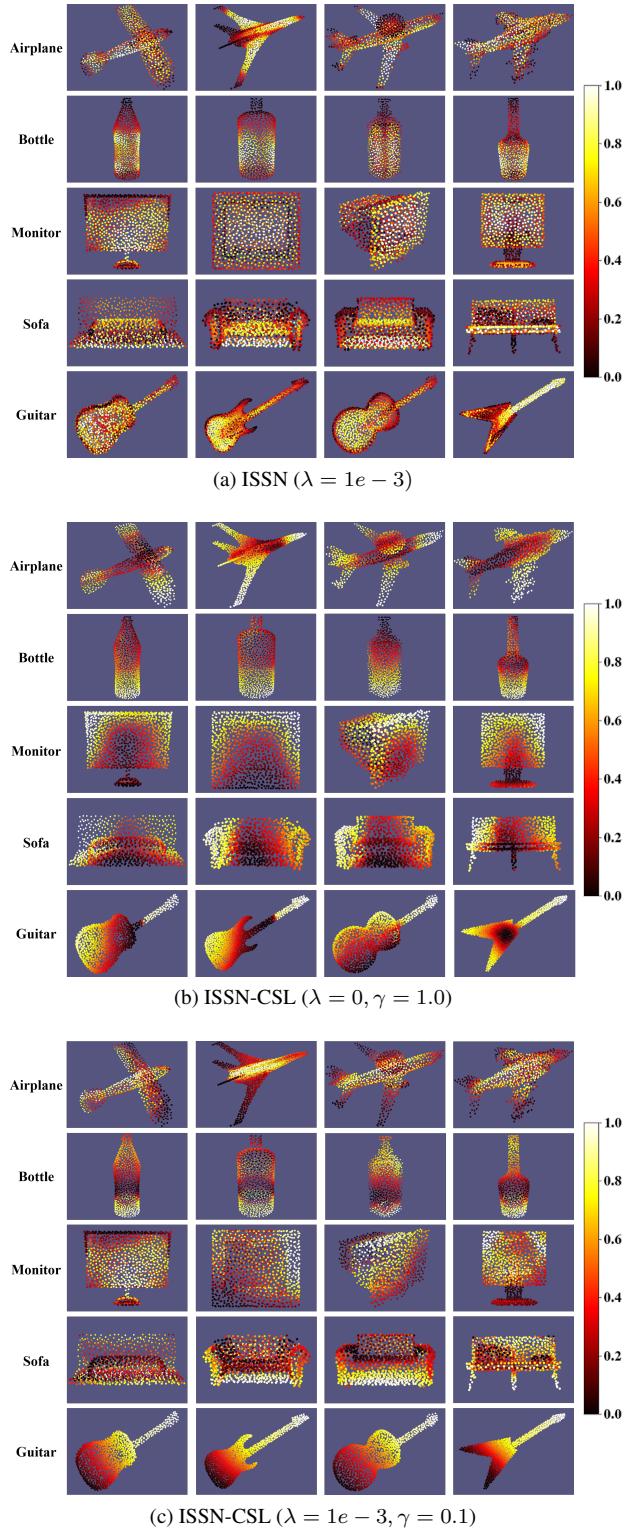
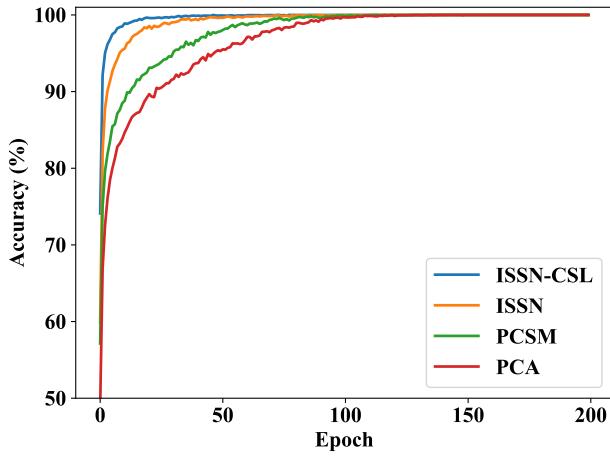
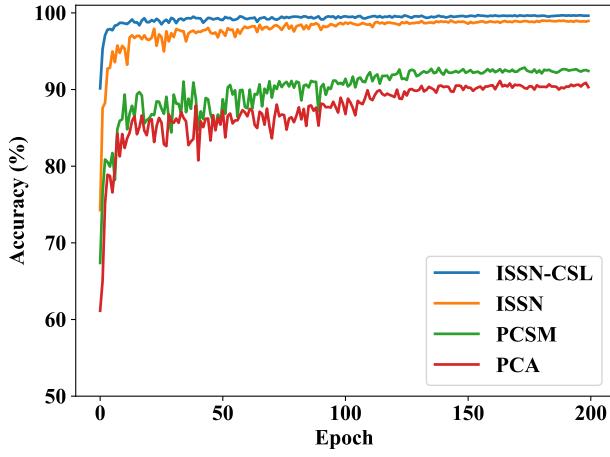


Figure 10: Saliency maps on point clouds. These models are from ModelNet40 [33].



(a) Training curves



(b) Testing curves

Figure 11: Overall accuracy curves of training and testing of saliency points classification on ModelNet40 [33] using DGCNN [31]. 512 saliency points of each object are used for training and testing.