Geometry-Aware 3D Salient Object Detection Network

Chen Wang[1], Liyuan Zhang[1], Le Hui[1,2]*, Qi Liu[1], Yuchao Dai[1]*

[1]Shaanxi Key Laboratory of Information Acquisition and Processing, Northwestern Polytechnical University
[2]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology
{chenw, zhangliyuannpu}@mail.nwpu.edu.cn, {huile, liuqi, daiyuchao}@nwpu.edu.cn

Abstract

Point cloud salient object detection has attracted the attention of researchers in recent years. Since existing works do not fully utilize the geometry context of 3D objects, blurry boundaries are generated when segmenting objects with complex backgrounds. In this paper, we propose a geometry-aware 3D salient object detection network that explicitly clusters points into superpoints to enhance the geometric boundaries of objects, thereby segmenting complete objects with clear boundaries. Specifically, we first propose a simple yet effective superpoint partition module to cluster points into superpoints. In order to improve the quality of superpoints, we present a point cloud class-agnostic loss to learn discriminative point features for clustering superpoints from the object. After obtaining superpoints, we then propose a geometry enhancement module that utilizes superpoint-point attention to aggregate geometric information into point features for predicting the salient map of the object with clear boundaries. Extensive experiments show that our method achieves new state-of-the-art performance on the PCSOD dataset.

Introduction

Salient object detection (SOD) focuses on segmenting the most attractive object from the surrounding background. As a pre-processing procedure, SOD has a variety of applications for many downstream tasks, such as semantic segmentation (Shi et al. 2021), object detection (Huang et al. 2020), and visual tracking (Mahadevan and Vasconcelos 2012). Many efforts (Fan et al. 2019; Liu et al. 2021; Zhou et al. 2020) are dedicated to image-based salient object detection, and plenty of well-known works have emerged. Recently, with the rapid development of 3D sensors, such as LiDAR and Kinect camera, acquiring 3D data has become more convenient, and growing numerous point cloud based applications. However, there are few works devoted to salient object detection on 3D point clouds. Due to the irregularity and sparsity of point cloud data, it is difficult to extend the methods designed for regular 2D images to point clouds. Therefore, there are also some unresolved issues in point cloud salient object detection.

Many excellent works (Wang 2020; Liu, Hou, and Cheng 2020; Huang, Xing, and Zou 2020) have emerged salient
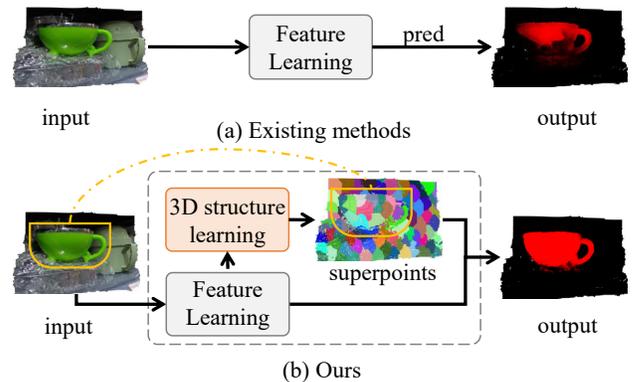
Figure 1: Comparison between our method and existing methods (PointSal (Fan, Gao, and Li 2022), EPFNet (Zhang et al. 2023) and PSOD-Net (Wei et al. 2024)) in 3D salient object detection task. We explicitly utilize the structural information of the point cloud by introducing superpoint representation. It can be observed that using superpoints can effectively characterize the structure of the green cup.

works on images. Current state-of-the-art methods (Wang et al. 2016; Liu and Han 2016; Hou et al. 2017; Chen and Li 2018; Zhang et al. 2020; Tu et al. 2021; Wang et al. 2021) usually use multiple enhancement strategies to improve the salient object detection performance. However, accurately segmenting the object boundaries from the image still remains challenging, due to the lack of effective spatial information. Therefore, researchers considered introducing different modals for studying multi-modal salient object detection. RGB-D salient object detection (Qu et al. 2017) uses depth maps to enhance the spatial information of objects for improving the accuracy of object boundaries. In addition, RGB-T salient object detection (Wang et al. 2018) uses RGB image combined with the thermal infrared image to locate and segment the common prominent object. Recently, researchers have begun to focus on point cloud salient object detection, *i.e.*, segmenting objects from the 3D point clouds. As a pioneer of point cloud salient object detection, Fan *et al.* (Fan, Gao, and Li 2022) proposed the first point cloud salient object detection dataset, and designed a simple multi-scale point network for 3D salient object detection task. Due to the lack of context information, it did not perform well on segmenting objects from complex backgrounds. Thus, Wei

et al. (Wei et al. 2024) proposed a point Transformer model to enhance contextual information of objects. Besides, in order to make full use of the complementary color information, Zhang *et al.* (Zhang et al. 2023) presented an enhanced point feature network that effectively fused the RGB information with 3D point clouds. Generally, the characteristic of the human vision system is its sensitivity to the boundaries and structural information of objects. Since these 3D salient object detection methods neglect important boundary and structural information, they are unable to segment the complete boundaries of objects from complex environments (as shown in Fig. 1(a)). Therefore, inspired by the human vision system, we consider clustering 3D point clouds into superpoints to learn the structural information of objects and fusing the superpoints with 3D points to enhance the learning of object boundaries. As depicted in Fig. 1(b), using superpoints can effectively characterize the structural information of the object (selected by the yellow line), thereby improving segmentation performance.

In this paper, we propose a geometry-aware 3D salient object detection network, called 3DGAS, which explicitly leverages the structural information of points to enhance object boundary segmentation. The entire network consists of two parts: a superpoint partition module and a geometry enhancement module. After extracting point cloud features, we construct a simple yet efficient superpoint partition module to segment the point cloud into superpoints, a set of 3D points that share similar local geometric information. The generated superpoints are used to embed the structural information into the point features, thereby improving the accuracy of object boundary segmentation. In this procedure, the quality of superpoints determines the accuracy of object boundaries. In order to improve the quality of the superpoints, we also propose a point cloud class-agnostic loss to learn point features, which can perceive the local geometric structure of point clouds without using semantic information. After clustering superpoints, we introduce a geometry enhancement module that uses superpoint-point attention to encode structural information of the point cloud into point features, thereby strengthening the recognition ability of object boundaries. To verify the effectiveness of the proposed method, we conduct experiments on the point cloud salient object detection dataset (PCSOD) (Fan, Gao, and Li 2022). Extensive experiments have shown that our approach is significantly superior to other methods and has a shorter running time. We conclude the contributions as follows:

- We propose a geometry-aware 3D salient object detection network, which explicitly utilizes the structural information of point clouds to enhance the segmentation of object boundaries on point clouds.

- We develop two simple yet efficient modules, a superpoint partition module and a geometry enhancement module, for 3D salient object detection. We also present a point cloud class-agnostic loss to learn the local geometric information of point clouds without using semantic information.

- Rich experiments show that the proposed method not only achieves new state-of-the-art performance on the

PCSOD dataset, but also has the shortest inference time.

Related Works

Deep learning on point clouds. Existing deep learning based point cloud processing methods can be roughly divided into four folds: point based (Lai et al. 2022; Jiang et al. 2020; Yan et al. 2020; Hu et al. 2020), graph based (Li et al. 2021; Ding et al. 2021; Lei, Akhtar, and Mian 2020), multi-view based (Li et al. 2020; Xu et al. 2023; Chen et al. 2020; Le, Bui, and Duan 2017) and voxelization based methods (Malik et al. 2021; Meng et al. 2019; Poux and Billen 2019). Qi *et al.* (Qi et al. 2017a) introduced PointNet as a pioneering method for learning features on point clouds, which directly handles point clouds with multilayer perceptron (MLP), max-pooling, and rigid transformations to achieve extracting features of point clouds. Following PointNet, the appearance of PointNet++ (Qi et al. 2017b) enhanced the ability to characterize the local geometric structures of point clouds. Subsequently, in order to extends the convolution operation from 2D images to 3D point cloud, Li *et al.* (Li et al. 2018) proposed a transformation to simultaneously weight and permute the input features in PointCNN. The graph-based methods (Cheng et al. 2021, 2020; Shen et al. 2018) regard point cloud data as a graph structure, where points represent nodes and relationships between points represent edges. These methods utilizing techniques such as graph convolutional networks to analyze and process point clouds in object recognition and segmentation tasks. For instance, DGCNN (Wang et al. 2019) is a commonly used graph-based method, which dynamically aggregate local geometric feature of point clouds. The methods based voxelization (Choy, Gwak, and Savarese 2019; Liu et al. 2019) divides point cloud data into regular voxels in space. Therefore, we can use the voxelized data as input to 3D convolutional networks for further processing and analysis. These networks are specifically designed to handle three-dimensional data, allowing us to extract meaningful features and patterns from the voxelized representation. Applying 3D convolutions can capture spatial relationships and learn hierarchical representations that are useful for various tasks such as object recognition, segmentation, and reconstruction. However, the sparsity of point cloud data often results in empty voxels, leading to the wastage of computational resources. The last multi-view based method (Wu et al. 2019; Chen et al. 2020) projects point clouds into different view spaces and utilizes these views as input to accomplish analysis and processing of point clouds. Although the above methods have made significant progress in 3D classification, segmentation, reconstruction, and generation tasks, their applicability to the 3D salient object detection task is not particularly strong.

Point cloud salient object detection. Point cloud salient object detection refers to accurately identifying and locating salient objects from point cloud data. Similar to salient object detection in images, it often serves as a valuable preprocessing step, providing better solutions for applications such as 3D scene understanding, object recognition, and robotic navigation. However, unlike traditional image-based salient object detection, salient object detection in
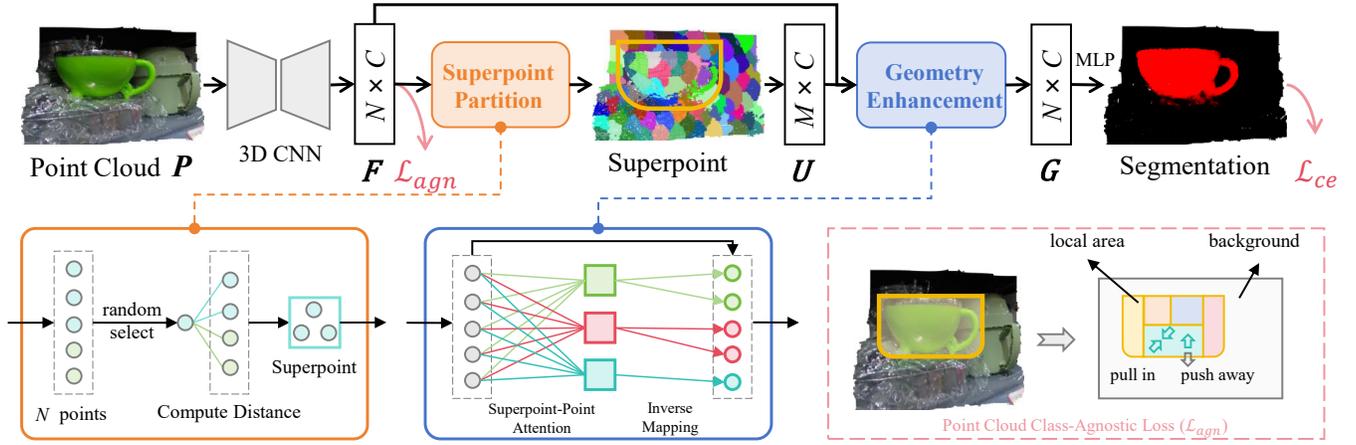
Figure 2: The pipeline of our geometry-aware 3D salient object detection network. Given a point cloud, we first use the 3D CNN backbone to extract point features. Then, we adopt the superpoint partition module to obtain superpoints. After that, we propose the geometry enhancement module to encode structural information into point clouds. In addition, we propose a point cloud class-agnostic loss $\mathcal{L}_{agn}$ to learn discriminative point features for improving superpoint quality.

point clouds requires consideration of factors such as spatial distribution, density, and shape of the points. On the other hand, point cloud data consists of a large number of irregularly distributed points, posing numerous challenges for salient object detection. In recent years, with the development of deep learning techniques, significant progress has been made in salient object detection in point clouds using neural network-based methods. In point cloud SOD, Fan *et al.* (Fan, Gao, and Li 2022) proposed PointSal is the pioneering work and it take a hand-labeled dataset PCSOD for advancing this important field. In this work, PointSal is a typical encoder-decoder architecture and design two important modules, *i.e.*, point perceptron block and saliency perception block, which local salient objects though taking full advantage of multi-scale features and global semantics. However, due to PointSal capture feature all by using MLPs, the capability of learning long-range feature representations of PointSal is very limted. Subsequently, Zhang *et al.* (Zhang et al. 2023) proposed an enhanced point feature network (EPFNet) for point cloud SOD, which take full advantage of color information available in point cloud for point cloud SOD. As the dominant frameworks Transform in natural language processing are applied to point clouds, Wei *et al.* (Wei et al. 2024) later proposed PSOD-Net, a model featuring two contextual transformer modules designed to effectively capture multi-scale contextual point information.

Superpoint representation of point cloud. Superpoints are similar to superpixels in 2D images, which refer to a collection of points within a point cloud that exhibit certain semantic and geometric similarities. Lin *et al.* (Lin et al. 2018) proposed a method that segments superpoints by utilizing locally crafted information to minimize an energy function. Guinard *et al.* (Guinard and Landrieu 2017) utilized handcrafted local descriptors to generate geometrically simple superpoints using a greedy graph-cut algorithm (Landrieu and Obozinski 2016). Landrieu (Landrieu and Boussaha 2019) proposed employing a deep network for extracting point embeddings instead of using handcrafted features to segment superpoints. Hui *et al.* (Hui et al. 2021) in-

troduced an end-to-end superpoint framework, which iteratively learns the correlation mapping between individual points and superpoints for the purpose of clustering. To handle LiDAR point clouds, Hui *et al.* (Hui et al. 2023) propose an efficient point cloud oversegmentation network by applying clustering on the range image.

Method

Architecture Overview

The overall pipeline of our geometry-aware 3D salient object detection network is shown in Fig. 2. Given a point cloud $\boldsymbol{P} \in \mathbb{R}^{N \times 6}$, where N is the number of points, each point is a 6-dimensional vector that contains 3D coordinates and RGB information. We first input the point cloud into the 3D CNN backbone (Yan et al. 2022) to extract point cloud features $\boldsymbol{F} \in \mathbb{R}^{N \times C}$, where C is the number of channels. Then, the feature $\boldsymbol{F}$ is used to calculate the point cloud class-agnostic loss for learning discriminative local geometric features without using semantic information. After that, we feed the point feature to the superpoint partition module to cluster the point cloud into M superpoints. By fusing the superpoint features $\boldsymbol{U} \in \mathbb{R}^{M \times C}$ into point features $\boldsymbol{G} \in \mathbb{R}^{N \times C}$, we propose a geometry enhancement module to embed the structural information of the point cloud. Finally, we predict the mask of the object from the enhanced point features.

Superpoint Partition

Inspired by the human vision system, we consider introducing structural information of point clouds to improve the performance of 3D salient object detection. Specifically, we introduce the superpoint representation of the point cloud and propose a superpoint partition module to generate superpoints. The quality of superpoints determines the quality of downstream object segmentation. Therefore, before clustering points into superpoints, we propose a point cloud class-agnostic loss to help improve the quality of superpoints. Therefore, in this subsection, we only introduce the

Algorithm 1: Superpoint Generation Algorithm

Input: Unclustered point set U, Queue size K,
       Distance threshold γ
Output: Superpoint sets S

1 **while** $U \neq \emptyset$ **do**
2   Randomly select a point $i \in U$ as cluster center;
3   Initialize queue Q as empty;
4   **for** *each point* $j \in U \setminus \{i\}$ **do**
5   Compute Euclidean distance d_{ij} ;
6   Enqueue (j, d_{ij}) into Q;
7   Sort Q by distance and keep the K nearest points;
8   **while** Q *is not empty* **do**
9   Extract point j from Q;
10   Compute feature distance $d(i,j)$;
11   **if** $d \leq \gamma$ **then**
12   Add j to the superpoint of i;
13   Remove j from U;
14   **else**
15   Clear queue Q;
16   **break**;

17 **return** S;

---

proposed simple yet effective superpoint partition approach strategy.

Similar to the concept of superpixels in 2D images, the superpoint is a set of 3D points that share similar local geometric information. To handle a large number of points, we adopt a very simple yet effective superpoint generation strategy. The core idea is to consider the similarity of point cloud features based on distance-based clustering, *i.e.*, if two points have high similarity, they belong to the same superpoint. The superpoint generation algorithm is shown in Algorithm 1. Note that in the experiment, we found that even without updating the feature of the cluster center, we can still generate good superpoints.

Geometry Enhancement

After obtaining the superpoints, we use a superpoint-point attention mechanism to encode the structural information of the superpoint into the point features.

Given the point feature $\boldsymbol{F} \in \mathbb{R}^{N \times C}$, we first obtain the superpoint feature $\boldsymbol{U} \in \mathbb{R}^{M \times C}$ by averaging the point features belonging to the same superpoint. It is worth noting that the averaging operation aggregates the local geometric information of the point cloud. Thus, the superpoint feature embeds the 3D structural information. Then, we design a superpoint-point attention mechanism to learn the correlation between the superpoints and points, which is formulated as:

$$\boldsymbol{U}^{'} = \text{CrossAttention}(\boldsymbol{U}, \boldsymbol{F}, \boldsymbol{F}) \quad (1)$$

where "Query" is the superpoint feature and "Key" and "Value" are point features. In this way, we fuse the fine-level point features into superpoint features. To encode superpoint features into point features, we directly inverse map superpoint features to point features based on the indexes between

the points and superpoints, which is given by:

$$\boldsymbol{G} = \text{Inv}(\boldsymbol{U}^{'}) + \boldsymbol{F} \quad (2)$$

where $\text{Inv}(\cdot)$ is the inverse mapping function that maps superpoint features into point features. In addition, $\boldsymbol{G} \in \mathbb{R}^{N \times C}$ is the obtained new point features. Finally, we directly use the point-level feature $\boldsymbol{G}$ to predict the category of each point, whether it belongs to the object or background.

Point Cloud Class-Agnostic Loss

Due to the fact that the superpoint quality is crucial for encoding geometric information into point clouds, we propose a point cloud class-agnostic loss for improving superpoint quality. In the 3D salient object detection task, each point only has a binary mask, so we cannot effectively learn semantic information from the binary mask of the point cloud. Thus, it is not possible to use semantic information to learn discriminative point features for clustering superpoints.

Formally, the core idea of our point cloud class-agnostic loss is to consider the local area of the point cloud, rather than the overall point cloud. Compared with the whole point cloud, local areas usually have similar geometric structures. Therefore, we can utilize the similarity of local geometric structures in point clouds to constrain the similarity of local point cloud features. Specifically, given the ground truth mask of a point cloud, we first generate local areas within the object. Then, we constrain the points that are located in the same local area to have similar features to each other. Assuming that we have the i-th local area $\mathcal{N}_i$ within the object, we force the point $j \in \mathcal{N}_i$ close to the meaning embedding of the local area, *i.e.*, we minimizing the point feature to the meaning embedding $\boldsymbol{y}_i$ of the local area, which is given by:

$$\mathcal{L}_{pull} = \frac{1}{Z} \sum_{i=1}^{Z} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} [||\boldsymbol{f}_j - \boldsymbol{y}_i||_2 - \alpha]_+^2 \quad (3)$$

where $\boldsymbol{y}_i$ is obtained by averaging the point features. Z is the number of local areas. We also push the point features within the object away from the background, which is written as:

$$\mathcal{L}_{push} = \frac{1}{Z} \sum_{i=1}^{Z} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} [2\beta - ||\boldsymbol{f}_j - \boldsymbol{b}||_2]_+^2 \quad (4)$$

where $\boldsymbol{b} \in \mathbb{R}^C$ is the background feature, which is computed by averaging the background point features. The point cloud class-agnostic loss is given by $\mathcal{L}_{agn} = \mathcal{L}_{pull} + \mathcal{L}_{push}$. Note that we empirically set $\alpha = 0.01$ and $\beta = 0.2$. Finally, the final loss function for training the network is written as $\mathcal{L}_{final} = \mathcal{L}_{ce} + \mathcal{L}_{agn}$, where $\mathcal{L}_{ce}$ is the cross-entropy for category prediction of each point.

Experiments

Experimental Setup

Implementation details. Our method is implemented with the PyTorch on an NVIDIA RTX 3090 Ti GPU. For the model, the input dims of points is 9 channels, which consist of spatial coordinates, RGB information, and normalized spatial coordinates. Since the increase in point count doesn't significantly affect our experimental speed, we opt
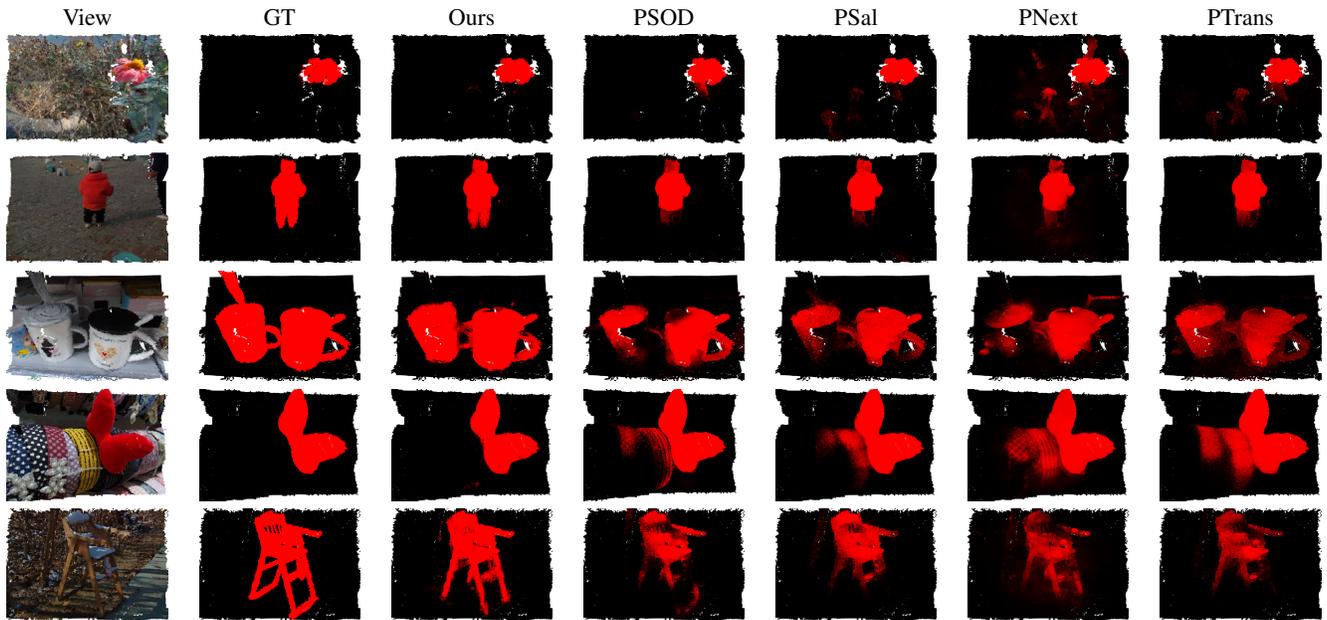
Figure 3: Visualization results of five methods under different views in the test set of the PCSOD dataset. Note that "GT","PSOD", "PSal", "PNext", and "PTrans" represent the ground truth, PSOD-Net (Wei et al. 2024), PointSal (Fan, Gao, and Li 2022), PointNeXt (Qian et al. 2022), Point Transformer (Zhao et al. 2021), respectively.

to utilize all 240,000 points for both the training and testing phases. For the training process, our model is trained using the Adam optimizer for a total of 300 epochs, taking approximately 13 hours. This represents a significant time saving compared to PointSal (Fan, Gao, and Li 2022) and PSOD-Net (Wei et al. 2024), which requires training for 3000 epochs (taking roughly 3 days) and 800 epochs(taking roughly 40 hours), respectively. In our training process, in addition to using cross-entropy loss, we also introduced a point cloud class-agnostic loss to facilitate better feature learning. The initial learning rate is 1e-3 and weight decay is 1e-4. Additionally, data augmentation is applied using transformation with standard deviations of [3, 3, 3]. The parameter spatial shape is set to [150, 100, 75], indicating the shape of the spatial grid with dimensions 150, 100, and 75 along the x, y, and z axes, respectively.

Dataset. Following previous methods (Fan, Gao, and Li 2022; Wei et al. 2024; Zhang et al. 2023), we adopt the point cloud salient object detection dataset (PCSOD) (Fan, Gao, and Li 2022) as the benchmark to conduct experiments. PCSOD is a challenging dataset, which has 53.4% difficult samples. It contains more than 100 daily life scenes, such as signs, daily necessities, flowers, and trees. There are a total of 2,873 scenarios in this dataset, including simple targets, multiple targets, small targets, and some targets with complex structures. According to the official partition, there are 2,000 samples for training and 872 samples for evaluation. Note that each point is described as a six-dimensional vector, which contains 3D coordinate information (x, y, z) and 3D color information (r, g, b).

Evaluation metrics. To assess and compare the effectiveness of various methods, we adopt the same met-

rics as in PointSal (Fan, Gao, and Li 2022) for performance evaluation, including mean absolute error (MAE), F-measure (Margolin, Zelnik-Manor, and Tal 2014), E-measure (Fan et al. 2018), and intersection over union (IoU). The MAE estimates the distance between the point-by-point predicted value and the corresponding true value, which is calculated as MAE $= \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$, where y_i is the true value and $\hat{y}_i$ is the predicted value for each point. F-measure is a weighted harmonic average of precision and recall, which is formulated as F-measure $= (1 - \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$. For a fair comparison, we follow (Fan, Gao, and Li 2022) and set β=0.3 for controlling the weighting relationship between precision and recall. E-measure provides a comprehensive evaluation metric for assessing the performance of saliency segmentation models. It simultaneously considers both local details and global structure, thus offering a more holistic assessment of model effectiveness. IoU calculates the overlap ratio of predict and the truth, $i.e.$, the ratio of their intersection and union. For both F-measure and E-measure, we use stochastic thresholds to calculate both values to ensure that our model show excellent results at different threshold distributions.

Results

We conduct a comparative analysis of performance by comparing our method with three state-of-the-art point cloud salient object detection methods, including PointSal (Fan, Gao, and Li 2022), EPFNet (Zhang et al. 2023) and PSOD-Net (Wei et al. 2024). We also compare with two point cloud segmentation methods, including PointNeXt (Qian et al. 2022) and Point Transformer (Zhao et al. 2021).

Data: [xyz, rgb]				
Methods	MAE↓	F-measure↑	F-measure↑	IoU↑
Point Transformer	0.075	0.762	0.848	0.670
PointNeXt	0.066	0.779	0.859	0.680
PointSal	0.069	0.769	0.851	0.656
PSOD-Net	0.058	0.805	0.878	0.711
EPFNet	0.047	0.820	0.898	0.727
3DGAS (ours)	**0.042**	**0.848**	**0.912**	**0.763**
Data: xyz				
Point Transformer*	0.108	0.643	0.756	0.513
PointNeXt*	0.092	0.652	0.752	0.501
PointSal*	0.090	0.693	0.797	0.565
PSOD-Net*	0.087	0.698	0.816	0.587
3DGAS (ours)	**0.050**	**0.814**	**0.895**	**0.720**

Table 1: Comparison results of different models on the test set of the PCSOD dataset. "[xyz, rgb]" means that the color information is concatenated with the 3D coordinates. Please note that EPFNet is a multi-modal method, which uses 2D image information. "*" means that the results are obtained by running the official codes. The best results are highlighted in **bold**.
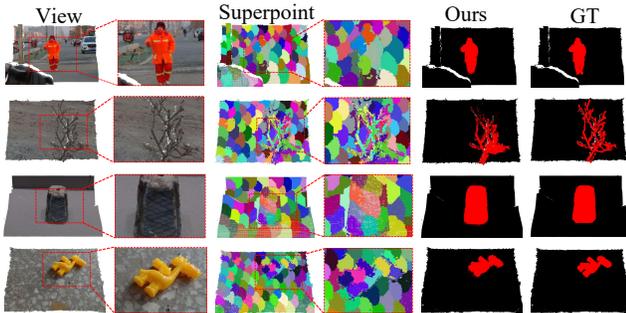

Figure 4: Visualization results of the generated superpoints by our method in the test set of the PCSOD dataset. Please note that superpoints are randomly colored.

Quantitative comparison. We list the result of the quantitative comparison in Table 1. Note that the PCSOD dataset contains 6-dimensional information ($xyz + rgb$). When using 3D coordinate plus color information, it can be observed that our method 3DGAS outperforms all compared methods on all four evaluation metrics. Compared with the previous state-of-the-art method EPFNet (Zhang et al. 2023), our 3DGAS surpasses it by 0.005 in MAE, 0.028 in F-measure, 0.014 in E-measure, and 0.036 in IoU. It is worth noting that except for EPFNet, all methods (Point Transformer (Zhao et al. 2021), PointNeXt (Qian et al. 2022), PointSal (Fan, Gao, and Li 2022), PSOD-Net (Wei et al. 2024)) concatenate 3D xyz information with rgb information as input. However, EPFNet is a multi-modal method that extracts color information from the point cloud as a RGB image, and uses an image network to learn features. Therefore, RGB images play an important role in the performance of EPFNet. Although our method is based on a single modal, our method can still achieve the highest per-

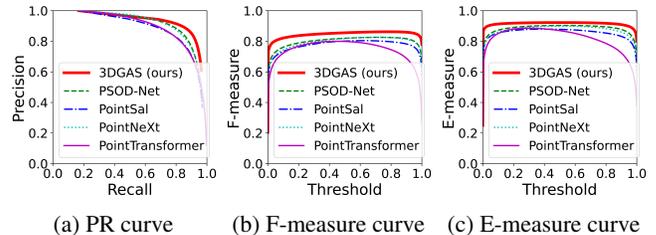
(a) PR curve　(b) F-measure curve　(c) E-measure curve

Figure 5: Precision-recall (PR), F-measure, and E-measure curves of different methods on the test set of the PCSOD dataset.

formance, which further demonstrates the effectiveness of the proposed method.

To compare the learning of point cloud structures of different methods in 3D salient object detection, we conduct experiments only using the 3D coordinate information as input and discarding color information. Since the codes of EPFNet have not been released, for a fair comparison, we conduct experiments by running the released codes of Point-Transformer, PointNext, PointSal, and PSOD-Net for experiments. The quantitative results are shown in Table 1. According to the results, our method is superior to other methods without using rgb information. Even compared with the methods that uses additional color information, our method still surpasses Point Transformer, PointNext, PointSal, and PSOD-Net. The comparison results indicate that our method can better learn the structural information of objects from the point clouds.

To present the results more intuitively, in Fig. 5, we plot the precision-recall (PR) curves, F-measure curves, and E-measure curves under different thresholds of different methods, respectively. From Fig. 5(a), it can be found that the PR curve of our 3DGAS is higher than other methods at different thresholds, which shows that the performance of our method is better than others. According to Fig. 5(b) and (c), it can be observed that the proposed 3DGAS consistently outperforms other methods in terms of both F-measure and E-measure at any given thresholds. Additionally, the curves of our method are smoother compared to others, indicating that our model exhibits more stable performance.

Visualization. To further illustrate the effectiveness of our proposed method, we present visual results of predictions on challenging scenes for different methods. As shown in Fig. 3, the first row depicts small object detection, the second row shows the salient object detection results in simple scenes, the third row shows multi-object detection, the fourth row illustrates salient object detection in complex backgrounds, and the fifth row shows the complex structure. From the results in Fig. 3, it can be observed that our method not only outperforms other methods in terms of localization and object segmentation integrity across various scenarios, but also exhibits favorable characteristics in edge handling. It rarely mistakes the surrounding non-target background as part of the recognition.

To reveal the contribution of the superpoints used in our method. In Fig. 4, we visualize the generated superpoints by

Methods	MAE↓	F-measure↑	E-measure↑	IoU↑
Baseline	0.116	0.632	0.768	0.519
+ SP	0.050	0.810	0.881	0.720
+ SP + GE	0.047	0.832	0.903	0.738
+ SP + GE + CA	**0.042**	**0.848**	**0.912**	**0.763**

Table 2: The results of ablation study on the test set of different modules. "SP", "GE", and "CA" indicate the superpoint partition, geometry enhancement, and point cloud class-agnostic loss, respectively.

Methods	N=240,000	N/2=120,000	N/4=60,000
PointSal	0.069	0.073	0.078
PSOD-Net	0.058	0.062	0.086
3DGAS (ours)	**0.042**	**0.045**	**0.049**

Table 3: MAE of different PSCOD methods under varying point cloud densities

our method in the test set of the PCSOD dataset. By comparing the visualization of view and superpoints, which are zoomed in on the second and fourth columns, it can be observed that the salient objects can be segmented into individual superpoints. For complex scenes, we can obtain good superpoints along the geometric structure of the object, such as the tree branches in the second row of Fig. 4. By introducing the structural information of point clouds through superpoints, the accuracy of boundary segmentation will be improved.

We also measure the computation costs of different methods in terms of parameters, FLOPs, and inference time. For a fair comparison, all methods are evaluated under the same experimental environment. The results are as follows: 7.1M/7.6G/0.22s (PointNeXt), 7.8M/2.8G/0.31s (Point-Transformer), 4.8M/**1.4G**/1.6s (PointSal), 8.2M/4.1G/0.8s (PSOD-Net), and **0.2M**/2.6G/**0.06s** (ours). Our method achieves the lowest number of parameters due to the use of sparse convolution and lower feature channels. Although our method has higher FLOPs than PointSal, it significantly outperforms other methods in terms of performance. Additionally, by reducing the number of point clouds through voxelization, our model achieves the fastest inference time. In summary, compared to existing methods designed for 3D salient detection, our method offers faster speed and higher performance.

To evaluate the performance of point cloud salient object detection models under varying point cloud densities, we conducted a comparative experiment. Specifically, we tested PointSal, PSOD-Net, and ours on three different point cloud sizes: the original size N, half the original size N/2, and one-quarter the original size N/4. The results of MAE are present in Tabel 3. The results indicate that our method is more robust to changes in point cloud density, making it better suited for handling point cloud data at varying resolutions in real-world applications.

Ablation Study

To analyze the effectiveness of each proposed module, we conduct comprehensive ablation study experiments on the PCSOD dataset.

Effect of superpoint partition. To analyze the effectiveness of the proposed superpoint partition, we conduct the ablation study experiment by comparing the baseline model. As shown in Table 2, when comparing the first row (baseline model) and second row, it can be observed that the

proposed superpoint partition (dubbed as "SP") can further improve the baseline performance. **Effectiveness of geometry enhancement.** Compared with previous methods, our method introduces the structural information of the point cloud through superpoints. We plus the geometry enhancement module on the learned features to conduct an ablation experiment. From Table 2, we can observe that using the geometry enhancement module (dubbed as "GE") can efficiently improve the performance. Due to the fact that the superpoint is a set of points that share similar local geometric information, it contributes to distinguishing the boundaries between the background and the object.

Point cloud class-agnostic loss analysis. In order to generate high-quality superpoints, we propose the point cloud class-agnostic loss to learn discriminative point features for clustering points. By comparing the third row and fourth row of Table 2, it can be found that using point cloud class-agnostic loss (dubbed as "CA") can significantly improve the performance of salient object recognition. Since it can enhance the discriminative ability of point cloud features on the local geometric structure of the salient object, the quality of generated superpoints is better.

Conclusion

In this paper, we present a geometry-aware 3D salient object detection network for point cloud salient object detection. To enhance the accuracy of object boundaries, we explicitly leverage the structural information of points by constructing superpoints. Specifically, after extracting point cloud features, we construct a simple yet efficient superpoint partition module to segment the point cloud into superpoints. The generated superpoints are used to embed the structural information into the point features, thereby improving the accuracy of object boundary segmentation. In order to ensure high-quality superpoints, we also propose a point cloud class-agnostic loss to learn discriminative point features for clustering points into superpoints. Extensive experiments demonstrate that the proposed method achieves new state-of-the-art performance and cost the shortest inference time.

Acknowledgments

References

Chen, H.; and Li, Y. 2018. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3051–3060.

Chen, X.; Zhang, C.; Lin, G.; and Han, J. 2020. Compositional prototype network with multi-view comparison for few-shot point cloud semantic segmentation. *arXiv preprint arXiv:2012.14255*.

Cheng, M.; Hui, L.; Xie, J.; and Yang, J. 2021. SSPC-Net: Semi-supervised semantic 3D point cloud segmentation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

Cheng, M.; Hui, L.; Xie, J.; Yang, J.; and Kong, H. 2020. Cascaded non-local neural network for point cloud semantic segmentation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal Convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ding, Y.; Zhao, X.; Zhang, Z.; Cai, W.; and Yang, N. 2021. Graph sample and aggregate-attention network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, D.-P.; Wang, W.; Cheng, M.-M.; and Shen, J. 2019. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8554–8564.

Fan, S.; Gao, W.; and Li, G. 2022. Salient object detection for point clouds. In *European Conference on Computer Vision*.

Guinard, S.; and Landrieu, L. 2017. Weakly supervised segmentation-aided classification of urban scenes from 3D LiDAR point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42: 151–157.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3203–3212.

Hu, Z.; Zhen, M.; Bai, X.; Fu, H.; and Tai, C.-l. 2020. Jsenet: Joint semantic segmentation and edge detection network for 3D point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*.

Huang, J.; Yan, W.; Li, T.; Liu, S.; and Li, G. 2020. Learning the global descriptor for 3-D object recognition based on multiple views decomposition. *IEEE Transactions on Multimedia*, 24: 188–201.

Huang, R.; Xing, Y.; and Zou, Y. 2020. Triple-complementary network for RGB-D salient object detection. *IEEE Signal Processing Letters*, 27: 775–779.

Hui, L.; Tang, L.; Dai, Y.; Xie, J.; and Yang, J. 2023. Efficient LiDAR point cloud oversegmentation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18003–18012.

Hui, L.; Yuan, J.; Cheng, M.; Xie, J.; Zhang, X.; and Yang, J. 2021. Superpoint network for point cloud oversegmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5510–5519.

Jiang, L.; Zhao, H.; Shi, S.; Liu, S.; Fu, C.-W.; and Jia, J. 2020. PointGroup: Dual-set point grouping for 3D instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, 4867–4876.

Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; and Jia, J. 2022. Stratified Transformer for 3D point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Landrieu, L.; and Boussaha, M. 2019. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7440–7449.

Landrieu, L.; and Obozinski, G. 2016. Cut Pursuit: fast algorithms to learn piecewise constant functions. In *Artificial Intelligence and Statistics*, 1384–1393.

Le, T.; Bui, G.; and Duan, Y. 2017. A multi-view recurrent neural network for 3D mesh segmentation. *Computers & Graphics*, 66: 103–112.

Lei, H.; Akhtar, N.; and Mian, A. 2020. SegGCN: Efficient 3D point cloud segmentation with fuzzy spherical kernel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, L.; Zhu, S.; Fu, H.; Tan, P.; and Tai, C.-L. 2020. End-to-End learning local multi-view descriptors for 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, R.; Zhang, Y.; Niu, D.; Yang, G.; Zafar, N.; Zhang, C.; and Zhao, X. 2021. PointVGG: Graph convolutional network with progressive aggregating features on point clouds. *Neurocomputing*, 429: 187–198.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. PointCNN: Convolution on X-Transformed points. *Advances in Neural Information Processing Systems*, 31.

Lin, Y.; Wang, C.; Zhai, D.; Li, W.; and Li, J. 2018. Toward better boundary preserved supervoxel segmentation for 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 143: 39–47.

Liu, J.-J.; Hou, Q.; and Cheng, M.-M. 2020. Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *IEEE Transactions on Image Processing*, 29: 8652–8667.

Liu, N.; and Han, J. 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 678–686.

Liu, N.; Zhang, N.; Shao, L.; and Han, J. 2021. Learning selective mutual attention and contrast for RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9026–9042.

Liu, Z.; Tang, H.; Lin, Y.; and Han, S. 2019. Point-Voxel CNN for efficient 3D deep learning. *Advances in Neural Information Processing Systems*, 32.

Mahadevan, V.; and Vasconcelos, N. 2012. Biologically inspired object tracking using center-surround saliency mechanisms. *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, 35(3): 541–554.

Malik, J.; Shimada, S.; Elhayek, A.; Ali, S. A.; Theobalt, C.; Golyanik, V.; and Stricker, D. 2021. HandVoxNet++: 3D hand shape and pose estimation using voxel-based neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8962–8974.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Meng, H.-Y.; Gao, L.; Lai, Y.-K.; and Manocha, D. 2019. VV-Net: Voxel vae net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Poux, F.; and Billen, R. 2019. Voxel-based 3D point cloud semantic segmentation: Unsupervised geometric and relationship featuring vs deep learning methods. *ISPRS International Journal of Geo-Information*, 8(5): 213.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; and Ghanem, B. 2022. Pointnext: Revisiting PointNet++ with Improved Training and Scaling Strategies. *Advances in Neural Information Processing Systems*, 23192–23204.

Qu, L.; He, S.; Zhang, J.; Tian, J.; Tang, Y.; and Yang, Q. 2017. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5): 2274–2285.

Shen, Y.; Feng, C.; Yang, Y.; and Tian, D. 2018. Mining point cloud local structures by kernel correlation and graph pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shi, W.; Xu, J.; Zhu, D.; Zhang, G.; Wang, X.; Li, J.; and Zhang, X. 2021. RGB-D semantic segmentation and label-oriented voxelgrid fusion for accurate 3D semantic mapping. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 183–197.

Tu, Z.; Li, Z.; Li, C.; Lang, Y.; and Tang, J. 2021. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing*, 30: 5678–5691.

Wang, A. 2020. Three-stream cross-modal feature aggregation network for light field salient object detection. *IEEE Signal Processing Letters*, 28: 46–50.

Wang, G.; Li, C.; Ma, Y.; Zheng, A.; Tang, J.; and Luo, B. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*, 359–369.

Wang, J.; Song, K.; Bao, Y.; Huang, L.; and Yan, Y. 2021. CGFNet: Cross-guided fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2949–2961.

Wang, L.; Wang, L.; Lu, H.; Zhang, P.; and Ruan, X. 2016. Saliency detection with recurrent fully convolutional networks. In *European Conference on Computer Vision*, 825–841.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5): 1–12.

Wei, Z.; Chen, B.; Wang, W.; Chen, H.; Wei, M.; and Li, J. 2024. Point Transformer-based Salient Object Detection Network for 3D Measurement Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing*.

Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; and Keutzer, K. 2019. SqueezeSegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation*.

Xu, W.; Li, X.; Ni, P.; Guang, X.; Luo, H.; and Zhao, X. 2023. Multi-view Fusion Driven 3D Point Cloud Semantic Segmentation Based on Hierarchical Transformer. *IEEE Sensors Journal*.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2DPASS: 2D priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, 677–695.

Yan, X.; Zheng, C.; Li, Z.; Wang, S.; and Cui, S. 2020. PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5589–5598.

Zhang, M.; Ren, W.; Piao, Y.; Rong, Z.; and Lu, H. 2020. Select, supplement and focus for RGB-D saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3472–3481.

Zhang, Z.; Gao, P.; Peng, S.; Duan, C.; and Zhang, P. 2023. Enhanced point feature network for point cloud salient object detection. *IEEE Signal Processing Letters*.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhou, W.; Chen, Y.; Liu, C.; and Yu, L. 2020. GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images. *IEEE Signal Processing Letters*, 27: 800–804.