# Point Transformer-based Salient Object Detection Network for 3D Measurement Point Clouds

Zeyong Wei, Baian Chen, Weiming Wang, Honghua Chen, Mingqiang Wei, *Senior Member, IEEE* and Jonathan Li, *Fellow, IEEE*

***Abstract*—While salient object detection (SOD) on 2D images has been extensively studied, there is very little SOD work on 3D measurement surfaces. We propose an effective point transformer-based SOD network for 3D measurement point clouds, termed PSOD-Net. PSOD-Net is an encoder-decoder network that takes full advantage of transformers to model the contextual information in both multi-scale point- and scene-wise manners. In the encoder, we develop a Point Context Transformer (PCT) module to capture region contextual features at the point level; PCT contains two different transformers to excavate the relationship among points. In the decoder, we develop a Scene Context Transformer (SCT) module to learn context representations at the scene level; SCT contains both Upsampling-and-Transformer blocks and Multi-context Aggregation units to integrate the global semantic and multi-level features from the encoder into the global scene context. Experiments show clear improvements of PSOD-Net over its competitors and validate that PSOD-Net is more robust to challenging cases such as small objects, multiple objects, and objects with complex structures. Code is available at: https://github.com/ZeyongWei/PSOD-Net.**

***Index Terms*—PSOD-Net, 3D salient object detection, point transformer, 3D measurement point cloud**

## I. Introduction

Salient objects are the most attractive objects in contrast to their surroundings in the scene [1], [2]. Salient object detection (SOD) has a wide range of applications and can provide pre-processing results for various vision tasks, including 3D shape classification [3], compression [4], quality assessment [5], and many others. Distinct from the relevant tasks, i.e., saliency detection [6], [7] and object detection [8]–[10], SOD requires locating and completely segmenting salient objects, hence being more challenging.

While SOD for 2D images has been extensively studied [11], [12], there are very few efforts on SOD for 3D point clouds. This is even though the rapid development of 3D acquisition technologies has significantly simplified geometric modeling [13], and 3D point clouds become more and more popular with wide applications of autonomous driving, and Metaverse [14], [15].

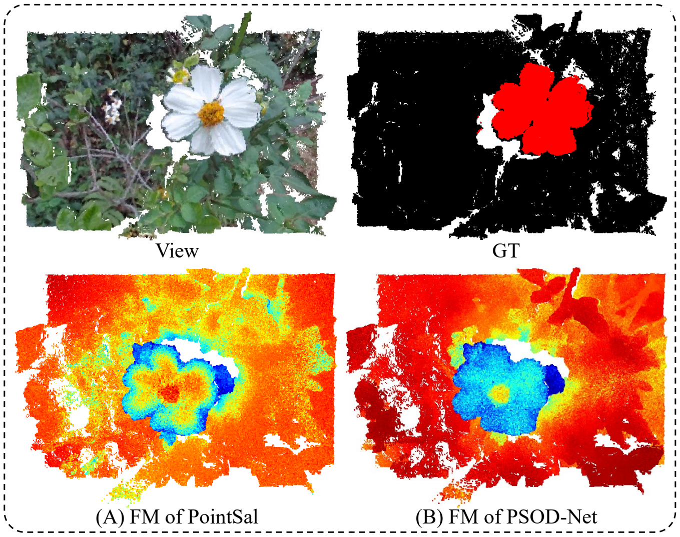


Fig. 1: 3D heatmap visualization of feature maps (FM). Points that belong to the same semantic part share more similar features by (B) PSOD-Net than (A) PointSal [16].

In contrast to images where salient objects remain unchanged, point clouds can easily undergo 3D view rotation. This implies that an object, during view rotation, may transition from being salient to becoming a non-salient object. PointSal [16] is the pioneering work of point cloud salient object detection (PCSOD). It provides a novel dataset, namely PCSOD, for point cloud salient object detection. To address the ambiguity caused by the aforementioned viewpoint changes, PCSOD divides the entire scene into different views from different perspectives, with salient objects in each view fixed. By combining the salient objects from the “given views”, we obtain a complete description of the salient objects for scenes in point clouds. Notably, PointSal takes full advantage of multi-scale features and global semantics to locate salient objects. However, all feature extraction modules of PointSal are implemented by multi-layer perceptrons (MLPs), which seriously limits the capability of learning long-range feature representations due to fixed receptive fields. When dealing with an object with complex structures (see Figure 1 (A)), PointSal [16] fails to capture the complete structure information.

To deal with the challenges of small/multiple objects and objects with complex structures, we propose a point transformer model for PCSOD, dubbed PSOD-Net. Considering

Z. Wei, B. Chen, H. Chen and M. Wei are with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, and also with the Shenzhen Institute of Research, Nanjing University of Aeronautics and Astronautics, Shenzhen, China (e-mail: weizeyong1@gmail.com; sx2116068@nuaa.edu.cn; chenhonghuacn@gmail.com; mingqiang.wei@gmail.com).

W. Wang is with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR (e-mail: wmwang@hkmu.edu.hk).

J. Li is with the Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada (e-mail: junli@uwaterloo.ca).

Corresponding authors: B. Chen and W. Wang.

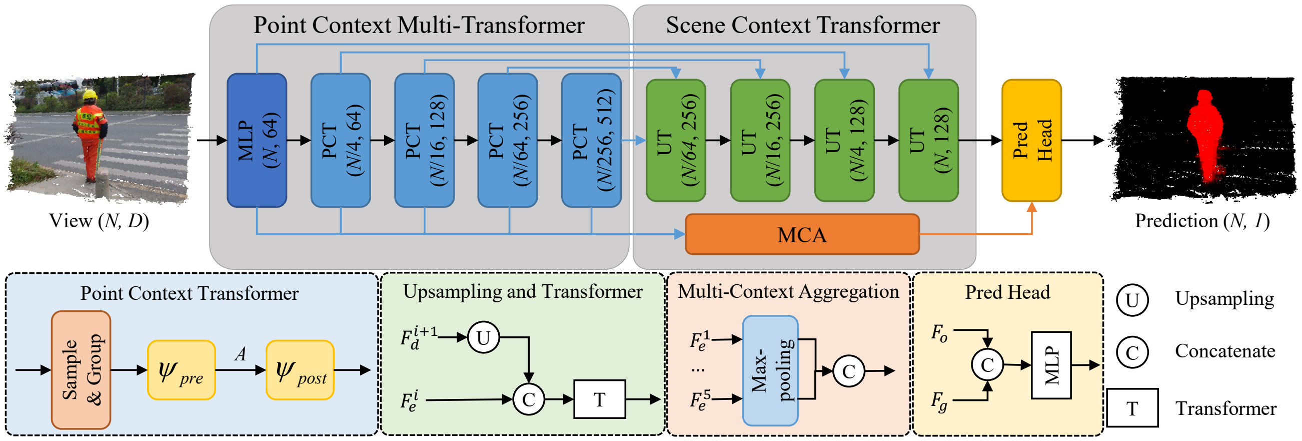


Fig. 2: Pipeline of PSOD-Net. PSOD-Net follows an encoder-decoder structure and mainly consists of a Point Context Transformer (PCT) and a Scene Context Transformer (SCT). The input measurement point cloud is first fed into the encoder to progressively capture multi-level semantic features by multiple continuous PCT modules. In PCT, two transformers ($\psi_{pre}$ and $\psi_{post}$) learn the point-wise context information in the local and global region. The decoder employs SCT to recover the resolution of the input point cloud by the Upsampling-and-Transformer (UT) block and integrate the scene contextual information via a Multi-Context Aggregation (MCA) module.

that transformer has been proven remarkably effective at the level of detailed object analysis and large-scale parsing of massive scenes [17], we improve the ability to detect salient objects by designing two types of transformer-based modules. The first type is the Point Context Transformer (PCT) and the second is the Scene Context Transformer (SCT). They can capture multi-scale point-wise and scene-wise contextual information simultaneously.

PCT can model point-wise relationships in the local and global regions by two different transformers. The captured local point-wise context information can describe the clear boundary structure, and the global point-wise context information assists in distinguishing the points belonging to different objects and classes. This ensures that our model can accurately detect multiple objects with complex structures. Besides, multiple continuous PCT modules in the encoder are employed to progressively capture multi-scale semantic features, which encourages our network to identify objects at different scales. SCT integrates the multi-scale features from the encoder to learn the global scene context information. The global scene context information is beneficial for learning the relationship between the salient objects and the background and locating the salient object from the interfering background.

Benefiting from PCT and SCT, PSOD-Net can effectively model the long-range dependencies and learn the contextual information to obtain more accurate detection results. As shown in Figure 1(B), it can be observed that the structural features of the 3D flower are completely distinguished from the complex backgrounds via PSOD-Net. Moreover, to verify the effectiveness of PSOD-Net, we compare it with PointSal [16] and several representative segmentation models on the PCSOD benchmark dataset and achieve the best performance. Besides, we successfully apply our method to LiDAR data with different structures.

Our main contributions are three-fold:

- We propose a point transformer model for 3D salient object detection (PSOD-Net). To the best of our knowledge, we are the first to try using the transformer to solve the salient object detection problem for 3D point clouds.
- We devise two types of transformers, i.e., Point Context Transformer and Scene Context Transformer, capturing contextual information at the point-and-scene levels. It successfully addresses the challenges of small objects, multiple objects, and objects with complex structures.
- Experiments verify that PSOD-Net outperforms PointSal and several representative segmentation models on the PCSOD dataset. Besides, our method is successfully applied to LiDAR data with different structures.

## II. Related Works

### A. Salient Object Detection

Early works [18]–[21] capture low-level cues by handcrafted features. Due to the lack of global semantic information, these methods cannot achieve satisfactory performance. Thanks to the emergence of convolution neural networks (CNNs) which have the powerful capability of feature representations, the CNNs-based models overcome the defects of traditional methods and make great progress. Hou et al. [22] design a skip-layer structure by the short connections, which fuses the multi-level and multi-scale features to provide advanced representations. Siris et al. [23] design a scene context-aware network to capture the global semantic information to assist in locating the salient objects. Although RGB image-based methods [24]–[26] have achieved great performance, they are restricted to understand complex scenes due to the lack of spatial geometric information. Subsequently, efforts are extended to depth images [27]–[31] for the task of RGB SOD, where depth cues can assist in recognizing the most attractive objects. Fu et al. [30] design a Siamese network to simultaneously extract RGB and depth features and fuse them in a cross-level and cross-modal manner. Zhang et al. [31] utilize the fine edge cues to locate salient objects by a complementary fusion network.

Regarding the saliency study on point clouds, early works [3], [7], [32], [33] attempt to compute the attention distribution by simulating the human visual system. However, these methods only generate a heatmap to describe the salient regions while not completely segmenting the salient objects. Fan et al. [16] first dive into studying the feasibility of point clouds salient object detection (PCSOD), and propose a benchmark dataset and a baseline model for PCSOD.

### B. Transformers for Computer Vision

Transformers have achieved great progress in the natural language processing task, which attracts the attention of the computer vision field. After the emergency of the vision Transformer [34], many works have introduced the transformer-based structure in the image understanding task [35]–[41]. Alexey et al. [41] take image patches as a token to train the transformer by a large amount of data for the image classification task. Further, Wang et al. [38] design a pyramid transformer network to capture the multi-scale feature representations and effectively decrease the cost of computation and memory by progressive reduction attention. Carion et al. [35] propose an end-to-end transformer detector that directly generates the bounding boxes from the CNN features by a transformer encoder-decoder. Lately, several Transformer-based fusion models have been propose for other types of image fusion tasks [42]–[45]. In [43], Zhang et al. [43] present a Transformer-based pan-sharpening method for redundancy reduction and global information exploitation. Tang et al. [44] propose a multiscale adaptive Transformer for multimodal medical image fusion in which an adaptive convolution and an adaptive Transformer are designed for global feature extraction. Later, Tang et al. [45] present YDTR to achieve improved performance, which is a Y-shape dynamic transformer-based network.

In the 3D vision field, transformers have also been employed in various tasks [17], [46]–[52]. Zhao et al. [17] extract the local features by a transformer block which applies self-attention in the local region of each input point. Guo et al. [46] propose an offset-attention module to calculate the offset between the input features and the attention features to optimize the original self-attention. Misra et al. [48] propose a 3D end-to-end detector built on top of [35], where a transformer encoder directly extracts features on the point cloud and a transformer decoder predicts the bounding boxes. Mao et al. [50] introduce a Transformer-based framework to capture long-range relationships between voxels. Lately, Zhang et al. [52] propose an enhanced point feature network (EPFNet) for point cloud salient object detection by aggregating image features with the point cloud.

## III. Methodology

### A. Overview

We propose a novel transformer-based salient object detector for point clouds. PSOD-Net models the context-dependent feature representations in both point and scene levels via Point Context Transformer (PCT) and Scene Context Transformer (SCT). It uses an encoder-decoder structure (see Figure 2).

Formally, given a fixed view $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ includes $N$ points with original features (e.g., $xyz$ location and RGB colors), where $v \in R^{d_{in}}$. The encoder first extracts multi-level features $\{F_e^l\}_{l=1}^5$ from the raw point clouds $\mathcal{V}$. The first-level features are extracted by multi-layer perceptron (MLP) layers and the feature dimension is fixed to 64 for subsequent transformers. The remaining features are extracted by multiple continuous PCT modules, the $l^{th}$ level features $F_e^l$ have $N_l = \frac{N}{4^{l-1}}$ aggregated points with double the feature dimension compared to the $F_e^{l-1}$ (except the second level features have the same feature dimension with the first level features).

To obtain the probabilities $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ of each point in the raw point cloud that belongs to salient points, the encoded features $\{F_e^l\}_{l=1}^5$ are fed to the decoder to progressively recover the resolution to the original size $N$. Each UT module takes the output from the previous layer and the corresponding encoder block as input and upsamples the high-level features to fuse with low-level features. The last UT module outputs the multi-scale perceptual aware features $F_o$. To alleviate the dilution of high-level features, we aggregate multi-level features $\{F_e^l\}_{l=1}^5$ into the global scene context $F_g$ via MCA. $F_o$ and $F_g$ are subsequently concatenated together to predict the final result by the prediction head.

### B. Point Context Transformer

To construct the hierarchical feature representation for understanding the semantic information of whole scenes, we follow PointNet++ [53] to build PCT in a pyramid manner. As shown in Fig. 3, PCT consists of two different transformers $\psi_{pre}$ and $\psi_{post}$. $\psi_{pre}$ extracts local context in each sampled group while $\psi_{post}$ models global context from the whole pooled point cloud.

Given an input point cloud $\mathcal{P} = \{p_1, p_2, ..., p_N\}$, we first generate a new subset $\{p_{c_1}, p_{c_2}, ..., p_{c_M}\}$ with $M$ points by the furthest point sampling (FPS) operation. In this new set, each point is recognized as the centroid to choose $K$ closest points in the local region within a given radius to form the point group. We denote $F_i = \{f_j | j \in \mathcal{N}(p_{c_i})\}$ as the $i_{th}$ group with the centroid $p_{c_i}$, where $F_i \in R^{K \times d}$ represents features of points in the $i_{th}$ group. The transformer is formulated as

$$q_i = F_i W_q, \quad k_i = F_i W_k, \quad v_i = F_i W_v, \tag{1}$$

$$F_i' = Softmax(q_i \cdot k_i / \sqrt{d}) v_i, \tag{2}$$

$$Trans(F_i) = F_i + FFN(F_i'), \tag{3}$$

where $W_q$, $W_k$, and $W_v$ are linear projections for query, key, and value terms, $d$ represents the feature dimension of key and value vectors, $FFN(\cdot)$ is a feed-forward network.

These groups are first fed into the Feature Normalization (FN) module to normalize the feature distributions in the local region. Then, the transformer block $\psi_{pre}$ takes these groups as input to model the context dependencies between points in the group, which describes the structural features on the boundary.

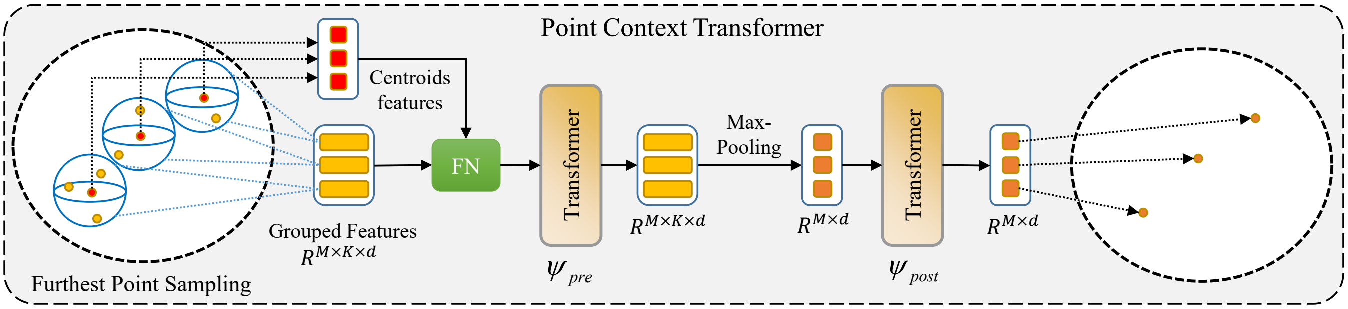


Fig. 3: Point Context Transformer (PCT). We first apply FPS and ball query to generate the local point groups. The Feature Normalization (FN) module is designed to simulate different feature distributions among diverse groups. Subsequently, two transformers are employed to capture the relationships of elements in the local region and global sampled points.

The max-pooling function aggregates the features of neighbor points into the centroid to reduce the resolution and expand the receptive field of the sampled points. Subsequently, the aggregated features are fed into the transformer block $\psi_{post}$ to learn the correlations among the sampled points, which assists in interfering with the category of each point.

The whole operation is formulated as

$$f_i' = \mathcal{A}(\psi_{pre}(F_i)), \tag{4}$$

$$F_e = \psi_{post}(\{f_i' | i \in [1, M]\}), \tag{5}$$

where $F_i = \{f_j | j \in \mathcal{N}(i)\} \in R^{K\times d}$ is the grouped set with the centroid $f_i$. $\mathcal{A}$ represents the max-pooling aggregation function. $\psi_{pre}$ and $\psi_{post}$ are both transformers in Eq. 3, which extract local context and global context, respectively. $\{f_i' | i \in [1, M]\} \in R^{M\times d}$ represent the extracted local context features of all groups, and $F_e$ is the output of PCT.

### C. Feature Normalization Module

The self-attention layer is the key component of the Transformer, which has a strong ability to learn the feature representations. But simply stacking self-attention layers to learn the deeper features will decrease the accuracy and robustness of performance. This is because points are sparse and irregular and feature distributions among the different local groups are diverse, while shared self-attention layers treat these groups equally. Inspired by [54], we utilize a Feature Normalization (FN) module to assign different weights to diverse groups. Let $\{f_{i,j}\}_{j=1,\dots,K} \in \mathbb{R}^{K\times d}$ be the grouped set with the centroid $f_i \in \mathbb{R}^d$, we transform the local grouped set by

$$\sigma = \sqrt{\frac{1}{M \times K \times d} \sum_{i=1}^{M} \sum_{j=1}^{K} (f_{i,j} - f_i)^2}, \tag{6}$$

$$\{f_{i,j}\} = \alpha \odot \frac{\{f_{i,j}\} - f_i}{\sigma + \epsilon} + \beta, \tag{7}$$

where the standard deviation $\sigma$ of all point cloud features are calculated to describe the offset across all groups and feature channels, $\alpha$ and $\beta$ are learnable parameters to simulate the distribution in different groups, $\odot$ is dot production and $\epsilon$ is a small number to maintain the numerical stability [55], [56].

### D. Scene Context Transformer

The high-level semantics and the multi-scale features are crucial for salient detection tasks [57]–[59]. Thereby we integrate the global semantic and multi-level features $\{F_e^l\}_{l=1}^5$ from encoders into the global scene context via the Scene Context Transformer (SCT) module. Scene context that describes the object distributions can locate the salient object. SCT mainly includes two components, i.e., Upsampling-and-Transformer (UT) and Multi-Context Aggregation (MCA).

UT upsamples the output $F_d^{i+1}$ from the previous UT block and concatenates it with features $F_e^i$ of the corresponding PCT module using the short link. Later, a transformer block excavates the inner relationships of the whole features as

$$F_d^i = Trans(C(U(F_d^{i+1}), F_e^i)), \tag{8}$$

where $U(\cdot)$ is an upsampling function and $C(\cdot,\cdot)$ is a concatenation function. MCA directly takes outputs $\{F_e^l\}_{l=1}^5$ from all PCT modules as input. We concatenate them together as the global scene context to assist in predicting the salient object. Specifically, we first adopt a channel compression operation for each output, which consists of the MLP layer and max-pooling function. MLP compresses outputs to an identical feature dimension and the max-pooling function is employed to generate different vectors for succeeding concatenation as

$$F_g = C(\{\mathcal{A}(\mathrm{MLP}(F_e^i)) | i \in [1, 5]\}), \tag{9}$$

where $C(\cdot,\cdot)$ is a concatenation function and $\mathcal{A}$ is the max-pooling function.

## IV. Experiments

### A. Experimental Setup

**Datasets.** PCSOD [16] is a benchmark dataset for 3D salient object detection. It includes 2,873 3D views from over one hundred scenes, where each view has 240,000 points. Each salient object is hierarchically labeled as three levels of annotations, i.e., class, bounding boxes, and segmentation map. Following the widely used split ratio of 7:3, PCSOD is randomly split into 2,000 training samples and 872 testing samples. Moreover, PCSOD contains a certain amount of challenging samples, such as multiple objects, small objects, complex structures, and low illumination.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3355968

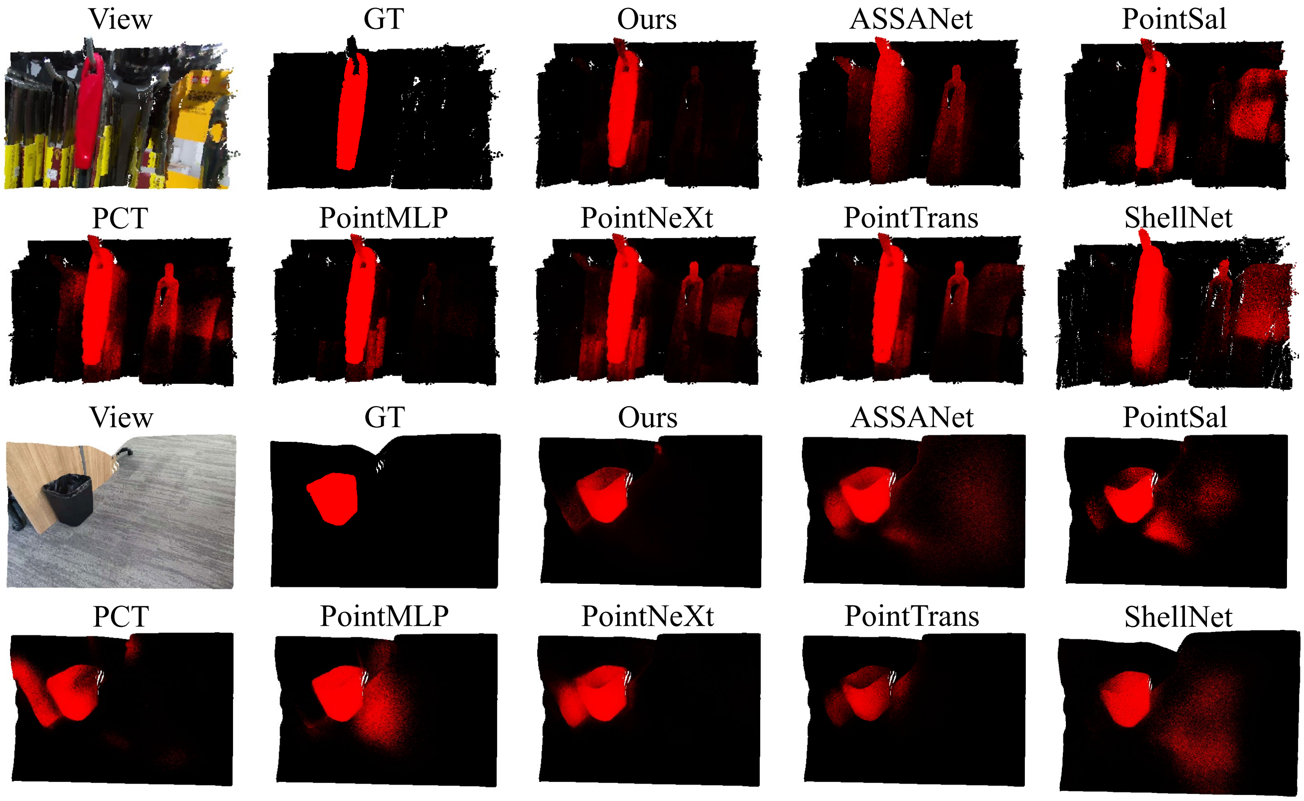


Fig. 4: The qualitative of different methods for PCSOD in the case of simple samples.

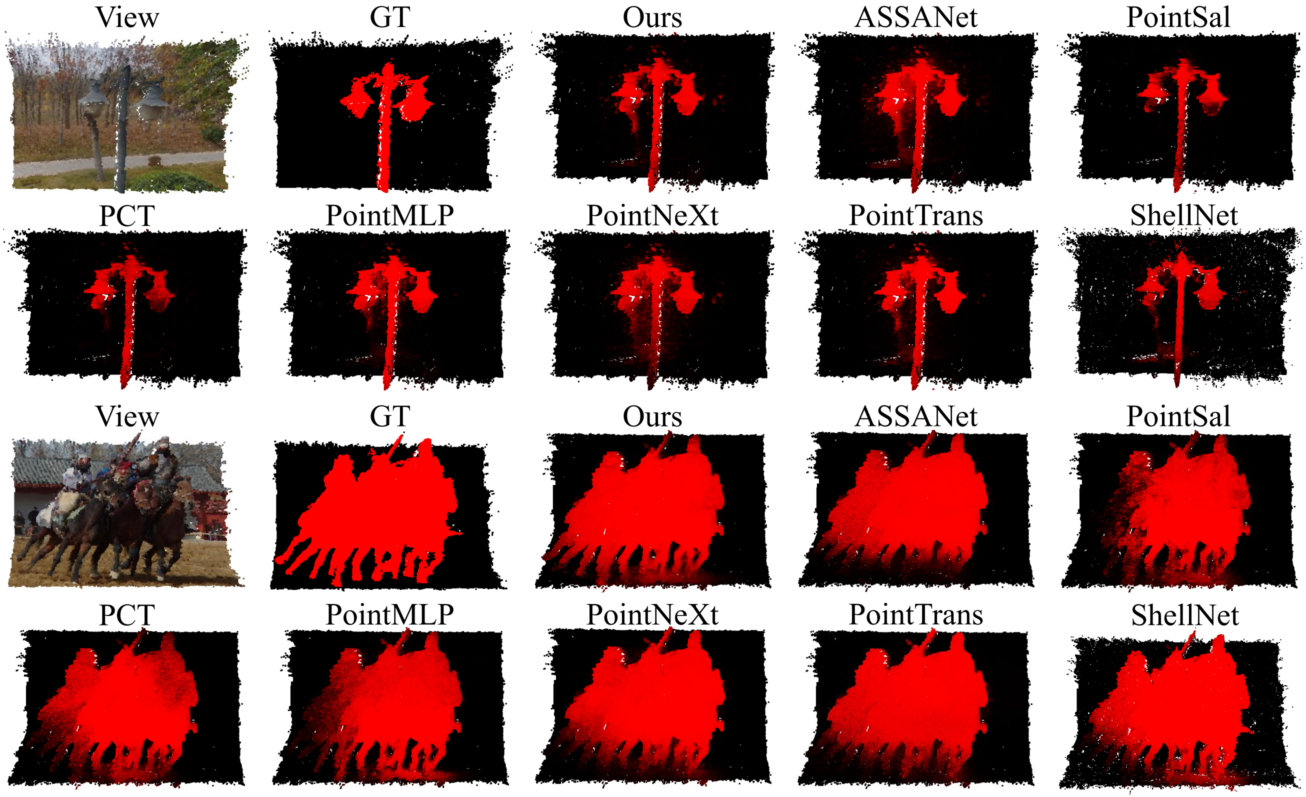


Fig. 5: The qualitative of different methods for PCSOD in the case of complex structures.

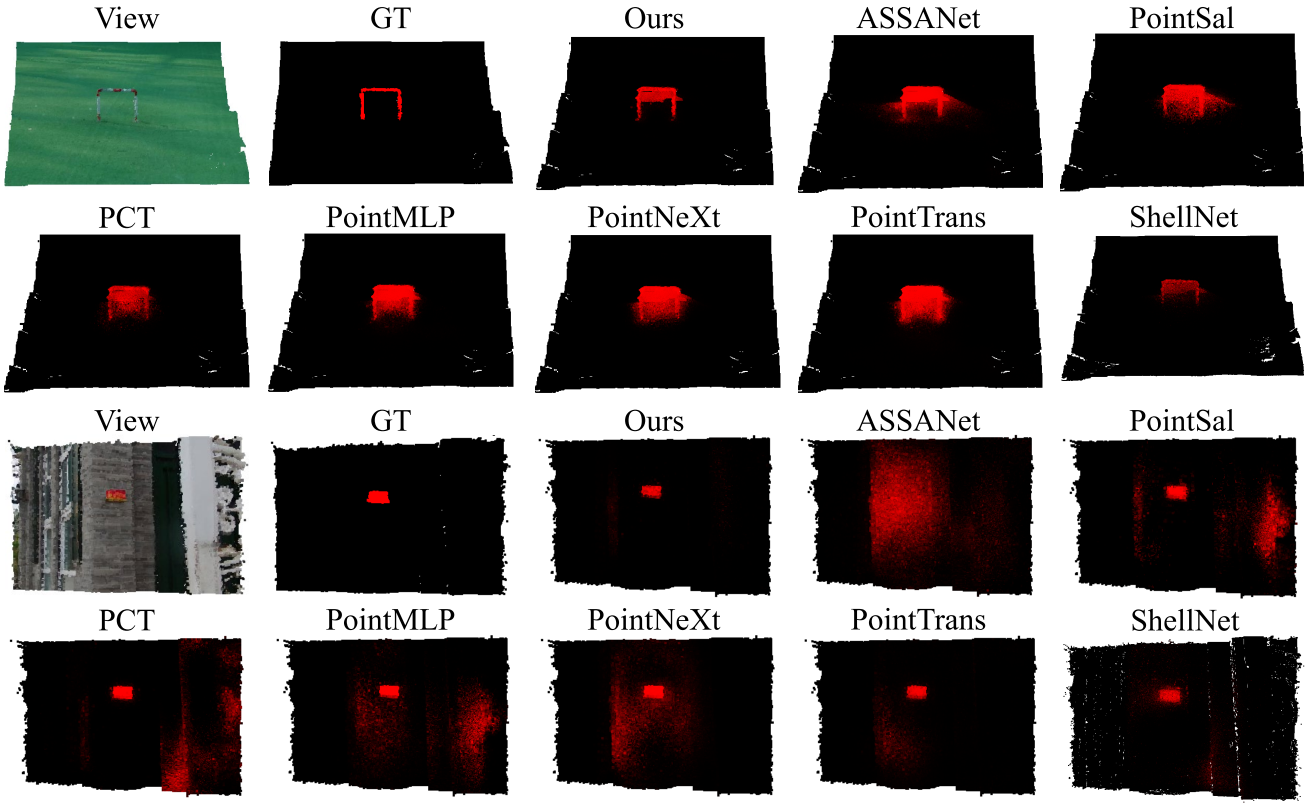


Fig. 6: The qualitative of different methods for PCSOD in the case of small objects.

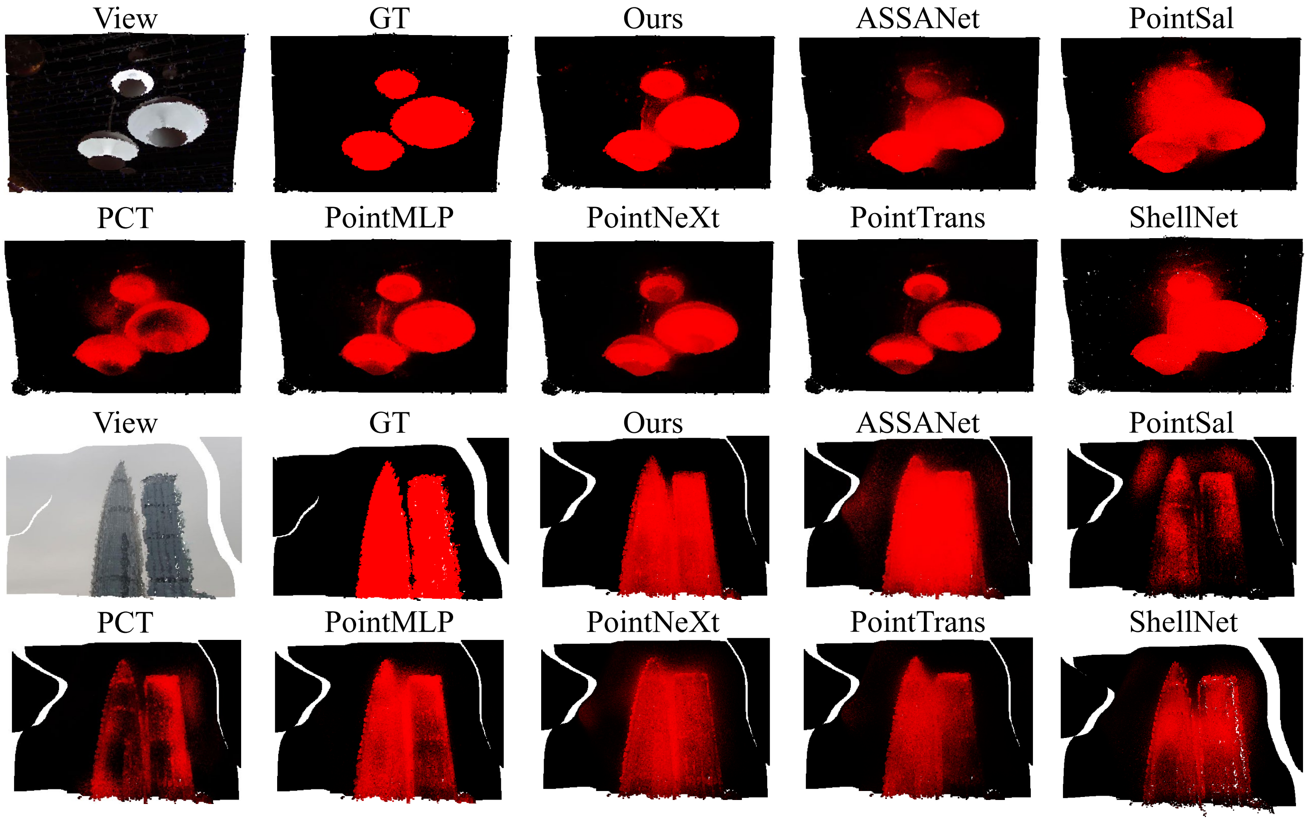


Fig. 7: The qualitative of different methods for PCSOD in the case of multi-objects.

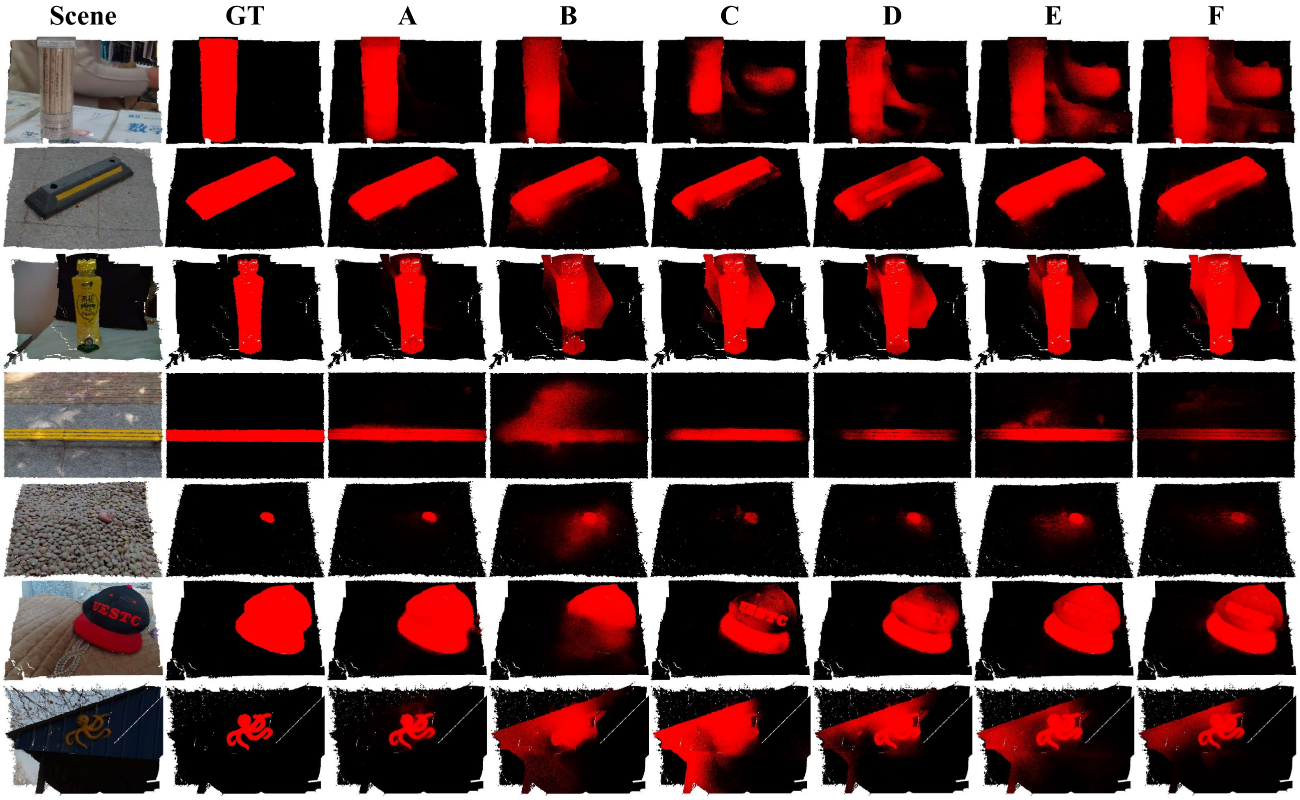


Fig. 8: The visual examples from the PCSOD dataset for ablation studies. For different configurations, “A”: PSOD-Net (*full modules*), “B”: PSOD-Net (*w/o FN*), “C”: PSOD-Net (*w/o* $\psi_{pre}$), “D”: PSOD-Net (*w/o* $\psi_{post}$), “E”: PSOD-Net (*w/o UT*), “F”: PSOD-Net (*w/o MCA*).

TABLE I: Comparison with the state-of-the-art methods on the PCSOD dataset.

| Methods | MAE ↓ | F-measure ↑ | E-measure ↑ | IoU ↑ |
|---|---|---|---|---|
| ASSANet [60] | 0.089 | 0.709 | 0.814 | 0.606 |
| PointTransformer [17] | 0.075 | 0.762 | 0.848 | 0.670 |
| PCT [46] | 0.069 | 0.770 | 0.846 | 0.652 |
| PointMLP [54] | 0.065 | 0.792 | 0.875 | 0.702 |
| PointNeXt [61] | 0.066 | 0.779 | 0.859 | 0.680 |
| ShellNet [62] | 0.074 | 0.753 | 0.848 | 0.648 |
| PointSal [16] | 0.069 | 0.769 | 0.851 | 0.656 |
| Ours | **0.058** | **0.805** | **0.878** | **0.711** |

**Evaluation Metrics.** We utilize four widely recognized evaluation metrics for performance benchmarking to compare the results of different methods. These metrics include the mean absolute error (MAE), F-measure, E-measure, and intersection over union (IoU).

The MAE metric evaluates the degree of point-wise approximation between the predicted segmentation maps and their corresponding ground truths. MAE can be formulated as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - g_i|, \tag{10}$$

where $p_i \in \mathcal{P}$ and $g_i \in \mathcal{G}$ are the prediction and ground truth, respectively.

F-measure is calculated as the harmonic mean of the precision (prec) and recall (reca), which is calculated by comparing the saliency map with the corresponding ground truth as

$$F - measure = \frac{(1-\beta^2)prec \cdot reca}{\beta^2 prec + reca}, \tag{11}$$

where $\beta^2$ is set to 0.3 to emphasize the importance of precision.

E-measure is a comprehensive evaluation metric that takes into account both local matching and global statistics information of segmentation maps for assessment [63], which can be formulated as

$$E - measure = \frac{1}{W \times H}\sum_{y=1}^{H}\sum_{x=1}^{W}\phi_{FM}(x,y), \tag{12}$$

where $H$ and $W$ are the height and width of the saliency map, and $\phi_{FM}$ is the enhanced alignment matrix.

IoU is a performance metric that quantifies the extent of overlap between two segmentation maps, which is defined as

$$IoU = \frac{intre}{union}, \tag{13}$$

where $inter$ and $union$ indicate the intersection and union of two segmentation maps, respectively.

**Implementation.** We implement our model with Pytroch on an NVIDIA RTX 2080ti GPU. The point clouds include 9-dimensional features that consist of spatial coordinates, RGB colors, and normalized coordinates. We randomly split the complete 3D view into patches with 4,096 points, and treat these patches as input. We train our model with the Adam optimizer in an end-to-end manner. The total training epochs are 800 and the initial learning rate is $5e^{-4}$.

## B. Comparison and Analysis

We compare our method with PointSal [16] and five representative segmentation methods [17], [46], [54], [60], [61].

**Quantitative results.** From Table I, our method achieves the best performance among all the methods on PCSOD. Our PSOD-Net outperforms the suboptimal model PointMLP [54] in all metrics on the PCSOD testing set. Point Transformer [17] and PCT [46] design various transformer modules on the point feature extraction. Although these sophisticated local feature extractors already learn the local context well, their performances are not effective enough due to the lack of representation of the global scene context. PointMLP enhances the ability to learn the point cloud feature representation by stacking more residual feed-forward MLPs. However, the fixed receptive field restricts the representation power of MLPs.

**Qualitative results.** To further verify the effectiveness of PSOD-Net, we visualize the predicted results on some challenging views, $e.g.$, structure-complex objects (Figure 5(b)), small objects (Figure 6(c)) and multi-objects (Figure 7(d)).

***Simple Samples in Fig. 4(a).*** While some other methods such as ASSANet, PointSal, and PCT may segment part of the background point cloud as a salient object in simpler scenes, our PSOD-Net avoids such interference from the background.

***Complex Structure in Fig. 5(b)***. Some salient objects have complex and sharp edges. Other methods, such as PointSal and PCT, will lose part of the edges or simply smooth complex edges into a whole. While our PSOD-Net can accurately segment these complex edges.

***Small Objects in Fig. 6(c)***. Small salient objects, due to their smaller point count, are easily disturbed by more numerous background points. Therefore, most methods tend to segment the adjacent background parts together with the salient object. However, our method can more accurately separate small salient objects from the background.

***Multi-objects in Fig. 7(d)***. A scene may contain multiple salient objects, which are very close to each other with boundaries tightly adjacent. Existing methods cannot accurately segment the boundaries of multiple objects, resulting in the blurring of multiple objects into whole, or being affected by background interference, resulting in only partial segmentation of salient objects. However, our method more accurately segments multiple salient objects and their boundaries.

The above results indicate that our method can accurately detect small objects, multiple objects, and objects with complex structures. This is attributed to our transformer-based PCT and SCT modules, which can improve the ability to detect salient objects. Specifically, firstly, the transformer is still effective in the PCSOD task, which can be proven by our experiment. Secondly, our PCT module can learn point-wise context information in the local and global regions. The local point-wise context information can describe the clear boundary structure, while the global point-wise context information can distinguish the points belonging to different objects and classes. This assists in accurately detecting multiple objects with complex structures. Besides, our network can identify objects at different scales by continuously using multiple PCT modules in the encoder to progressively capture multi-scale features. Thirdly, our SCT module can extract the global scene context information from the multi-scale features of the encoder. The global scene context information contains the potential relationship between salient objects and the background and is beneficial for locating the salient object from the interfering background. For these reasons, benefiting from our PCT and SCT modules, PSOD-Net can achieve better results.

TABLE II: Ablation study on the PCSOD testing set.

| Methods | MAE ↓ | F-measure ↑ | E-measure ↑ | IoU ↑ |
|---|---|---|---|---|
| PSOD-Net (w/o FN) | 0.071 | 0.755 | 0.842 | 0.649 |
| PSOD-Net (w/o $\psi_{pre}$) | 0.066 | 0.789 | 0.864 | 0.687 |
| PSOD-Net (w/o $\psi_{post}$) | 0.063 | 0.799 | 0.873 | 0.700 |
| PSOD-Net (w/o UT) | 0.069 | 0.775 | 0.858 | 0.678 |
| PSOD-Net (w/o MCA) | 0.061 | 0.801 | 0.873 | 0.702 |
| PSOD-Net (full) | **0.058** | **0.805** | **0.878** | **0.711** |

## C. Ablation Study

We split out five main components from our model and remove them one by one to verify their effectiveness. They are FN, $\psi_{pre}$, $\psi_{post}$ self-attention layers in PCT and UT, and MCA modules in SCT. All results are reported in Table II.

**Efficiency of Feature Normalization.** Point clouds are sparse and irregular and feature distributions among various local groups are diverse. However, the self-attention layer in transformers treats these groups equally. Simply stacking self-attention layers results in reduced accuracy and robustness. Therefore, we introduce the feature normalization module to assign different weights to different groups, balancing the feature difference between different groups. As shown in Table II, the performance of *PSOD-Net (w/o FN)* is decreasing.

**Efficiency of Point Context Transformer.** PCT includes two transformers that model the local and global point-wise relationships. $\psi_{pre}$ describes the boundary structure, while $\psi_{post}$ helps in distinguishing the points belonging to different objects and classes. To demonstrate the significance of PCT, we remove $\psi_{pre}$ and $\psi_{post}$ in PCT respectively. The performance of modified model *PSOD-Net (w/o $\psi_{pre}$)* and *PSOD-Net (w/o $\psi_{post}$)* is shown in Table II. It can be observed that after eliminating each of the two modules separately, the performance of the model drops to varying degrees.

**Efficiency of Scene Context Transformer.** SCT explores scene-wise context by UT and MCA modules. UT progressively recovers the resolution of the point cloud via the upsample and transformer, MCA takes the output of all encoders as input to concatenate them as the global scene context to assist in predicting the salient object. To evaluate the efficiency of SCT, we remove UT and MCA respectively, resulting in a modified model denoted as *PSOD-Net (w/o UT)* and *PSOD-Net (w/o MCA)*. As shown in Table II, the model performs poorly when lacking the scene-wise context.

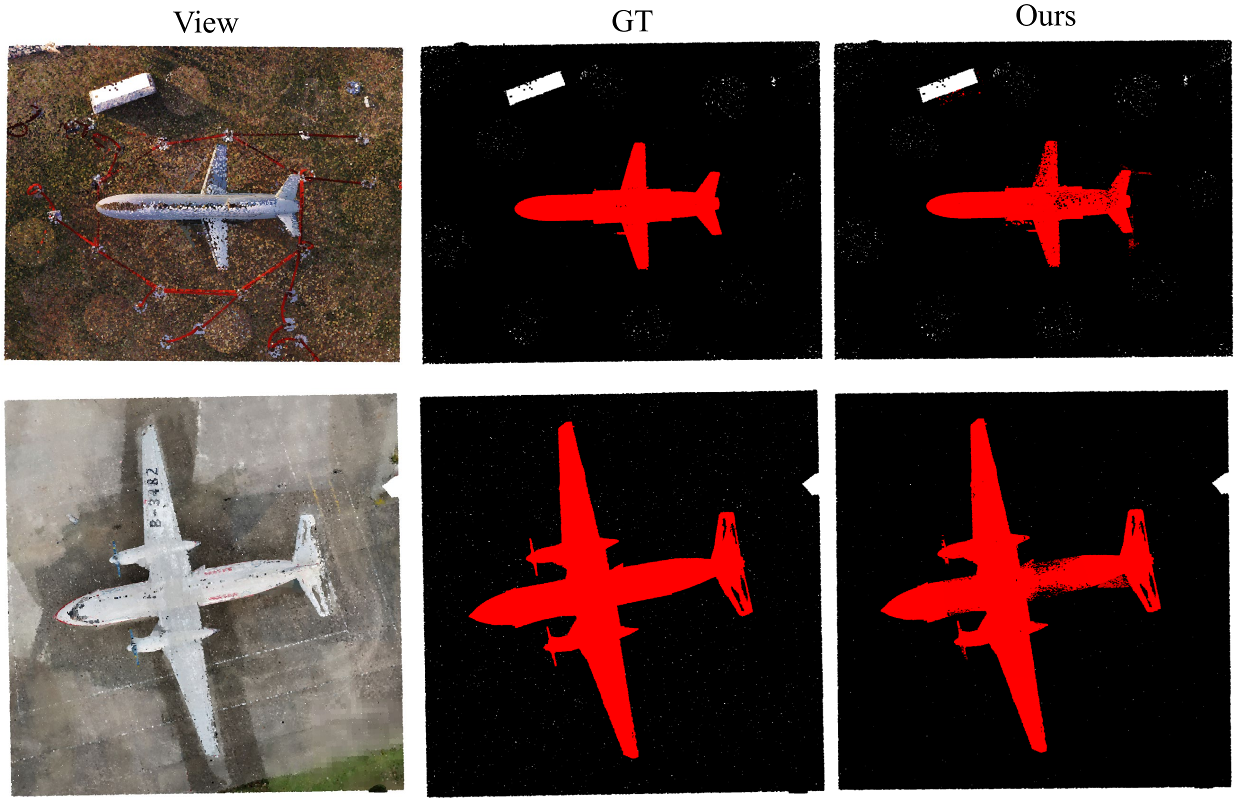


Fig. 9: The visual results of aircraft remote sensing data. The two input data are scanned by Leica BLK360 and UAV aerial Photogrammetry, respectively. The GT data is manually labeled. It can be seen that our method can detect the aircraft accurately.

| Method | #parameters ↓ | FLOPs ↓ | MAE ↓ |
|---|---|---|---|
| PointTransformer [17] | 7.8M | 2.8G | 0.075 |
| PCT [46] | 2.9M | 2.3G | 0.069 |
| PointMLP [54] | 12.6M | 9.8G | 0.065 |
| PointNeXt [61] | 7.1M | 7.6G | 0.066 |
| PointSal [16] | 4.8M | 1.4G | 0.069 |
| Ours | 8.2M | 4.1G | 0.058 |

TABLE III: Comparison of model parameters, floating point operations, and MAE of different methods.

Therefore, the contextual information captured by PCT and SCT indeed improves the performance of our method.

### D. Efficiency and Model complexity

To compare the complexity of different methods, we show the parameter numbers, floating point operations (FLOPs), and MAE in Table III. Floating point operations are tested on 4096 points. It can be seen that our method is competitive in terms of space and time efficiency.

### E. Application

In this subsection, we develop a practical application on the aircraft remote sensing data, which achieves favorable results. Specifically, we test our approach on two data are two real-scanned remote sensing data in Fig. 9. These two data are scanned by Leica BLK360 and UAV aerial Photogrammetry, respectively. In the results, our PSOD-Net accurately segments the aircraft.

## V. Conclusion

3D salient object detection is a new topic, remaining many non-trivial problems to solve. For the first time, we propose a transformer model for 3D salient object detection from point clouds, namely PSOD-Net. Our PSOD-Net enhances the ability to learn context-dependent feature representations at the point and scene levels by introducing two different types of transformers. The proposed Point Context Transformer (PCT) models hierarchical context-aware features at the point level by two different transformer blocks. The proposed Scene Context Transformer (SCT) captures the global scene context by integrating the multi-scale contextual information from different-level PCT modules. Thus, PSOD-Net is robust to the cases of small objects, multiple objects, and objects with complex structures. Extensive experiments verify the effectiveness of our method over its competitors. In the future, we will apply our method to more data with different structures and information, and extend our idea to other 3D task, such as feature extraction and segmentation.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. T2322012, No. 62172218, No. 62032011), the Shenzhen Science and Technology Program (No. JCYJ20220818103401003, No.

JCYJ20220530172403007), the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010170), the Research Grants Council of the Hong Kong Special Administrative Region, China (No. UGC/FDS16/E14/21) and Hong Kong Metropolitan University Research Grant (No. RD/2022/2.13).

## References

[1] Z. Wang, J. Guo, C. Zhang, and B. Wang, "Multiscale feature enhancement network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–19, 2022.

[2] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–15, 2022.

[3] F. P. Tasse, J. Kosinka, and N. Dodgson, "Cluster-based point set saliency," in *ICCV*, 2015, pp. 163–171.

[4] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *TIP*, vol. 19, no. 1, pp. 185–198, 2009.

[5] E. Alexiou and T. Ebrahimi, "Point cloud quality assessment metric based on angular similarity," in *ICME*, 2018, pp. 1–6.

[6] X. Ding, W. Lin, Z. Chen, and X. Zhang, "Point cloud saliency detection by local and global feature fusion," *TIP*, vol. 28, no. 11, pp. 5379–5393, 2019.

[7] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren, "Pointcloud saliency maps," in *ICCV*, 2019, pp. 1598–1606.

[8] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, pp. 1–11, 2023.

[9] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *CVPR*, 2018, pp. 7652–7660.

[10] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *CVPR*, 2021, pp. 11 784–11 793.

[11] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[12] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, no. 2, pp. 117–150, 2019.

[13] P. O'Leary, M. Harker, and M. Janko, "Instrumentation and surface modeling for the measurement of disks, circular- and cylindrical-strips," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 5, pp. 1181–1189, 2014.

[14] D. Yu, J. Xiao, and Y. Wang, "Registration method for point clouds of complex rock mass based on dual structure information," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–18, 2022.

[15] Y. Song, F. He, Y. Duan, T. Si, and J. Bai, "LSLPCT: an enhanced local semantic learning transformer for 3-d point cloud analysis," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–13, 2022.

[16] S. Fan, W. Gao, and G. Li, "Salient object detection for point clouds," in *ECCV*, vol. 13688, 2022, pp. 1–19.

[17] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021, pp. 16 259–16 268.

[18] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *ICCV*, 2013, pp. 2976–2983.

[19] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012, pp. 733–740.

[20] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 1915–1926, 2011.

[21] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*, 2012, pp. 29–42.

[22] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017, pp. 3203–3212.

[23] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Scene context-aware salient object detection," in *ICCV*, 2021, pp. 4156–4166.

[24] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[25] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *CVPR*, 2019, pp. 7479–7489.

[26] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *ICCV*, 2019, pp. 8779–8788.

[27] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *ECCV*, 2020, pp. 275–292.

[28] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *CVPR*, 2020, pp. 8582–8591.

[29] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *ECCV*, 2020, pp. 235–252.

[30] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *CVPR*, 2020, pp. 3052–3062.

[31] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *CVPR*, 2020, pp. 3472–3481.

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

[33] G. Kim, D. Huber, and M. Hebert, "Segmentation of salient regions in outdoor scenes using imagery and 3-d data," in *IEEE Workshop on Applications of Computer Vision*, 2008, pp. 1–8.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *ICLR*, 2021, pp. 1–16.

[37] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *CVPR*, 2020, pp. 10 076–10 085.

[38] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[39] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021, pp. 10 347–10 357.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021, pp. 1–12.

[42] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–15, 2022.

[43] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-sharpening based on transformer with redundancy reduction," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[44] W. Tang, F. He, Y. Liu, and Y. Duan, "Matr: Multimodal medical image fusion via multiscale adaptive transformer," *TIP*, vol. 31, pp. 5134–5149, 2022.

[45] W. Tang, F. He, and Y. Liu, "YDTR: infrared and visible image fusion via y-shape dynamic transformer," *IEEE Trans. Multim.*, vol. 25, pp. 5413–5428, 2023.

[46] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.

[47] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3d object detection with pointformer," in *CVPR*, 2021, pp. 7463–7472.

[48] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3d object detection," in *ICCV*, 2021, pp. 2906–2917.

[49] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3d object detection via transformers," in *ICCV*, 2021, pp. 2949–2958.

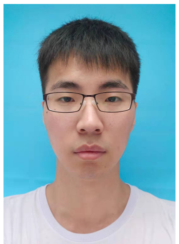
[50] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *ICCV*, 2021, pp. 3164–3173.

[51] X. Liu, Z. Han, Y. Liu, and M. Zwicker, "Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network," in *AAAI*, 2019, pp. 8778–8785.

[52] Z. Zhang, P. Gao, S. Peng, C. Duan, and P. Zhang, "Enhanced point feature network for point cloud salient object detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 1617–1621, 2023.

[53] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, 2017, pp. 5099–5108.

[54] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," in *ICLR*, 2021, pp. 1–13.
[55] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.
[56] Y. Wu and K. He, "Group normalization," in *ECCV*, 2018, pp. 3–19.
[57] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *AAAI*, 2020, pp. 10 599–10 606.
[58] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *CVPR*, 2019, pp. 3917–3926.
[59] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *CVPR*, 2020, pp. 9413–9422.
[60] G. Qian, H. Hammoud, G. Li, A. Thabet, and B. Ghanem, "Assanet: An anisotropic separable set abstraction for efficient point cloud representation learning," *NeurIPS*, vol. 34, pp. 28 119–28 130, 2021.
[61] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," *NeurIPS*, vol. 35, pp. 23 192–23 204, 2022.
[62] Z. Zhang, B.-S. Hua, and S.-K. Yeung, "Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics," in *ICCV*, 2019, pp. 1607–1616.
[63] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018, pp. 698–704.



**Zeyong Wei** is a Ph.D. candidate at Nanjing University of Aeronautics and Astronautics (NUAA). He received his M.Sc. degree from NUAA. He has published several papers on SIGGRAPH, IJCV, TPAMI, etc. His research interests include computer vision and learning-based geometry processing.

**Baian Chen** is now pursuing his PhD degree at Nanjing University of Aeronautics and Astronautics (NUAA), China. He received his B.Sc. degree from China University of Mining and Technology. His research interests include 3D vision and learning-based geometry processing.

**Weiming Wang** is an Assistant Professor in Hong Kong Metropolitan University. He received his PhD degree from The Chinese University of Hong Kong in 2014. He was an assistant researcher at Shenzhen Institutes of Advanced Technology from 2014 to 2015. After that, he worked in high-tech companies for a few years. His research interests include image processing, point cloud processing and deep learning. He is currently an Associate Editor for The Visual Computer.

**Honghua Chen** received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2022. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interest is 3D Measurement/Vision and point cloud processing.

**Mingqiang Wei** (Senior Member, IEEE) received his Ph.D degree (2014) in Computer Science and Engineering from the Chinese University of Hong Kong (CUHK). He is a professor at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). Before joining NUAA, he served as an assistant professor at Hefei University of Technology and a postdoctoral fellow at CUHK. He was a recipient of the CUHK Young Scholar Thesis Awards in 2014. He is now an Associate Editor for ACM TOMM, The Visual Computer, Journal of Electronic Imaging, and a Guest Editor for IEEE Transactions on Multimedia. His research interests focus on 3D vision, computer graphics, and deep learning.

**Jonathan Li** (Fellow, IEEE) received the Ph.D. degree in geomatics engineering from the University of Cape Town, South Africa, in 2000. He is a Professor with the Department of Geography and Environmental Management and cross-appointed with the Department of Systems Design Engineering, University of Waterloo, Canada and a Fellow of the Engineering Institute of Canada. His main research interests include image and point cloud analytics, mobile mapping, and AI-powered information extraction from LiDAR point clouds and earth observation images. He has co-authored over 500 publications, including 300+ in refereed journals and 200+ in conference proceedings. Dr. Li is a recipient of the 2021 Geomatica Award, 2020 Samuel Gamble Award, and 2019 Outstanding Achievement Award in Mobile Mapping Technology. He is currently serving as the Editor-in-Chief of the International Journal of Applied Earth Observation and Geoinformation, Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and Canadian Journal of Remote Sensing.