Salient Object Detection for Point Clouds

Songlin Fan[1,2], Wei Gao[1,2,✉], and Ge Li[1]

[1] Peking University Shenzhen Graduate School
[2] Peng Cheng Laboratory
{slfan, gaowei262, geli}@pku.edu.cn
https://git.openi.org.cn/OpenPointCloud/PCSOD

Abstract. This paper researches the unexplored task—point cloud salient object detection (SOD). Differing from SOD for images, we find the attention shift of point clouds may provoke saliency conflict, *i.e.*, an object paradoxically belongs to salient and non-salient categories. To eschew this issue, we present a novel view-dependent perspective of salient objects, reasonably reflecting the most eye-catching objects in point cloud scenarios. Following this formulation, we introduce ***PCSOD***, the first dataset proposed for point cloud SOD consisting of 2,872 in-/out-door 3D views. The samples in our dataset are labeled with hierarchical annotations, *e.g.*, super-/sub-class, bounding box, and segmentation map, which endows the brilliant generalizability and broad applicability of our dataset verifying various conjectures. To evidence the feasibility of our solution, we further contribute a baseline model and benchmark five representative models for a comprehensive comparison. The proposed model can effectively analyze irregular and unordered points for detecting salient objects. Thanks to incorporating the task-tailored designs, our method shows visible superiority over other baselines, producing more satisfactory results. Extensive experiments and discussions reveal the promising potential of this research field, paving the way for further study.

Keywords: Salient object detection, point cloud, dataset, baseline.

1 Introduction

Salient objects describe the most attractive objects with respect to their surroundings. Due to its myriad applications, salient object detection (SOD) can provide the pre-processing results for many vision tasks, such as 3D shape classification [46], compression [32], and quality assessment [30], to name a few. Distinct from the relevant task [6, 43, 61] for predicting eye fixation positions, namely saliency detection, SOD demands locating salient objects and completely segmenting them further, thus being more challenging. Most existing SOD works [9, 12, 15, 26, 37, 58] devote their efforts to analyzing salient objects on regular images. With the fast revolution of 3D collection equipment, point clouds as the raw output of many devices (such as LiDAR and depth sensors) have a growing presence in research and applications. Compared with the adoption of alternative 3D formats, data processing directly on point clouds avoids

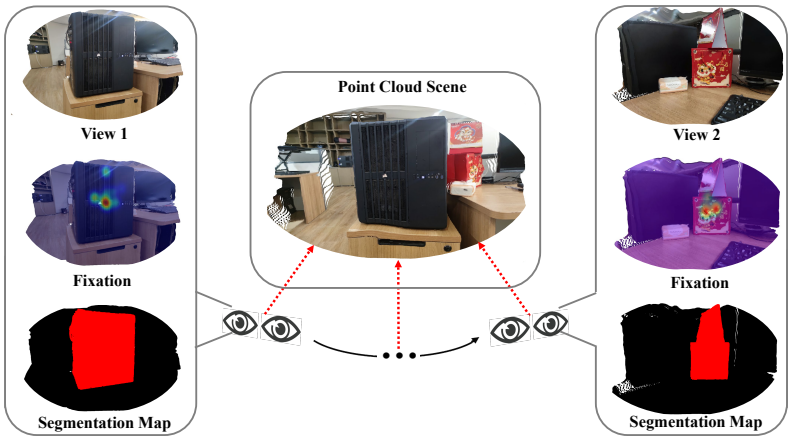

Fig. 1: Illustration of saliency conflict. The variation of attention allocated to the black computer causes a contradiction that the black computer simultaneously belongs to the salient and non-salient objects for this scene. We, therefore, propose to analyze the salient objects of point clouds according to the views.

information loss and computational redundancy in format conversion that may induce performance drops. Despite the flourishing advance of many point-based tasks, *e.g.*, classification [3], object detection [42], and segmentation [16], point cloud SOD is still in its infancy, and many issues have not been discussed yet.

As immersive visual media, point clouds offer a watching experience with six degrees of freedom (6DOF). Unlike the watching of static images, the attention allocation of humans varies when the view changes. The research community dubs the phenomenon that attention being allocated from one region to another as the attention shift [9, 44]. However, we find that the attention shift of point clouds may trigger a new thorny problem that we name saliency conflict, *i.e.*, an object paradoxically belongs to salient and non-salient categories for different views of one point cloud scene sample. Fig. 1 shows an example of an office scene recorded by point clouds. The attention allocated to the black computer varies as the view changes, which causes the black computer to go from being the salient object to the non-salient object. **Then is the black computer the salient object of this office scene?** The answer matters not only the definition of salient objects in point cloud scenarios but also the relevant dataset construction.

In this paper, we argue that the manifestations of salient objects in point clouds depend on the views, and point cloud SOD is to compute the salient objects of any given view in 3D space. The union of salient objects (segmentation maps) of "given views" indicates the complete description of salient objects for scenes in point clouds. For Fig. 1, different segmentation maps correspond to different views, and the union of segmentation maps represents the salient objects of this office scene. Firstly, this formulation makes it easier to grasp the nature of the SOD problem due to the fact that humans actually observe only one view at

a time while the viewpoint is free. Broadly speaking, the image is a special case of only a single view. Secondly, this formulation avoids the complex modeling to handle the whole 3D scene with saliency conflict phenomenon, which can benefit the design of simpler models capable of analyzing different views. Thirdly, this formulation eases the dataset construction with the human-annotated most attractive objects via subjective experiments, since the subjective experiment results of different views sometimes cannot be reflected into a large-scale point cloud sample (such as the office scene sample in Fig. 1) simultaneously without our view-dependent saliency analysis.

Following our formulation of point cloud SOD, we introduce *PCSOD*—the first versatile dataset for point cloud SOD with densely annotated labels. Our dataset contains 2,872 frequent 3D views that belong to over one hundred in-/out-door scenes. The manual data collection phase lasts over one year, and the samples reflect a wide range of scenarios in our lives. Detailed statistics show that our dataset has 138 object categories and 53.4% difficult samples, which ensures its brilliant generalizability. To extend the applicability of this new dataset, we provide hierarchical annotations for each sample, including super-/sub-class, bounding box, and segmentation map. The proposed dataset as a comprehensive platform can conveniently support research on multi-task learning [48] and other valuable vision tasks, not limited to point cloud SOD.

Since point clouds record 3D information in the format of irregular and unordered points, existing SOD models [9, 12, 15, 26, 58] for images cannot be transferred for point cloud processing. Additionally, though several representative point-based models [3, 16, 25, 38, 59] have been developed for other segmentation tasks, they are incapable of performing well in SOD. These models for other segmentation tasks fail to consider the particularities of SOD, *i.e.*, the benefits of multi-scale features [35] and the refinement of global semantics [4,27]. To prove the feasibility of our solution, we further develop a baseline model and benchmark five representative segmentation models for comparison and analysis of point cloud SOD. Owing to incorporating the task-tailored designs, the proposed baseline model can take full advantage of the multi-scale features and global semantics to locate salient objects and accurately separate them. Extensive experiments verify the effectiveness of our solution for point cloud SOD.

In summary, we conclude the contributions as follows:

1) We propose a novel view-dependent perspective of point cloud SOD. Our formulation avoids the saliency conflict, emphasizes the nature of SOD, and reasonably reflects the most eye-catching objects in point cloud scenarios.
2) We construct the first versatile dataset for point cloud SOD, termed *PCSOD*. Our dataset has brilliant generalizability and broad applicability, expected to be a catalyst for point cloud SOD and many other vision tasks.
3) We develop a baseline model for point cloud SOD. Our baseline model has a full consideration of the particularities of SOD, outperforming other baseline models by a clear margin.
4) We establish the first benchmark of point cloud SOD, conduct a thorough analysis, and bring a new perspective toward point cloud SOD.

2 Related Work

Salient Object Detection. Following the pioneer attempt [17], many early works [24, 36, 47, 53] design hand-crafted features to exploit low-level cues. These methods cannot obtain satisfactory accuracy because of the lack of semantic cues. Thanks to the powerful capability of neural networks in abstracting semantics, the bottleneck of traditional methods is broken. Hou *et al.* [15] introduce short connections into a skip-layer structure. The advanced representations at multiple layers thus can be fully utilized. Siris *et al.* [45] propose a semantic scene context-aware framework to capture sufficient high-level semantics for locating salient objects. To rich the semantic information diluted during the top-down transmission, some recent works [4, 27, 35] explicitly extract global semantics and append them into low-level features, achieving visible performance improvement. Despite the gratifying achievements of existing RGB image-based methods [28, 40, 55, 60], they still have difficulty understanding complex scenes for lacking spatial geometry information. Consequently, researchers begin extending the task of SOD on 3D images, such as RGB-D images [10, 15, 20, 22, 26, 56, 58] and light field images [23, 29, 50, 57], which show significant potential. A detailed description of these image-based methods is beyond the scope of this article. Please refer to the relevant surveys [11, 51, 62] for more introduction. We can conclude that all these efforts are confined to the image domain. This work will disentangle the limitation and probe SOD on point clouds.

Regarding the attention modeling on point clouds, we also learn that a few methods [6, 13, 18, 43, 46, 61] are developed to automatically compute the human attention distribution. The algorithms of these methods merely produce a heatmap of the attention distribution, while the SOD task we study demands completely segmenting the salient objects, thus being more challenging.

Deep Learning on Point Clouds. Processing point clouds has long been a significant challenge. Previous works [19, 21] tend to first rearrange raw points via octree or kdtree. The emergence of PointNet/PointNet++ [3, 38] shows us a new approach for raw point processing. They employ shared multilayer perceptrons (MLPs) to extract point-wise features and achieve state-of-the-art performance across many vision tasks. Following PointNet, three directions are mainly adopted to improve the performance further, *i.e.*, powerful convolution [25, 54], effective neighborhood connection [52, 59], and advanced reduction [16, 39]. Li *et al.* [25] propose to learn an X-transformation from raw points by imitating the typical convolution, while Wu *et al.* [54] regard the typical convolution as the combination of weight and density functions. ShellNet [59] arranges neighbors into concentric spherical shells that have a convolution order from the inner to the outer shells. Wang *et al.* [52] propose a simple operation known as EdgeConv, which extracts local geometric features while retaining permutation invariance. To explore more advanced reduction operations, Hu *et al.* [16] and Qian *et al.* [39] resort to attentive pooling and anisotropic reduction, respectively. However, these methods are not initially developed for SOD, ignoring the particularities of SOD.

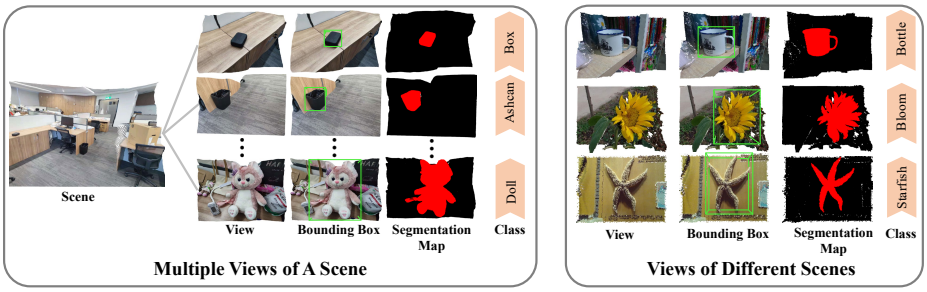

Fig. 2: Examples from our *PCSOD* dataset with hierarchical annotations.

3 Proposed Dataset

Datasets [9, 23, 41] have become the driving force behind many vision tasks, especially with the emergence of deep learning. With this in mind, we introduce *PCSOD* for: (1) probing a new challenging task, (2) facilitating research on new issues, and (3) verifying new conjectures. Next, we will elaborate more details about our dataset. Besides, some visual examples are shown in Fig. 2 and Fig. 3.

3.1 Dataset Construction

Data Collection. Point clouds in existing datasets [1, 2, 5, 14, 34] are often collected for specific scenes (such as outdoor road or indoor office scenes). In contrast, a high-quality SOD dataset [49] demands rich scenes, which motivates us to collect diverse data by ourselves. The data collection phase takes over one year, and we collect 2,872 3D views from over one hundred preset scenes across dozens of cities. Each 3D view has 240,000 points. This process can also simulate the 3D view acquisition when "travelling" in an off-shelf large-scale point cloud sample (such as an office or even a city). As shown in Fig. 2, the 3D views of a scene constitute a series of watching descriptions of this scene whose salient objects can be obtained from subjective experiments without saliency conflict.

Data Annotation. Referring to the determination of salient objects in images [49, 50], we employ thirty professional annotators to label the salient objects from given views. Before the labeling, every annotator is pre-trained over fifteen samples. To ensure the annotation accuracy, we divide these thirty annotators into ten groups. Three annotators in one group jointly determine the salient objects, then cross-validated by other groups. An object is regarded as a positive label only if more than eighty percent of annotators verify it. The recently released datasets [9, 34] indicate that offering hierarchical annotations benefits the applicability of a new dataset. As shown in Fig. 2, we hierarchically label the determined salient objects and provide three levels of annotations, *i.e.*, class, bounding box,

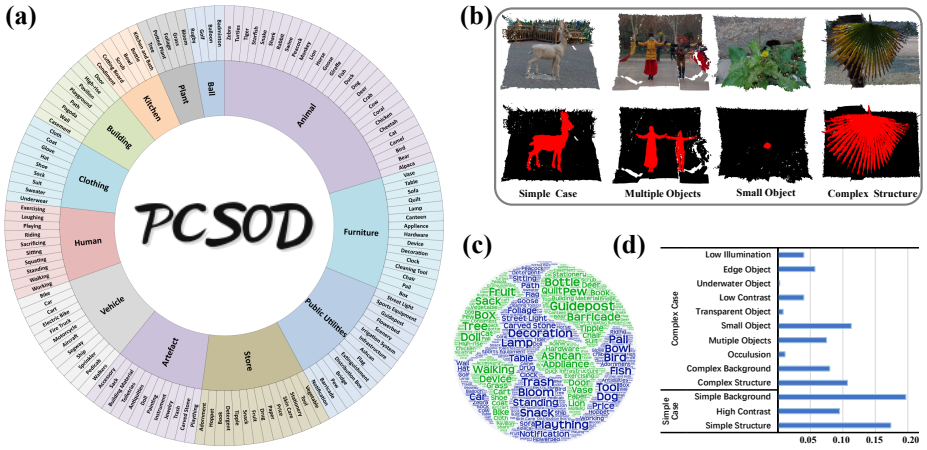

Fig. 3: Statistics of our *PCSOD* dataset. (a) Categories of salient objects. (b) Illustration of challenging samples. (c) Word cloud of salient objects. (d) Histogram distribution of challenging samples.

and segmentation map. Each level of annotations is obtained through corresponding professionals. Furthermore, at least two passes of verification are performed for each annotation to ensure its quality.

Data Split. Having a standard dataset split [9,50] is conducive to fairly studying and comparing the pros and cons of algorithms. Following the ratio of 7:3 adopted by many datasets [50], our *PCSOD* is randomly split into 2,000 samples for training and 872 samples for testing.

3.2 Dataset Statistics

Diverse Object Categories. A diverse SOD dataset should have broad coverage of scenes in the real world to ensure brilliant generalizability. Our *PCSOD* covers a wide range of scenarios in our lives. As shown in Fig. 3(a) and Fig. 3(c), the salient object categories have a heterogeneous variety. Specifically, objects in our dataset can be categorized into 12 super-classes, *e.g.*, human, animal, plant, *etc.* These 12 super-classes are further comprised of 138 sub-classes, fully covering the daily situations. The diverse salient object categories enable a comprehensive understanding of the attention allocation of humans in real-world scenes.

Rich Annotations. A versatile dataset should not only support the study of existing issues but also adapt to new research directions. As shown in Fig. 2, our *PCSOD* offers hierarchical annotations, *e.g.*, super-/sub-class, bounding box, and segmentation map. These annotations help researchers understand each sample of our dataset from different aspects (such as object property, object proposal, and

scene parsing), sparking novel ideas. Besides, our annotations are very precise. The segmentation maps accurately reflect the structures of objects in 3D scenes, even though some are very complex (see the complex structure case in Fig. 3(b)).

Difficult Samples. A valuable dataset should contain a certain amount of difficult samples and dive into the problems. The difficult samples benefit the performance of models confronting various complex scenes. With this consideration, we add many challenging samples to our dataset, including multiple objects, small objects, complex structures, low illumination, *etc.* Some visual examples are shown in Fig. 3(b). Fig. 3(d) further details the proportion of samples with each attribute. Statistics indicate that our dataset has 53.4% difficult samples, which evidences that the proposed *PCSOD* is very challenging.

4 Proposed Method

Extending the concept of salient objects in images to point clouds, we formulate that the salient objects of views from a scene indicate the complete description of salient objects in this scene. Point cloud SOD aims to identify the salient objects of any given view. While various methods have been developed for image-based SOD, they cannot handle irregular and unordered point clouds. Moreover, existing point-based segmentation models for other tasks cannot guarantee the performance of identifying salient objects. These circumstances motivate us to design a baseline model and excavate potential directions for point cloud SOD.

4.1 Overall Architecture

As shown in Fig. 4, the proposed baseline model inherits a typical encoder-decoder architecture. The encoder extracts multi-level features from raw points, while the decoder enhances and fuses the extracted features to predict salient objects. To illustrate the effectiveness of our designs, we introduce the classical PointNet++ [38] as the encoder. It has been studied [27] that high-level features will be gradually diluted when transmitted to low-level ones. To address this issue, some recent image-based methods [4, 27, 35] explicitly extract global semantics and append them into low-level features, observing gratifying performance improvement. Inspired by the philosophical designs of these methods, we design two key modules, *i.e.*, Point Perception Block (PPB) and Saliency Perception Block (SPB), to take full advantage of the benefits of multi-scale features and the refinement of global semantics for locating salient objects.

Formally, let $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ represent a view of N points with associated point-wise features (*e.g.*, RGB colors), where $v \in \mathbb{R}^{d_{in}}$. To obtain the probabilities $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ of corresponding points being salient, the encoder first extracts multi-level features $\{F_l\}_{l=1}^{4}$ from raw points $\mathcal{V}$. The l^{th} level features $F_l = \{f_1^l, f_2^l, ..., f_{N_l}^l\}$ have $N_l = \frac{N}{4^l}$ aggregated points with doubling the feature dimension compared with F_{l-1} (except the feature dimension of the first level

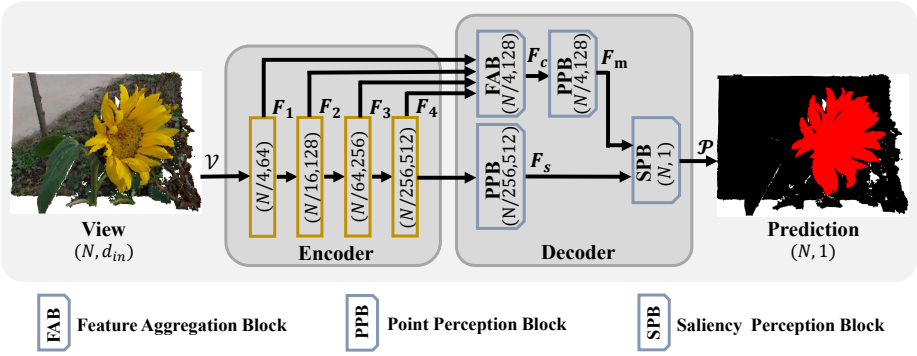

Fig. 4: Overall architecture of the proposed baseline model, which has a typical encoder-decoder architecture.

is fixed to 64). Then we aggregate multi-level features $\{F_l\}_{l=1}^{4}$ into the compact representations F_c via the Feature Aggregation Block (FAB). As shown in Fig. 5, the operations in FAB are very straightforward, *i.e.*, upsampled high-level features are sequentially fused with low-level features. We adopt the common trilinear interpolation as the upsampling operation to match the spatial size of different level features, while the fusion operation we employ is concatenation along the feature dimension followed by MLPs. Following previous works [16, 38], the feature concatenation can simultaneously retain the originality of the fused two level features and is proved to be very effective for point cloud feature fusion. Note that the feature fusion in all modules is uniformly through concatenation unless otherwise stated. To prevent the dilution of high-level features, the PPB is proposed to abstract global semantics and strengthen the multi-scale representations. We obtain global semantics F_s and multi-scale features F_m from the highest-level features F_4 and the compact representations F_c, respectively, using two PPBs with different configurations. The global semantics can supplement the diluted high-level features in multi-scale features and alleviate the distraction of non-salient background. To achieve this, we further develop the SPB to integrate multi-scale features F_m and global semantics F_s, and produce the final prediction $\mathcal{P}$. Next, we will elaborate on the details of our PPB and SPB.

4.2 Proposed Modules

Point Perception Block. The global semantics and multi-scale features are important for SOD [4, 27, 35]. The former helps to locate the positions of salient objects, while the latter is conducive to recognizing salient objects of different sizes. Besides, the acquisition of them demands enlarging the receptive fields of features and capturing the context information. Inspired by the widely used Receptive Field Block [31], we introduce the PPB to achieve this goal.

As shown in Fig. 5, the PPB consists of five branches to capture the context information of point-wise features. The first four branches with similar structures

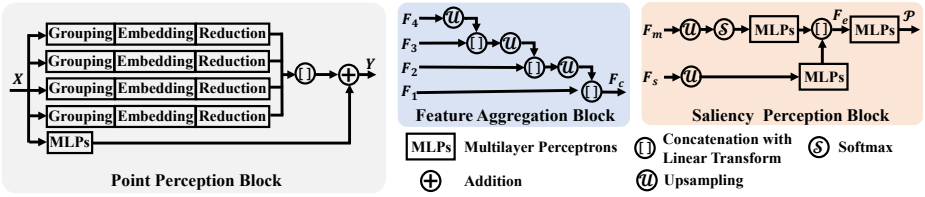

Fig. 5: Details of the components in the proposed baseline model, *i.e.*, Point Perception Block, Feature Aggregation Block, and Saliency Perception Block.

encode center points by their local regions of different sizes. Each branch has three sub-units, *i.e.*, grouping, embedding, and reduction. To be more specific, let $X^p = \{x_1^P, x_2^P, ..., x_M^p\}$ denote the spatial coordinates of input points X with intermediate learned features $X^f = \{x_1^f, x_2^f, ..., x_M^f\}$. M indicates the number of points. For each point $x_i^p \in X^p$, the grouping sub-unit gathers its k nearest neighbors $\mathcal{N}(x_i^p) = \{x_{i,1}^p, x_{i,2}^p, ..., x_{i,k}^p\}$ by K-nearest neighbours (KNN). The spatial size of the local region $\mathcal{N}(x_i^p)$ centered on x_i^p varies as k takes different values. To learn local geometric representations, the embedding sub-unit embeds the relative spatial position between x_i^p and its neighbor $x_{i,j}^p$ as

$$e_i^j = MLPs([x_i^p, x_{i,j}^p, x_i^p - x_{i,j}^p, \mathcal{D}(x_i^p, x_{i,j}^p)]), \tag{1}$$

where $\mathcal{D}(\cdot)$ and [] denote the Euclidean distance between two points and the concatenation operation, respectively. Because e_i^j merely contains the geometric features and lacks associated point-wise features, we concatenate e_i^j with corresponding point-wise features x_i^f to obtain the advanced representations a_i^j. All advanced representations $\mathcal{A}_i = \{a_i^1, a_i^2, .., a_i^k\}$ of k neighbors express each of their semantic contributions to the center point x_i^p. The reduction sub-unit aggregates the neighborhood semantic contributions by a Mean-max reduction operation

$$\hat{x}_i^f = MLPs([max(\mathcal{A}_i), mean(\mathcal{A}_i)]), \tag{2}$$

where $max(\cdot)$ and $mean(\cdot)$ denote the max function and mean function, respectively. Compared with the input features X^f, the branch outputs $\hat{X}^f = \{\hat{x}_1^f, \hat{x}_2^f, ..., \hat{x}_M^f\}$ have enlarged receptive fields and capture the context information in local regions. Finally, we fuse the output features $\{\hat{X}_b^f\}_{b=1}^4$ of the first four branches and further introduce a skip connection of the fifth branch to retain the original features

$$\hat{Y}^f = MLPs([\hat{X}_1^f, \hat{X}_2^f, \hat{X}_3^f, \hat{X}_4^f]) + MLPs(X^f). \tag{3}$$

Similar to the Receptive Field Block, by setting $K = \{k_1, k_2, k_3, k_4\}$ for the first four branches reasonably, the global semantics and multi-scale features can be obtained, respectively. Besides, the input points X and corresponding outputs Y of our PPB share the same feature size. Therefore, our PPB can be easily embedded in various networks to improve their performance.

Saliency Perception Block. The utilization of our PPB allows the acquisition of global semantics and multi-scale features. Subsequently, how to seamlessly merge the two kinds of features and obtain the final prediction is still open.

As shown in Fig. 5, our SPB enhances the multi-scale features using the global semantics. The global semantics can effectively alleviate the distraction of non-salient background in multi-scale features and emphasize the salient regions (see Fig. 8). The enhanced multi-scale features are then used to predict the salient objects. Specifically, the SPB first upsamples the global semantics F_s and multi-scale features F_m to the spatial size of the input $\mathcal{V}$. The upsampling operation is followed by MLPs to reduce the aliasing effect. Then we use the upsampled global semantics to enhance the multi-scale features

$$F_e = MLPs([MLPs(\mathcal{U}(F_s)), \mathcal{S}(MLPs(\mathcal{U}(F_m)))]), \tag{4}$$

where $\mathcal{U}$ and $\mathcal{S}$ denote the upsampling and softmax operations, respectively. F_e is the enhanced multi-scale features. In this approach, the enhanced multi-scale features include both the accurate positions and fine-grained structures of salient objects. Finally, we use a prediction layer (MLPs) to predict salient objects $\mathcal{P}$ from the enhanced multi-scale features F_e.

5 Experiments

5.1 Experimental Setup

Implementation Details. We use the popular Pytorch framework to implement our method on an NVIDIA TITAN XP GPU. The points in the inputs are represented by nine-dimensional vectors ($d_{in} = 9$) consisting of spatial coordinates, RGB colors, and normalized spatial coordinates. Due to the limitations of memory capacity, we randomly sample $N = 4,096$ points with replacement from inputs in the training stage, while the sampling operations in the testing are without replacement for testing all 240,000 points in a 3D view. We use random rotation to augment data. The parameters K of the PPB for abstracting global semantics are $\{1, 4, 9, 16\}$ while those of another PPB are $\{1, 9, 25, 49\}$. Our loss function is defined on the standard cross-entropy loss. We train the proposed baseline model by Adam optimizer with an initial learning rate of 5e-4 and a weight decay of 1e-4. The total training epochs are 3,000, with a batch size of 32. A three-time voting strategy [38] is adopted to produce the predictions in the testing phase.

Evaluation Metrics. To compare the results of different methods, we adopt four popular evaluation metrics for performance benchmarking, *i.e.*, mean absolute error (MAE), F-measure [33], E-measure [8], and intersection over union (IoU). MAE estimates the point-wise approximation degree between predicted segmentation maps and corresponding ground truths. It can be formulated as $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|p_i - g_i|$, where $p_i \in \mathcal{P}$ and $g_i \in \mathcal{G}$ are the prediction and ground truth, respectively. F-measure is the harmonic mean value of the precision

Methods	Years	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
PointNet [3]	CVPR'17	0.116	0.632	0.768	0.519
PointNet++ [38]	NeurIPS'17	0.077	0.738	0.816	0.608
PointCNN [25]	NeurIPS'18	0.142	0.409	0.575	0.265
ShellNet [59]	ICCV'19	0.074	0.753	0.848	0.648
RandLA [16]	TPAMI'21	0.127	0.633	0.740	0.517
Ours	-	**0.069**	**0.769**	**0.851**	**0.656**

Table 1: Benchmarking results of six representative baseline models on our *PCSOD* dataset. "↑"/"↓" suggests that larger/smaller is better. Note that the best results are shown in **boldface**

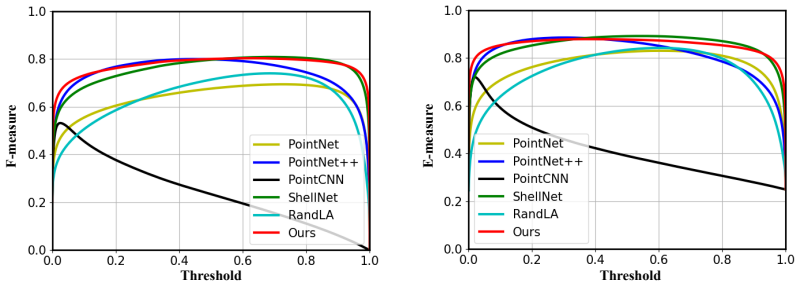

Fig. 6: F-measure and E-measure under different thresholds.

(*prec*) and recall (*reca*), *i.e.*, F-measure $= \frac{(1-\beta^2)prec \cdot reca}{\beta^2 prec + reca}$, where β^2 is set to 0.3 for emphasizing the importance of precision. E-measure captures both the local matching and region-level matching information of segmentation maps for assessment. IoU is a metric describing the extent of overlap between two segmentation maps. It is defined as IoU $= \frac{inter}{union}$, where *inter* and *union* indicate the intersection and union of two segmentation maps, respectively. Note that the relevant concepts of S-measure [7] in 3D space may change, thus being ignored.

5.2 Comparison and Analysis

To the best of our knowledge, there is no deep learning-based method designed for point cloud SOD. Consequently, we introduce five representative baseline models [3, 16, 25, 38, 59] from others segmentation tasks for comparison and analysis. PointNet [3] and its improved version, namely PointNet++ [38], are the two most representative models in point cloud processing. PointCNN [25], ShellNet [59], and RandLA [16] indicate three promising directions of point cloud processing, *i.e.*, powerful convolution, effective neighborhood connection, and advanced reduction. For a fair comparison, we retrain these models on our *PCSOD* dataset according to the recommended parameter settings and produce the final results by the same voting strategy as our method.

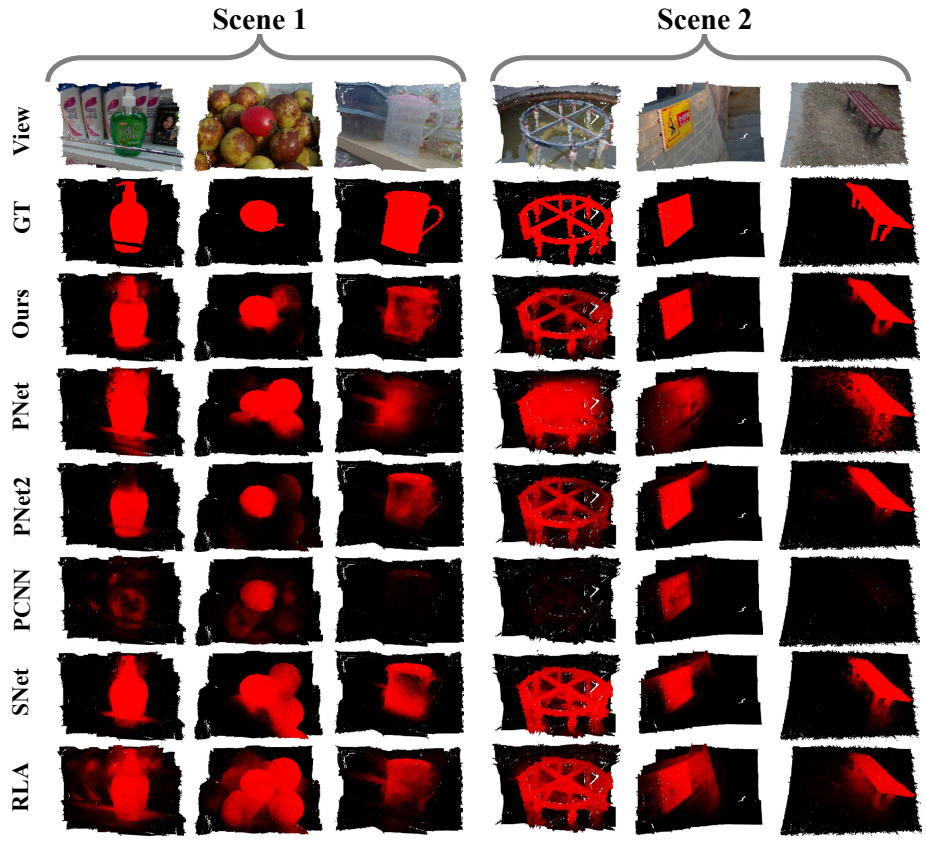

Fig. 7: Qualitative comparison of six baseline models on views of two common scenes, *i.e.*, a supermarket (Scene 1) and a park (Scene 2). Note that "GT", "PNet", "PNet2", "PCNN", "SNet", and "RLA" mean the ground truth, PointNet [3], PointNet++ [38], PointCNN [25], ShellNet [59], and RandLA [16], respectively.

Quantitative Comparison. In Tab. 1, we list the results of six baseline models on four evaluation metrics. We can learn that the proposed method achieves state-of-the-art performance and outperforms all competitors by a clear margin. Specifically, our model surpasses the second-best model ShellNet by 6.8%, 2.1%, 0.4%, and 1.2% on MAE, F-measure, E-measure, and IoU. RandLA is a recently proposed model with significant superiority over PointNet++ for semantic segmentation. However, experiments in Tab. 1 show that RandLA has no advantage on SOD and even performs worse than PointNet++, indicating designing tailored models for point cloud SOD is non-trivial. Though our baseline model has the best performance, there is still considerable room for performance improvement, which demands further efforts from the research community. To study the generalizability of these baseline models under different thresholds, we plot the F-measure scores and E-measure scores by taking different thresholds.

No.	Methods	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
1	PointNet++ [38]	0.077	0.738	0.816	0.608
2	+SPB	0.076	0.748	0.828	0.624
3	+SPB, +PPB 1	0.073	0.754	0.840	0.639
4	+SPB, +PPB 1, +PPB 2	**0.069**	**0.769**	**0.851**	**0.656**
5	Mean Reduction	0.071	0.764	0.843	0.649
6	Max Reduction	0.070	0.765	0.843	0.651
7	Attentive Reduction [16]	0.074	0.758	0.847	**0.658**
8	Mean-max Reduction	**0.069**	**0.769**	**0.851**	0.656

Table 2: Ablation analysis of the proposed point cloud SOD model. No.1-No.4 study the effectiveness of our SPB and PPB, respectively. "PPB 1" and "PPB 2" denote the PPBs for producing global semantics and multi-scale features, respectively. No.5-No.8 investigate the alternative reduction operations.

As shown in Fig. 6, the results of our method are much flatter at most thresholds, which demonstrates that our method has excellent generalizability.

Qualitative Comparison. To further reveal the feasibility of our solution predicting salient objects of any given 3D views, we illustrate the results of several frequent views from two common scenes in Fig. 7. Scene 1 is a supermarket (indoor scene), while Scene 2 is a park (outdoor scene), both of which are unseen by these models. It can be seen that most baseline models can locate the salient objects of given views, except for PointCNN. Though some views are very challenging, *e.g.*, cluttered background (column 2), transparent object (column 3), complex structure (column 4), and random view with non-central object (column 5 and 6), our method can consistently produce accurate and complete segmentation maps with high contrast, which evidences the superiority of our method.

5.3 Ablation Study

To analyze the fundamentals of our baseline model, we conduct extensive ablation experiments in Tab. 2. The ablation experiments are based on the encoder PointNet++ [38], studying the effectiveness of the designs in our decoder, *i.e.*, key modules and feature reduction operations. In each experiment, only one influential factor is changed as the others keep the same for a fair comparison.

To investigate the contributions from our SPB and PPB separately, we first load the SPB into the encoder. By comparing No.1 and No.2 in Tab. 2, we can learn that the introduction of our SPB can help promote the performance of our model in locating salient objects. However, because the high-level features from the encoder have limited receptive fields, directly utilizing them as the semantics can only achieve suboptimal performance. As demonstrated in Tab. 2 (No.3), a properly configured PPB helping acquire semantics with global receptive fields can unlock the potential of the SPB. Besides, the PPB with a different configuration can also strengthen the multi-scale representations of features,

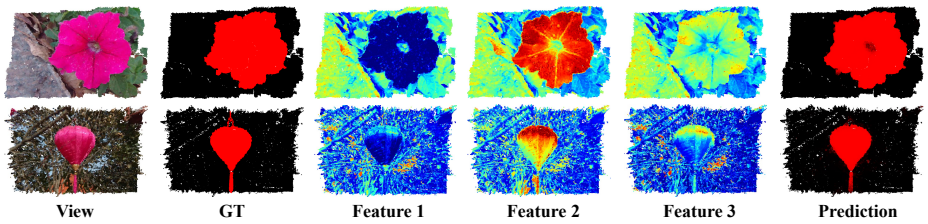

Fig. 8: 3D heatmap visualization of feature maps. Feature 1, Feature 2, and Feature 3 represent the multi-scale features, global semantics, and enhanced multi-scale features, respectively.

which benefits the perception of objects of different sizes. Therefore, another PPB in the ablation No.4 can bring orthogonal contributions to SOD. Fig. 8 further shows how the feature maps change. Due to the dilution of high-level features, multi-scale features incorrectly focus on the non-salient background, whereas the global semantics have an accurate perception of salient objects. The SPB can correct the deviation of multi-scale features by combining global semantics and obtain the enhanced multi-scale features. The ablations No.4-No.8 in Tab. 2 study various reduction manners. It can be seen that our Mean-max reduction can outperform the individual Mean reduction or Max reduction. Furthermore, compared to the attentive reduction [16], our method has a better performance without increasing the number of network parameters.

6 Conclusion

In this paper, we present the first comprehensive study on point cloud SOD, involving its formulation, dataset construction, and baseline design. To avoid the saliency conflict, we propose a novel view-dependent perspective of salient objects. Our formulation can reasonably reflect the salient objects in point cloud scenarios. Then we elaborately construct a high-quality dataset, namely *PCSOD*, and contribute a baseline model for point cloud SOD. Our dataset has excellent generalizability and broad applicability, expected to boost the advance of SOD and many other vision tasks. We conduct extensive experiments on our dataset to verify the feasibility of our solution. Experimental results show that our baseline model has significant superiority and produces visually favorable predictions. Our work reveals the potential of point cloud SOD and pave the way for further study.

Acknowledgements. This work was supported by National Key R&D Program of China (2020AAA0103501) , The Major Key Project of PCL, Natural Science Foundation of China (61801303, 62031013), Guangdong Basic and Applied Basic Research Foundation (2019A1515012031), Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003), and Shenzhen Science and Technology Plan Basic Research Project (JCYJ20190808161805519).

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1534–1543 (2016)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: IEEE/CVF International Conference on Computer Vision. pp. 9297–9307 (2019)
3. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 77–85 (2017). https://doi.org/10.1109/CVPR.2017.16
4. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 10599–10606 (2020)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
6. Ding, X., Lin, W., Chen, Z., Zhang, X.: Point cloud saliency detection by local and global feature fusion. IEEE Transactions on Image Processing **28**(11), 5379–5393 (2019). https://doi.org/10.1109/TIP.2019.2918735
7. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4548–4557 (2017)
8. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
9. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8546–8556 (2019). https://doi.org/10.1109/CVPR.2019.00875
10. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision. pp. 275–292. Springer (2020)
11. Fu, K., Jiang, Y., Ji, G.P., Zhou, T., Zhao, Q., Fan, D.P.: Light field salient object detection: A review and benchmark. arXiv preprint arXiv:2010.04968 (2020)
12. Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., Lin, W.: Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2021). https://doi.org/10.1109/TCSVT.2021.3082939
13. Guo, Y., Wang, F., Xin, J.: Point-wise saliency detection on 3d point clouds via covariance descriptors. The Visual Computer **34**(10), 1325–1338 (2018)
14. Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M.: Semantic3d. net: A new large-scale point cloud classification benchmark. arXiv preprint arXiv:1704.03847 (2017)
15. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(4), 815–828 (2019). https://doi.org/10.1109/TPAMI.2018.2815688
16. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Learning semantic segmentation of large-scale point clouds with random sampling.

IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3083288
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998). https://doi.org/10.1109/34.730558
18. Kim, G., Huber, D., Hebert, M.: Segmentation of salient regions in outdoor scenes using imagery and 3-d data. In: IEEE Workshop on Applications of Computer Vision. pp. 1–8. IEEE (2008)
19. Klokov, R., Lempitsky, V.: Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In: IEEE/CVF International Conference on Computer Vision. pp. 863–872 (2017). https://doi.org/10.1109/ICCV.2017.99
20. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: European conference on computer vision. pp. 101–115. Springer (2012)
21. Lei, H., Akhtar, N., Mian, A.: Octree guided cnn with spherical kernels for 3d point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9623–9632 (2019). https://doi.org/10.1109/CVPR.2019.00986
22. Li, C., Cong, R., Piao, Y., Xu, Q., Loy, C.C.: Rgb-d salient object detection with cross-modality modulation and selection. In: European Conference on Computer Vision. pp. 225–241. Springer (2020)
23. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(8), 1605–1616 (2017). https://doi.org/10.1109/TPAMI.2016.2610425
24. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: IEEE/CVF International Conference on Computer Vision. pp. 2976–2983 (2013). https://doi.org/10.1109/ICCV.2013.370
25. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. Advances in Neural Information Processing Systems **31** (2018)
26. Liao, G., Gao, W., Jiang, Q., Wang, R., Li, G.: Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection. In: ACM International Conference on Multimedia. pp. 2436–2444 (2020)
27. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3917–3926 (2019)
28. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 678–686 (2016). https://doi.org/10.1109/CVPR.2016.80
29. Liu, N., Zhao, W., Zhang, D., Han, J., Shao, L.: Light field saliency detection with dual local graph learning and reciprocative guidance. In: IEEE/CVF International Conference on Computer Vision. pp. 4712–4721 (2021)
30. Liu, Q., Su, H., Duanmu, Z., Liu, W., Wang, Z.: Perceptual quality assessment of colored 3d point clouds. IEEE Transactions on Visualization and Computer Graphics pp. 1–1 (2022). https://doi.org/10.1109/TVCG.2022.3167151
31. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: European conference on computer vision. pp. 385–400 (2018)
32. Ma, Y., Zhai, Y., Yang, C., Yang, J., Wang, R., Zhou, J., Li, K., Chen, Y., Wang, R.: Variable rate roi image compression optimized for visual quality. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1936–1940 (2021). https://doi.org/10.1109/CVPRW53098.2021.00221

33. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2014). https://doi.org/10.1109/CVPR.2014.39
34. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 909–918 (2019)
35. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9410–9419 (2020). https://doi.org/10.1109/CVPR42600.2020.00943
36. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 733–740 (2012). https://doi.org/10.1109/CVPR.2012.6247743
37. Piao, Y., Rong, Z., Zhang, M., Lu, H.: Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 11865–11873 (2020)
38. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
39. Qian, G., Hammoud, H., Li, G., Thabet, A., Ghanem, B.: Assanet: An anisotropic separable set abstraction for efficient point cloud representation learning. Advances in Neural Information Processing Systems **34** (2021)
40. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7471–7481 (2019). https://doi.org/10.1109/CVPR.2019.00766
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
42. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–779 (2019). https://doi.org/10.1109/CVPR.2019.00086
43. Shtrom, E., Leifman, G., Tal, A.: Saliency detection in large point sets. In: IEEE/CVF International Conference on Computer Vision. pp. 3591–3598 (2013). https://doi.org/10.1109/ICCV.2013.446
44. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Inferring attention shift ranks of objects for image saliency. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12130–12140 (2020). https://doi.org/10.1109/CVPR42600.2020.01215
45. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Scene context-aware salient object detection. In: IEEE/CVF International Conference on Computer Vision. pp. 4156–4166 (2021)
46. Tasse, F.P., Kosinka, J., Dodgson, N.: Cluster-based point set saliency. In: IEEE/CVF International Conference on Computer Vision. pp. 163–171 (2015). https://doi.org/10.1109/ICCV.2015.27
47. Tu, W.C., He, S., Yang, Q., Chien, S.Y.: Real-time salient object detection with a minimum spanning tree. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2334–2342 (2016). https://doi.org/10.1109/CVPR.2016.256

48. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3054719
49. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3796–3805 (2017). https://doi.org/10.1109/CVPR.2017.404
50. Wang, T., Piao, Y., Lu, H., Li, X., Zhang, L.: Deep learning for light field saliency detection. In: IEEE/CVF International Conference on Computer Vision. pp. 8837–8847 (2019). https://doi.org/10.1109/ICCV.2019.00893
51. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3051099
52. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions On Graphics **38**(5), 1–12 (2019)
53. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: European Conference on Computer Vision. pp. 29–42. Springer (2012)
54. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9613–9622 (2019). https://doi.org/10.1109/CVPR.2019.00985
55. Yang, S., Lin, W., Lin, G., Jiang, Q., Liu, Z.: Progressive self-guided loss for salient object detection. IEEE Transactions on Image Processing **30**, 8426–8438 (2021). https://doi.org/10.1109/TIP.2021.3113794
56. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8579–8588 (2020). https://doi.org/10.1109/CVPR42600.2020.00861
57. Zhang, J., Liu, Y., Zhang, S., Poppe, R., Wang, M.: Light field saliency detection with deep convolutional networks. IEEE Transactions on Image Processing **29**, 4421–4434 (2020). https://doi.org/10.1109/TIP.2020.2970529
58. Zhang, M., Fei, S.X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: European Conference on Computer Vision. pp. 374–390. Springer (2020)
59. Zhang, Z., Hua, B.S., Yeung, S.K.: Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In: IEEE/CVF International Conference on Computer Vision. pp. 1607–1616 (2019). https://doi.org/10.1109/ICCV.2019.00169
60. Zhao, J., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: IEEE/CVF International Conference on Computer Vision. pp. 8778–8787 (2019). https://doi.org/10.1109/ICCV.2019.00887
61. Zheng, T., Chen, C., Yuan, J., Li, B., Ren, K.: Pointcloud saliency maps. In: IEEE/CVF International Conference on Computer Vision. pp. 1598–1606 (2019). https://doi.org/10.1109/ICCV.2019.00168
62. Zhou, T., Fan, D.P., Cheng, M.M., Shen, J., Shao, L.: Rgb-d salient object detection: A survey. Computational Visual Media **7**(1), 37–69 (2021)