MSV-PCT: Multi-Sparse-View Enhanced Transformer Framework for Salient Object Detection in Point Clouds

Zihao Wang[1], Yiming Huang[1], Gengyu Lyu[1], Yucheng Zhao[1], Ziyu Zhou[1], Bochen Xie[2], Zhen Yang[1], Yongjian Deng[1]*

[1]College of Computer Science, Beijing University of Technology
[2]Department of Mechanical Engineering, City University of Hong Kong
rex.wangzihao@gmail.com,huangyiming2002@126.com,
lyugengyu@bjut.edu.cn,yzhao836@gatech.com, zzhou651@connect.hkustgz.edu.cn,
boxie4-c@my.cityu.edu.hk,yangzhen@bjut.edu.cn, yjdeng@bjut.edu.cn

Abstract

Salient object detection (SOD) methods for 2D images have great significance in the field of human-computer interaction (HCI). However, as a common data format in HCI, the SOD research in the form of 3D point cloud data remains limited. Previous works commonly treat this task as point cloud segmentation, which perceives all points in the scene for prediction. However, these methods neglect that SOD is designed to simulate human visual perception where human can only see the surfaces rather than occluded point clouds. Thereby, these methods may fail when meet such situations. This paper aims to solve this problem by approximately simulating the perception paradigm of humans towards 3D scenes. Thus, we propose a framework based on the 3D visual point cloud backbone and its multi-view projection named MSV-PCT. Specifically, instead of relying solely on general point cloud learning frameworks, we additionally introduce multi-sparse-view learning branches to supplement the SOD perception. Furthermore, we propose a novel point cloud edge detection loss function to effectively address artifacts, enabling the accurate segmentation of the edges of salient objects from the background. Finally, to evaluate the generalization of point cloud SOD methods, we introduce a new approach to generate simulated PC-SOD datasets from RGBD-SOD data. Experiments on the simulated datasets show that MSV-PCT achieves better accuracy and robustness.

Introduction

Salient object detection (SOD) aims to precisely identify and segment visually distinctive regions (Borji et al. 2019). These regions typically attract human attention due to their contrast with the rest of the image or scene. As an important image preprocessing step, salient object detection is widely used in various computer vision tasks, such as action recognition (Abdulmunem, Lai, and Sun 2016), semantic segmentation (Ghariba, Shehata, and McGuire 2022), and image retrieval (Ma et al. 2021). Most of the existing SOD methods are applied to salient objects on conventional images or RGBD images, with relatively less work on point cloud data.

The powerful spatial modeling capability of LiDAR devices and the rapid growth in the deployment of devices equipped with LiDAR have led to a significant improvement in the availability of point cloud data. These 3D capture sensors not only replicate the geometric perception similar to the human binocular vision system but also maintain reliable semantic accuracy in both high and low light conditions. Unlike image-based methods, point clouds record the precise 3D structure of objects without interference from external conditions such as lighting. Although the sparsity of point clouds makes semantics discontinuous, the original 3D information contained avoids the computational resource loss and feature errors associated with dense rendering and extensive post-processing in image-based methods. This provides more accurate semantic information for point cloud salient detection. Unlike objective factual judgments, SOD emphasizes human subjectivity, simulating human visual perception and attention to a scene to better fit practical application scenarios. However, existing point cloud SOD methods generally follow the design principles of point cloud deep learning networks, assuming that observers have an unidirectional perspective, enabling them to see all points within the scene simultaneously and make salient judgments. This assumption overlooks the limitations of human binocular vision in actual observation, leading to shifts and artifacts in the model's prediction results.

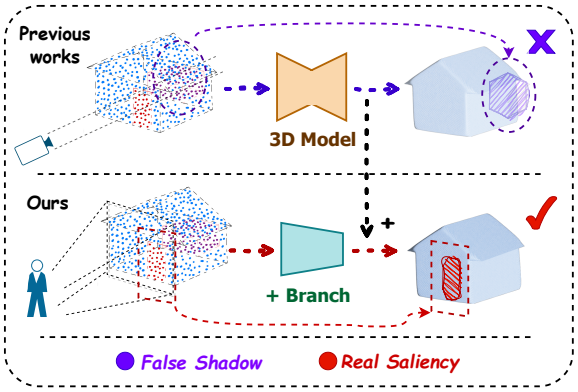

Figure 1: Illustration of our motivation.

To overcome these drawbacks, we designed a **m**ulti-**s**parse-**v**iew (MSV) feature fusion approach inspired by hu-

*Corresponding Author

man visual perception. As Fig. 1 shows, instead of rendering and then projecting, we simulate human visual attention by projecting point clouds onto different views, with each view representing a distinct simulation of human perception from various angles. By ignoring occlusion points within the human field of view, our model focuses on accessible and relevant features, mirroring how humans naturally disregard occlusions and prioritize visible information. By extracting and fusing features from these multiple views, a more comprehensive representation of the 3D scene and its salient objects is achieved. This human-inspired approach improves the accuracy of SOD, aligning the detection process with the natural way humans perceive and prioritize visual information.

Thereby, we propose a new framework for 3D Point cloud SOD with multi-view augmentation, dubbed PSOD-Net (Gao et al. 2022). To improve the robustness of saliency detection, we propose a point cloud multi-view branch framework for aggregating the features of points in different views to emphasize saliency viewpoints from different perspectives. To address the issue of poor edge decision-making for salient objects caused by semantic inconsistency in point clouds and thus eliminate artifacts in point cloud prediction target edges, we propose a saliency detection loss function based on point cloud edge loss, which calculates the spatial weight of each point in the point cloud to determine the edges of salient objects. And a point cloud SOD simulation dataset was proposed to verify the generalization of point cloud SOD. The entire network can be trained end-to-end and has advantages over existing point cloud SOD methods. Our main contributions are threefold:

1. We propose a new point cloud SOD framework with multi-view enhancement, which can effectively utilize the geometry information carried in the point cloud.
2. We propose a salient detection loss function for point clouds based on edge loss, which effectively diminishes artifacts in the prediction outcomes.
3. We convert several RGBD saliency detection datasets into point clouds, creating a simulated point cloud dataset SOD-RGBD2PC, along with annotations to assess model generalization.

Related Work

Salient Object Detection

Early attempts to perform salient object detection employ manually create features to exploit low-level cues (Song et al. 2023). Hou et al. introduce the integration of short connections within a skip-layer architecture, enabling the full utilization of advanced representations across multiple layers (Hou et al. 2017). Siris et al. propose a framework that is contextually aware of semantic scenes to effectively capture high-level semantics for identifying prominent objects (Siris et al. 2021). Recent studies explicitly extract overarching semantics and integrate them with low-level features, resulting in noticeable enhancements in performance (Wang et al. 2022). The current RGB image-based methods also have achieved satisfactory results (Wang et al. 2021). Despite their challenges with complex scenes due to a lack of spatial geometry information, we designed our framework to address the deficiency in human subjectivity in current 3D Salient Object Detection. Additionally, we converted RGB-D datasets to this format for thorough testing and further validation of our framework.

3D Transformers

Transformers have made significant advancements in sequence modeling (Zhou et al. 2024), which attracted the attention of the computer vision area. Following the emergence of the vision transformer (Kolesnikov et al. 2021), Transformers have also been applied to numerous tasks in the 3D vision area (Lahoud et al. 2022). Zhao et al. (Zhao et al. 2021) utilize a transformer block to extract local features, employing self-attention within the local region of each input point. Misra et al. (Misra, Girdhar, and Joulin 2021) propose a 3D end-to-end detector, wherein a transformer encoder directly extracts features from the point cloud and a transformer decoder predicts bounding boxes. Mao et al. (Mao et al. 2021) present a transformer-based framework to capture distant relationships among voxels. Recently, Zhang et al. (Zhang et al. 2023) propose Enhanced Point Feature Network for point cloud Salient Object Detection (SOD) by combining image features with the point cloud. Building on previous findings, we recognize that transformer architectures, despite their complexity, effectively mine comprehensive objective information. To address the deficiency in subjective saliency, a simpler MLP-based architecture suffices for the transformer backbone of our framework.

Multi-View Methods on Point Clouds

The concept of utilizing 2D images to interpret the 3D scene was first introduced by Bradski in 1994 (Bradski and Grossberg 1994). This intuitive multi-view approach was integrated with deep learning techniques to enhance 3D understanding, as seen in MVCNN (Su et al. 2015). Subsequent research has built upon this foundation, refining the aggregation of features from different image views for tasks such as classification and retrieval. Notable contributions in this domain advance the state of the art in multi-view feature aggregation (Kanezaki, Matsushita, and Nishida 2018) (Esteves et al. 2019) (Cohen and Welling 2016) (Wei, Yu, and Sun 2020) (Hamdi, Giancola, and Ghanem 2021). Our research integrates the concept of multi-view directly into the 3D point cloud structure, which carries out multi-view projection on the point cloud and retains the format of the point cloud for further feature extraction, so as to retain the original information of the point cloud as much as possible, and fuse the multi-view features in the form of point cloud, which facilitates view-based learning for point cloud.

Methodology

Overall Architecture

Considering our reflection that semantics from human subjectivity will be beneficial to the SOD tasks, we utilize customed human views to construct our **M**ulti-**S**parse-**V**iew enhanced **P**oint **C**loud **T**ransformer —**MSV-PCT** framework

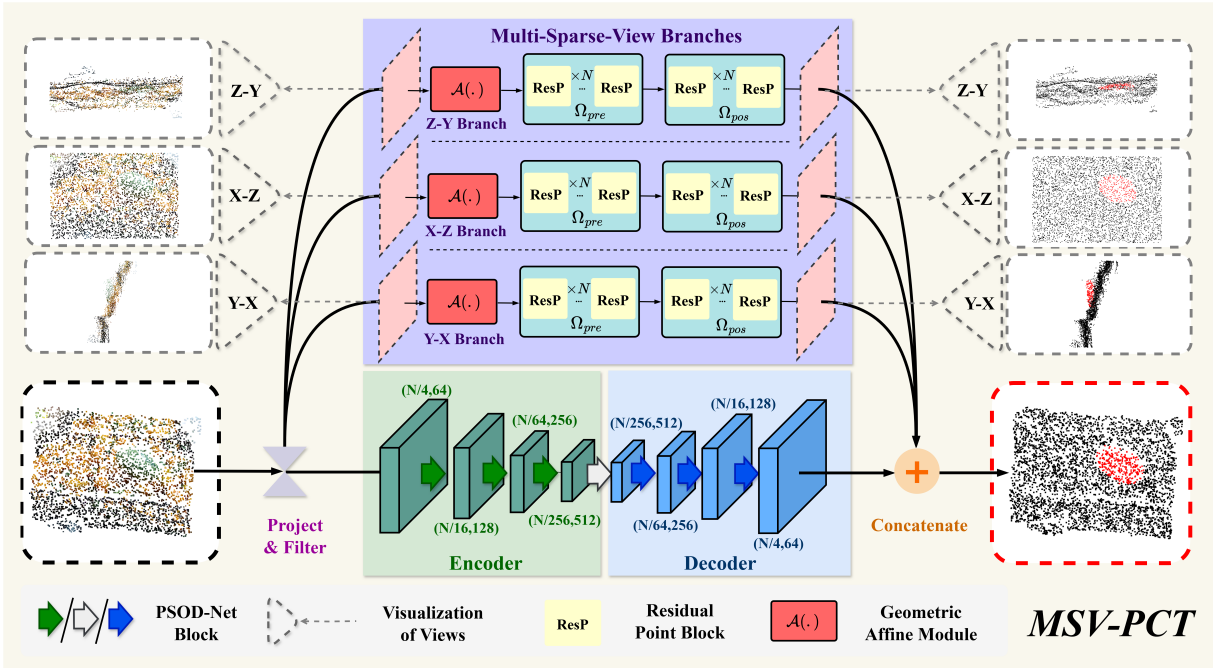

Figure 2: Detailed design of the **M**ulti-**S**parse-**V**iew enhanced **P**oint **C**loud **T**ransformer —**MSV-PCT** framework we proposed.

as Fig. 2 shown. In this framework, the point cloud data is initially projected to discrete point set in the planes in respective views. These views are the most common three views in human daily perception system, which are the front view, the side view, and the top view. After projection enhances the human subjective semantics, these directional discrete point set in the planes are sent to their corresponding branch. These three branches are designed to the same MLP-based architecture, which efficiently helps the framework strengthen human-centered 3D comprehensiveness with few parameters. Meanwhile, to digest the main patterns in conventional 3D point cloud data, we adopt the typical design of point cloud transformers as the backbone. In the end, we provide the fusion block in our framework to concatenate enhanced multi-view semantics and main patterns from 3D point cloud data. In this way, we suppose human-comprehensive flaws of conventional 3D patterns will be compensated for better performance in the SOD tasks. Moreover, we also present the elaborate matching loss to promote the learning process of the framework. The Loss clearly demonstrates the edge loss in more explicit human perceived saliency.

Projection for Multi-Sparse-View

Projection is the most significant procedure for generating semantics about saliency. Saliency is precepted by human perspective and we defined 'salient' from the surface of the scene, not the inside. In other words, the all-round, unidirectional, and thorough machine-received 3D point cloud downplays the saliency in human eyes. That is the fundamental reason why previous plain point-cloud-based works like PSOD-Net (Gao et al. 2022) are hard to fit labels containing strong human semantics. Therefore, we decide to augment human subjectivity from the semantics of human-centered projected views. Meanwhile, our projection gets rid of the rendering delay as the previous projection methods (Schütz, Kerbl, and Wimmer 2021) have.

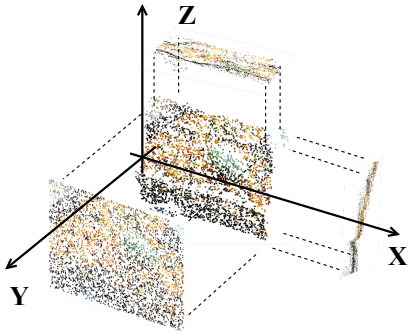

Figure 3: The project operation for discrete points.

As this design principle demonstrated, we believe the projection angles must selected for the best human comprehensiveness in daily life visual perceptional systems. Naturally, it comes to the three-view orthographic projection — the front view, the side view, and the top view. To be specific, consider a 3D point set $P = \{p_i\}^N$ (N is point number) with $p_i = (x_i, y_i, z_i, r_i, b_i, g_i)$ includes three coordinates values and three color values and a plane with its analytic expression $Ax + By + Cz + D = 0$, and the convert is expected to give a discrete point set $Q = \{q_i\}_N$ with $q_i = (u_i, v_i, r_i, b_i, g_i)$ represents point with its color values in projected 2D plane of three views, u_i, v_i is the corre-

sponding coordinates values. This projection is released as follows:

$$x_i' = x_i - \frac{A(Ax_i + By_i + Cz_i + D)}{|N|^2}, \quad (1)$$

$$y_i' = y_i - \frac{B(Ax_i + By_i + Cz_i + D)}{|N|^2}, \quad (2)$$

$$z_i' = z_i - \frac{C(Ax_i + By_i + Cz_i + D)}{|N|^2}. \quad (3)$$

Here, u_i, v_i are instanced from two of three x_i', y_i', z_i' in accordance with the plane. The projection process is illustrated in Fig. 3. For each point $q_i \in Q$, we define a circular filter with radius R and the filter center is (u_i, v_i). The set of points within this filter region is formulated as:

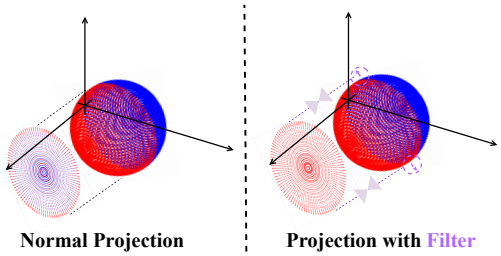

Figure 4: Our projection with filter.

$$N_i = \{q_j \in Q \mid \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \leq R\}. \quad (4)$$

The initial average depth $\bar{z}_i$ is:

$$\bar{z}_i = \frac{1}{|N_i|} \sum_{q_j \in N_i} z_j. \quad (5)$$

Set a dynamic threshold factor α, the threshold T_i is: $T_i = \alpha \bar{z}_i$, as shown in Fig. 4. For each point q_i, if its depth z_i meets the following condition, it is removed as a background point: $z_i >= T_i$. the coincident bottom background points. After this operation, the surface of the 3D scene will be projected into the plane of view we are familiar with, as Fig. 3 shows. Thus, we obtain a projected discrete point set in the planes to enhance the human semantics for the framework.

Multi-Sparse-View Branches

We suggest overabundant parameters are needless for multi-view branches because the multi-view information is comparatively supplementary to usual 3D point information processed by the powerful transformer-based backbone. As this idea goes, we adopt PointMLP (Ma et al. 2022) networks as branches, because of its simplicity and efficiency. The core module process functions as the equation below describes:

$$g_i = \Omega_{pos}(\mathcal{A}(\Omega_{pre}(f_{i,j}))), \; j = 1, 2, 3, ..., K, \quad (6)$$

where g_i is the processed output and $f_{i,j}$ is the j-th of the neighbor point feature obtained by plain K-Nearest-Neighbor (KNN) algorithm of i-th sampled point 2D-projection input, Ω_{pre} and Ω_{pos} is simple residual point MLP structured block after the geometric affine module $\mathcal{A}$.

In the above geometric affine module $\mathcal{A}$ of PointMLP, the following mapping operation helps pooling of saliency:

$$k_{i,j}' = \frac{\alpha \odot k_{i,j} - k_i}{\gamma + \theta} + \beta. \quad (7)$$

In this operation, $k_{i,j}$ is the output of Ω_{pre}, k_i is outcomes among each corresponding $k_{i,j}$, α and β are learnable parameters, γ is feature deviation across all groups and channels, θ is common minimum constants for numerical stability, $\odot$ is element-wise multiplication.

The effectiveness of this module lies in that geometric module $\mathcal{A}$ sparsely pre-aggregates saliency in 2D-human-viewed information and resulting simple residual point blocks Ω_{pre} & Ω_{pos} easily capture patterns about saliency from human subjectivity.

Backbone

In this study, we use PSOD-Net, which is composed of point context transformer (PCT) and scene context transformer (SCT), as the backbone model. As a three-dimensional point cloud segmentation model, the backbone has strong ability of context information modeling and efficient feature extraction. Specifically, we use PSOD-Net to extract multi-level semantic features from 3D point cloud data, and use these multi-source features for the final salient target detection.

Point Cloud Edge Detection Loss

In SOD, when the object and background region have complex contours, the model is often difficult to accurately detect the edge of the salient target. Background points may be incorrectly classified as significant objects, which will have a negative impact on the prediction results. Specifically, because the sensor is affected by the acquisition angle, object occlusion relationship and distance during the acquisition process, the point cloud data inside the object is relatively dense, while the point cloud data in the junction area between objects is relatively sparse, which makes it difficult for the model to balance the foreground object and background points in feature extraction. In order to meet these challenges, we design a loss function that emphasizes the edge of salient objects, and enhance the prediction accuracy of the boundary region between salient objects and background by using edge information. Our proposed loss function integrates edge detection and weight computation into the learning process, enabling the model to focus on important boundary regions within the point cloud.

The edge detection function identifies boundary points within the point cloud based on pairwise distances between points. For each point cloud in the batch, we compute the Euclidean distances between all pairs of points. Points are considered to belong to the inside of the object if their distance is below a predefined threshold τ. This results in an edge matrix E, where connected points are indicated by binary values:

$$E_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \tau \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where d_{ij} is the Euclidean distance between points i and j.

Weight Aggregation. To assign importance to each point, we compute Gaussian weights based on the distances between points. These weights reflect the similarity between points, with closer points receiving higher weights. The Gaussian weights W_{ij} are computed as:

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \tag{9}$$

where σ is a predefined scaling parameter. The Gaussian weights are combined with the edge information to produce the final weights w_i for each point. The weights for each point are computed by summing the combined Gaussian and edge weights:

$$w_i = \sum_j E_{ij} W_{ij}. \tag{10}$$

Weighted Cross Entropy Loss. The loss function integrates the computed weights into the cross entropy loss. This weighted loss ensures that points near the edges, which are typically more critical for distinguishing salient objects from the background, contribute more significantly to the overall loss. The weighted cross entropy loss $\mathcal{L}_{weighted}$ is:

$$\mathcal{L}_{\text{weighted}} = \tag{11}$$
$$-\frac{1}{B \cdot N} \sum_{b=1}^{B} \sum_{i=1}^{N} w_i \left[t_i \log(o_i) + (1 - t_i) \log(1 - o_i) \right],$$

where B is the batch size, N is the number of points, w_i is the weight for point *i*, t_i is the target label, and o_i is the predicted output.

By incorporating edge detection and weight computation into the training process, our method effectively enhances the model's capability to identify and focus on salient regions within complex point clouds. This edge-aware approach ensures that the model can better differentiate between foreground objects and the background, resulting in improved performance in salient object detection tasks.

SOD-RGBD2PC Dataset

To demonstrate the generalization of point cloud saliency detection methods, it is necessary to validate them on a broader dataset. The number of existing point cloud object saliency detection datasets is very limited. To the best of the author's knowledge, there is only one point cloud object saliency detection dataset available, namely the PCSOD dataset. To address this issue, this article proposes a SOD-RGBD2PC method that converts RGB-D saliency detection datasets or simple RGB-D datasets into point cloud saliency object detection datasets. Converting RGBD dataset to point cloud dataset mainly involves converting each pixel in the depth image into a point in three-dimensional space and associating it with the color information of the corresponding pixel in the RGB image. We need an inherent parameter matrix for RGB images, depth images, and cameras. RGB images contain color information for each pixel, while depth images contain distance values for each pixel, representing the distance from a point to the camera. The intrinsic parameter matrix contains the internal parameters of the camera, including focal length and optical center position.

Each pixel in the image is converted into three-dimensional coordinates in the camera coordinate system based on the camera's intrinsic parameter matrix. For each pixel in the depth image, its corresponding three-dimensional coordinates can be calculated as follows:

$$\begin{aligned} X &= \frac{(x - c_x) \cdot Z}{f_x}, \\ Y &= \frac{(y - c_y) \cdot Z}{f_y}, \\ Z &= \text{Depth}(x, y), \end{aligned} \tag{12}$$

f_x and f_y are the focal lengths of the camera. c_x and c_y are the coordinates of the optical center on the image. $Depth(x, y)$ is the depth value of $d(x, y)$ in the depth image. Associate the color information of corresponding pixels in RGB images with the generated three-dimensional point coordinates to form a point cloud dataset containing color information. Therefore, each point contains three-dimensional coordinates and corresponding color information.

Experiments

Experimental Setup

Implementation. We implement our model using pytorch on six NVIDIA RTX 3090 GPUs. The point cloud includes 9-d features composed of spatial coordinates, RGB colors and normalized coordinates. We randomly divide the entire 3D view into 4096 patches as the input. We use the Adam optimizer to train the model end-to-end. The total training cycle is 800, and the initial learning rate is 5e-4.

Baselines. Current methods can be sorted into two categories: 1) **General Point-Cloud-Used**, which mainly are traditional Point Cloud models — ASSANet (Qian et al. 2021), PointMLP (Ma et al. 2022), PointNeXt (Qian et al. 2022), ShellNet (Zhang, Hua, and Yeung 2019), PointTransformer (Zhao et al. 2021), PCT (Guo et al. 2021); 2) **3D-Saliency-Specialized**: PointSal (Fan, Gao, and Li 2022) and PSOD-Net (Wei et al. 2024).

Evaluation Metrics. This paper adopts four commonly acknowledged metrics to evaluate the proposed MSV-PCT. They are 1) MAE, Mean Absolute Error, demonstrates average absolute value difference; 2) F-measure, comprehensively considers precision and recall; 3) E-measure, enhanced F-measure for saliency problem; 4) IoU, Intersection over Union, judges accuracy of saliency selection.

Experiment on PCSOD Dataset

PCSOD Dataset. PCSOD (Point Cloud Salient Object Detection) dataset is a pioneering dataset tailored for the task of salient object detection in point clouds. Unlike traditional image-based SOD datasets, PCSOD focuses on the unique challenges posed by 3D point clouds, such as the attention shift phenomenon, where objects can be perceived as both salient and non-salient from different viewpoints. This

Methods	NPLR				RGB-D Scenes v2			
	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
ASSANet	0.095	0.710	0.822	0.605	0.098	0.705	0.815	0.600
PointTransformer	0.079	0.764	0.840	0.675	0.084	0.758	0.846	0.668
PCT	0.074	0.770	0.853	0.652	0.076	0.776	0.848	0.650
PointMLP	0.066	0.792	0.878	0.704	0.070	0.785	0.872	0.698
PointNeXt	0.071	0.780	0.860	0.683	0.073	0.785	0.865	0.680
ShellNet	0.078	0.755	0.845	0.645	0.081	0.750	0.842	0.640
PointSal	0.070	0.768	0.850	0.658	0.073	0.765	0.853	0.656
PSOD-Net	0.063	0.800	0.878	0.707	0.065	0.805	0.880	0.710
MSV-PCT (Ours)	**0.056**	**0.846**	**0.912**	**0.760**	**0.054**	**0.850**	**0.915**	**0.765**

Table 1: Performance of our MSV-PCT and baselines on simulation converted datasets NPLR and RGB-D ScenesV2.

dataset comprises 2,872 annotated 3D views from diverse indoor and outdoor scenes, each annotated with hierarchical labels, including super- and sub-classes, bounding boxes, and segmentation maps. The comprehensive annotations and diverse object categories ensure the dataset's applicability across various vision tasks, promoting robust generalizability and facilitating multi-task learning. PCSOD sets a new benchmark for point cloud SOD.

Comparsion on PCSOD. With prepared evaluation settings, we have conducted performance experiments in the PCSOD (Fan, Gao, and Li 2022) dataset and we have presented outcomes in Tab. 2. From the quantitative outcomes, our framework outperforms all the approaches on all the metrics and achieves a 5% boost on IoU compared to the previous SOTA PSOD-Net. To be specific, the results indicate that the 3D-Saliency-Specialized methods undoubtedly perform better than General Point-Cloud-Used methods, and our method promote this trend by adding human-comprehensive multi-view information which enhances its performance than PSOD-Net.

Methods	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
ASSANet	0.089	0.709	0.814	0.606
Point Transformer	0.075	0.762	0.848	0.670
PCT	0.069	0.770	0.846	0.652
PointMLP	0.065	0.792	0.875	0.702
PointNeXt	0.066	0.779	0.859	0.680
ShellNet	0.074	0.753	0.848	0.648
PointSal	0.069	0.769	0.851	0.656
PSOD-Net	0.058	0.805	0.878	0.711
MSV-PCT (Ours)	**0.052**	**0.845**	**0.908**	**0.763**

Table 2: Performance of our MSV-PCT framework and baselines on PCSOD dataset.

Experiments on SOD-RGBD2PC Dataset

To validate and demonstrate the effectiveness of the SOD-RGBD2PC method, we selected two common RGB-D datasets for transformation: NPLR (Peng et al. 2014) and RGB-D Scenes v2. The NPLR dataset for salient object detection consists of 1000 image pairs captured by a standard Microsoft Kinect with a resolution of 640×480. The RGB-D scene dataset v2 consists of 14 scenes, including furniture (chairs, coffee tables, sofas, tables) and a subset of objects in the RGB-D object dataset (bowls, lids, cereal boxes, coffee cups, and soda cans). However, due to the fact that not all 14 scenes in the RGB-D scene v2 dataset are suitable for salient detection tasks, 4 scenes with salient objects were selected. Scenarios 2, 4, and 13 are used as training sets, while Scenario 1 is used as the testing set. There are labels for salient object detection in the NPLR dataset, but not in the RGB-D scene v2 dataset.

To overcome this issue, we first use the complete pre-trained D3Net (Chen et al. 2021) architecture to perform salient detection on the RGB-D Scenes v2 dataset, and use the detection results as ground truth. Then, we convert these results into point cloud format. Through this approach, we are able to generate high-quality datasets for point cloud salient detection from the RGB-D dataset.

Comparison on SOD-RGBD2PC Dataset. We also accomplish similar experiments on our converted RGB-D datasets, the results are shown in Tab. 1. Based on numerical results, our framework successfully beats all other models on all metrics, which further proves the validity of our multi-view strategy and the 3D saliency superiority of our proposed framework. In addition, the saliency-specialized method and our plus edition are the leading positions in this comparison. This phenomenon demonstrates the necessity of specialized human-subjectivity multi-view information.

Ablation Study

Effectiveness of Multi-Sparse-Views Designation. We truly wonder whether multi-view information indeed helps SOD or not. Hence, we regard the most natural and the best way to verify this issue is that **gradually add different one single view, two views, and three views of the given projection outcome**. We execute this ablation study and depict the comparison of the different ablation settings in Tab. 3. Overall, the outcome indicates the escalation of views steadily increases the performance of our framework. In other words, **More Views, More Powerful!** Besides that, the outcome also illustrates the front view is the most pow-

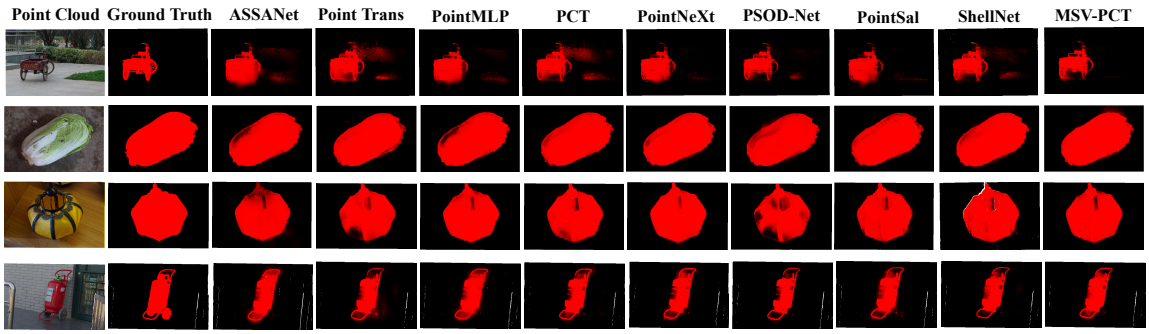

Figure 5: Visualization of the SOD results. The examples shown in the figure are randomly selected from the PCSOD dataset. Compared to other models, our proposed MSV-PCT can more accurately identify salient features in the point cloud.

Methods	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
MV-PCT + (**XY**)	0.055	0.813	0.899	0.719
+ (**YZ**)	0.060	0.799	0.889	0.698
+ (**XZ**)	0.056	0.805	0.892	0.715
+ (**XY+YZ**)	0.054	0.813	0.902	0.702
+ (**XY+XZ**)	0.055	0.818	0.900	0.730
+ (**YZ+XZ**)	0.056	0.823	0.893	0.724
+ (**XY+YZ+XZ**)	**0.052**	**0.845**	**0.908**	**0.763**

Table 3: Performance of MSV-PCT framework with different multi-view settings on PCSOD Dataset. **XY** represents the front view, **YZ** represents the top view, **XZ** represents the side view, and **+** means the combination of the views.

erful augment for our framework compared to different ablation settings. Since the front view is the most human-subjective among the three views, it is evident that subjectivity contained by human views do help the SOD task. In all, this ablation study sufficiently validates the effectiveness of our multi-view enhancing strategy.

The key issue of our proposed framework is that subjective multi-sparse-view information helps SOD tasks. We regard the most natural and the best way to verify this issue is that **gradually add different one single view, two views, and three views of the given projection outcome**.

Effectiveness of the Weighted Cross Entropy Loss. Besides verifying the multi-view strategy, we also carry out the ablation experiment to test the effectiveness of our proposed loss. The result is shown in Tab. 4. On the one hand, our model still outperforms the previous SOTA PSOD-Net without optimized by our proposed loss. On the other hand, the ascending performance with our loss proves that our loss indeed promotes the ability in the 3D saliency task.

Visualization Analysis

Beyond the numerical results presented by the ablation study above, we also provide visualizations for analyzing the potency of our method. We visualize several randomly selected saliency detection results produced by our method with others to compare. As Fig. 5 shows, severe false shadows and errors occur in previous methods while ours gives the closest

Methods	MAE ↓	F-measure ↑	E-measure ↑	IoU ↑
PSOD-Net	0.058	0.805	0.878	0.711
MV-PCT (**w/o loss**)	0.054	0.823	0.882	0.736
MV-PCT	**0.052**	**0.845**	**0.908**	**0.763**

Table 4: Performance of MSV-PCT framework with different loss settings on PCSOD Dataset.

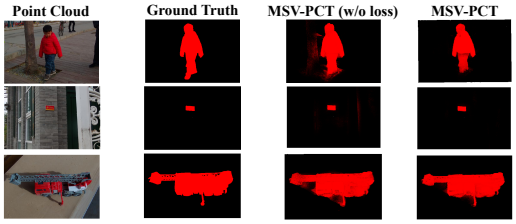

Figure 6: Visualization of the loss ablation.

result to the ground truth. Moreover, this relatively whole result also illustrates that our framework optimized by our loss (shown in Fig. 6) has reached a kind of high-level saliency with the corresponding explicit.

Conclusion

This paper present **MSV-PCT** for 3D point cloud salient object detection task with a multi-view enhanced strategy, which efficiently promotes the performance. And we also propose the matching loss to sharpen this framework. Additionally, we convert RGB-D data to point cloud saliency form for stressing and widening the 3D-saliency research. Proved by empirical experiments, our framework achieves the new SOTA in these three datasets (PCSOD dataset and two converted datasets). Furthermore, the ablation study and visualization deepen the discussion of the necessity for utilizing more views. All experimental results point out human subjectivity conveyed by multi-view information indeed helps 3D-saliency tasks. These issues will provide constructive direction for further study in 3D saliency solutions.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (62203024, 92167102, 61873220, 62102083, 62173286, 61875068, 62177018, 62306020), the Natural Science Foundation of Jiangsu Province (BK20210222), the R&D Program of Beijing Municipal Education Commission (KM202310005027), the Research Grants Council of Hong Kong (CityU11206122) and the Young Elite Scientist Sponsorship Program by BAST (BYESS2024199).

References

Abdulmunem, A.; Lai, Y.-K.; and Sun, X. 2016. Saliency guided local and global descriptors for effective action recognition. *Computational Visual Media*, 2: 97–106.

Borji, A.; Cheng, M.-M.; Hou, Q.; Jiang, H.; and Li, J. 2019. Salient object detection: A survey. *Computational visual media*, 5: 117–150.

Bradski, G.; and Grossberg, S. 1994. Recognition of 3-D Objects from Multiple 2-D Views by a Self-Organizing Neural Architecture. In *From Statistics to Neural Networks*, 349–375.

Chen, D. Z.; Wu, Q.; Nießner, M.; and Chang, A. X. 2021. D3Net: A Speaker-Listener Architecture for Semi-supervised Dense Captioning and Visual Grounding in RGB-D Scans. arXiv:2112.01551.

Cohen, T.; and Welling, M. 2016. Group equivariant convolutional networks. In *International conference on machine learning*, 2990–2999. PMLR.

Esteves, C.; Xu, Y.; Allen-Blanchette, C.; and Daniilidis, K. 2019. Equivariant multi-view networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1568–1577.

Fan, S.; Gao, W.; and Li, G. 2022. Salient object detection for point clouds. In *European conference on computer vision*, 1–19. Springer.

Gao, S.; Zhang, W.; Wang, Y.; Guo, Q.; Zhang, C.; He, Y.; and Zhang, W. 2022. Weakly-Supervised Salient Object Detection Using Point Supervision. In *AAAI*.

Ghariba, B.; Shehata, M. S.; and McGuire, P. 2022. Salient object detection using semantic segmentation technique. *International Journal of Computational Vision and Robotics*, 12(1): 17–38.

Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; and Hu, S.-M. 2021. PCT: Point cloud transformer. *Computational Visual Media*, 7: 187–199.

Hamdi, A.; Giancola, S.; and Ghanem, B. 2021. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.

Hou, Q.; et al. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3203–3212.

Kanezaki, A.; Matsushita, Y.; and Nishida, Y. 2018. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5010–5019.

Kolesnikov, A.; Dosovitskiy, A.; amd Georg Heigold, D. W.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Lahoud, J.; Cao, J.; Khan, F. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; and Yang, M.-H. 2022. 3D vision with transformers: A survey. *arXiv preprint arXiv:2208.04309*.

Ma, G.; Li, S.; Chen, C.; Hao, A.; and Qin, H. 2021. Rethinking image salient object detection: Object-level semantic saliency reranking first, pixelwise saliency refinement later. *IEEE Transactions on Image Processing*, 30: 4238–4252.

Ma, X.; Qin, C.; You, H.; Ran, H.; and Fu, Y. 2022. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. *arXiv preprint arXiv:2202.07123*.

Mao, J.; Xue, Y.; Niu, M.; et al. 2021. Voxel Transformer for 3D Object Detection. *ICCV*.

Misra, I.; Girdhar, R.; and Joulin, A. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2906–2917.

Peng, H.; Li, B.; Xiong, W.; Hu, W.; and Ji, R. 2014. RGBD Salient Object Detection: A Benchmark and Algorithms. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 92–109.

Qian, G.; Hammoud, H.; Li, G.; Thabet, A.; and Ghanem, B. 2021. ASSANet: An Anisotropical Separable Set Abstraction for Efficient Point Cloud Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; and Ghanem, B. 2022. PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Schütz, M.; Kerbl, B.; and Wimmer, M. 2021. Rendering Point Clouds with Compute Shaders and Vertex Order Optimization. *Computer Graphics Forum*, 40(4): 115–126.

Siris, A.; Jiao, J.; Tam, G. K.; Xie, X.; and Lau, R. W. 2021. Scene context-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4156–4166.

Song, Y.; et al. 2023. Towards End-to-End Unsupervised Saliency Detection with Self-Supervised Top-Down Context. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5532–5541.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 945–953.

Wang, J.; Zhao, Z.; Yang, S.; Chai, X.; Zhang, W.; and Zhang, M. 2022. Global contextual guided residual attention network for salient object detection. *Applied Intelligence*, 1–19.

Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; and Yang, R. 2021. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3239–3259.

Wei, X.; Yu, R.; and Sun, J. 2020. View-GCN: View-based graph convolutional network for 3D shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1850–1859.

Wei, Z.; Chen, B.; Wang, W.; Chen, H.; Wei, M.; and Li, J. 2024. Point Transformer-Based Salient Object Detection Network for 3-D Measurement Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–11.

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

Zhang, Z.; Hua, B.-S.; and Yeung, S.-K. 2019. ShellNet: Efficient Point Cloud Convolutional Neural Networks using Concentric Shells Statistics. In *International Conference on Computer Vision (ICCV)*.

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

Zhou, Z.; Lyu, G.; Huang, Y.; Wang, Z.; Jia, Z.; and Yang, Z. 2024. SDformer: Transformer with Spectral Filter and Dynamic Attention for Multivariate Time Series Long-term Forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 5689–5697.