2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Complex Encoding Transformer for 3D Sonar Target Detection

Tiancheng Cai[1], Dongdong Zhao[1], Peng Chen[1(✉)], Yiran Li[1], Xiang Tian[2], and Ronghua Liang[1]

[1] School of Computer Science and Technology,
Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China,
chenpeng@zjut.edu.cn
[2] College of Biomedical Engineering and Instrument Science,
Zhejiang University, Hangzhou 310009, Zhejiang, China,
xiang.t@163.com

Abstract. With the ongoing advancement of 3D sonar detection technology, research on underwater 3D target detection has gained increasing significance. Currently, there is substantial research on optical point clouds, while research on 3D sonar point clouds remains limited. Underwater 3D sonar recognition differs from optical recognition, facing challenges such as high sparsity, strong noise intensity, and inter-object coupling. However, traditional optical-based methods struggle with recognizing coupled targets like frogmen and bubbles. This paper proposed a detection method based on a dynamic complex encoding transformer. By combining the principles of sparse array 3D sonar imaging and complex decoupling based on prior knowledge, noise and sidelobe interference are effectively reduced. Addressing the challenges of detecting concealed targets, this paper proposed a novel 3D backbone based on complex-encoding, which effectively enhances additional information around targets, achieving efficient recognition of 3D sonar targets. Finally, our model achieved satisfactory performance through both qualitative and quantitative experiments.

Keywords: Underwater detection, 3D sonar, acoustic pointcloud, transformer, complex encoding.

1 Introduction

Intelligent underwater detection has gained increasing significance due to the escalating demand for ocean exploration resources. Among various techniques, underwater detection based on 3D sonar[1] has emerged as one of the most compelling technologies, exhibiting remarkable potential in environmental research [2], underwater robot vision systems, and marine engineering [3]–[5]. Underwater target detection based on 3D sonar point clouds [6] presents a critical challenge that has garnered growing attention. However, research on 3D sonar point clouds is relatively scarce, especially in the area of 3D sonar detection. Compared to optical point clouds, 3D sonar point clouds pose challenges such as high noise levels[7], coupling of foreground and background points, and severe loss of target information, as shown in **Fig. 1** To address these challenges, we propose a dynamic complex encoding-based 3D sonar target detection network that can effectively detect high-noise, heavily occluded, and arbitrarily shaped 3D sonar point cloud detection tasks.

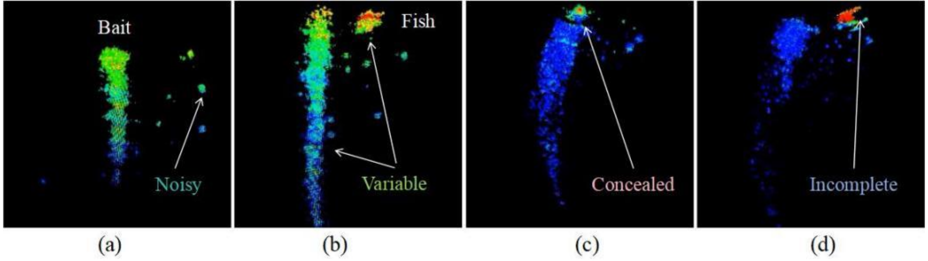

Fig. 1. Imaging of Baiting spot based on 3D sonar, where fish is attracted after the bait disperses. (a) shows a amount of noise in the scene.(b) shows fishes with variable bubbles and bait. (c) displays fish hidden among bubbles and bait. (d) highlights the issue of target incompleteness in the imaging.

In recent years, with the advent of PointNet [8] and the decreasing difficulty of acquiring 3D data, 3D object detection[9] research has emerged rapidly. For objects on the ground, 3D imaging is commonly performed using LiDAR, while underwater [1], [10], [11] objects are usually scanned and imaged using 3D sonar. There are three fundamental differences between underwater and road targets: the first is their location, with road targets typically appearing close to the ground (in a nearly 2.5D scene), while underwater targets can appear at any location. The second difference lies in the propagation medium. Due to factors such as different equipment, refractive index, reflectivity, wavelength, and attenuation, there are significant differences in the point cloud properties of underwater and road targets. The third difference is image quality, with 3D sonar images [12] having higher noise levels and lower resolution than 3D optical images, making 3D sonar image target detection more challenging.

Specifically, sonar targets are often surrounded by various interferences, such as side-lobe interference near the main lobe, noise, and other background objects, which causes deterioration of image quality. This paper is inspired by human observers who typically use factors[13] such as bubbles, wake flows, and side-lobe interference around three-dimensional sonar objects to assist in target identification. For example, frogmen exhale upward bubbles, while AUVs generate wake flows. Despite the high sparsity of 3D sonar data, it is also necessary to utilize surrounding information to assist in recognition. Traditional methods, such as [14][15], directly encode point clouds, making it difficult to separate target information and leading to model overfitting. Other methods, such as [16][17], that simply separate foreground and background points struggle with the challenge of insufficient information after separation. Therefore, this paper proposes a complex mixed encoding method for point cloud encoding. The real and imaginary parts of each point are defined to decouple the target from its surroundings, and features are encoded in parallel, effectively extracting information from both the target and surrounding interferences without altering the original point cloud properties.

3D sonar images often require surrounding information for additional assistance in judgement. For example, human observers usually judge frogmen[18] by the bubbles around them and distinguish them from fish through the wake of underwater submarines. Due to the effective application of the swin-transformer[19] in computer vision and the introduction of a 3D backbone[20] based on the transformer, global 3D information can be effectively extracted. However, it cannot effectively balance the target and its surrounding information, resulting in a high error rate in high noise conditions. To further process the target and surrounding information after complex decoupling, this paper proposes a parallel 3D backbone based on the parallel transformer. A pillar-based sliding window design is used to effectively extract global information, which can better connect the features between different targets and enhance the high correlation of features at different levels and locations. The main contributions of this research are summarized as follows:

(1) Inspired by human observation of 3D sonar images, this work proposes an underwater target detection method based on 3D sonar pointcloud, effectively addressing the challenges of high noise and hidden targets in 3D sonar images.

(2) We propose a complex mixed encoding method, achieving local information enhancement and effectively resolving the issue of traditional algorithms being unable to disentangle target coupling in small areas.

(3) We propose a group mapping 3D backbone based on transformer, utilizing complex sparse window attention to effectively leverage global information, reduce information loss during the transformation to BEV(bird's-eye view), and enhance the recognition of incomplete targets.

2 Relate Work

With the groundbreaking work of PointNet[8] on addressing the issues of point cloud disorder and permutation invariance, the task of 3D object detection has seen a surge in development. One of the first methods was VoxelNet[21], which used 3D convolutions to extract features, but its performance was limited. Subsequently, Yan et al.[22] proposed SECOND, pioneering the use of sparse convolutions to reduce complexity. Lang et al. introduced PointPillar[23], which further reduced complexity and improved performance based on the pillar method, but these methods faced challenges when dealing with objects of varying sizes. Yin et al.[24] proposed CenterPoint, utilizing Gaussian heat maps to eliminate the dependence on anchor sizes, while Wang et al.[20] introduced DSVT, using a dynamic sparse transformer to enhance the performance of the 3D backbone. However, the aforementioned methods still face challenges when dealing with high-noise and target coupling situations. Currently, there is no public 3d sonar dataset available, and 3D sonar faces challenges such as high noise, low resolution, and hidden targets, which have led to few existing target detection methods based on 3D point clouds. He et al. [25] proposed a two-branch point cloud detection network that utilizes graph attention and 3D sparse convolutions to extract detailed features. However, it faces challenges in extracting information from side lobes and noise in sonar images. Lee et al.[26] proposed a multi-view 3D detection method based on AlexNet, but since it does not directly use point clouds for processing, it can cause significant information loss. Kim et al. [27] proposed a PointNet-based method for sonar point cloud classification, but it is not applicable to 3D sonar target detection tasks. Hoang et al.[28] proposed a network inspired by resonant scattering that effectively classifies 3D sonar images, but it is mainly a classification task rather than a detection task. Henley et al.[29] proposed a method for 3D voxel recognition using 3D-Unet and 3D-Vet for 3D forward-looking sonar, but using direct 3D convolution results in weak performance.

3 The Proposed Method

Our network structure is shown in **Fig.2**. Considering the significant noise often present in 3D sonar point clouds, we utilize a simple preliminary network to decompose the original point cloud into real and imaginary parts. To ensure consistency, we assume that the square root of the intensities of the real and imaginary part point clouds equals the intensity of the original point cloud. Although DSVT[20] has effectively validated the usefulness of a transformer-based[19] 3D backbone for point cloud tasks, it faces the challenge of targets in 3D sonar being often coupled with noise. To better adapt to the characteristics of 3D sonar point clouds, we designed a dual-branch decoupled model. Finally, the final proposals are obtained through an anchor-free detection head.

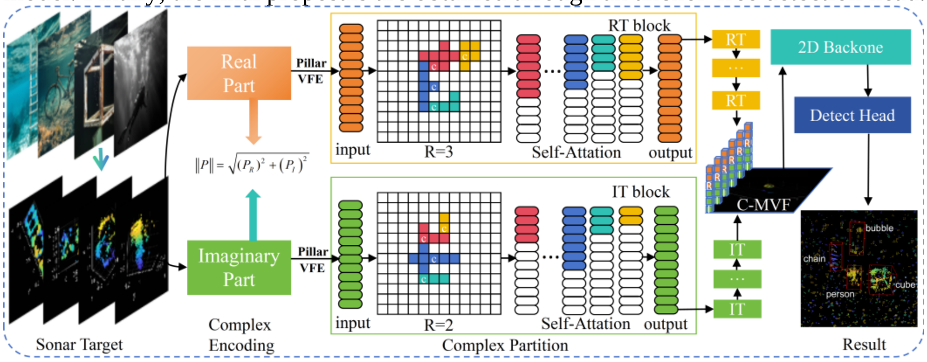

Fig.2. Overview of proposed framework. Here, ***R*** is the grouping radius. If the effective distance from the center point ***C*** is less than ***R***, the points are grouped together. If the number of groups exceeds the maximum allowed, a new set is created directly. Tokens of different colors represent the pillars at corresponding positions, while the same color indicates the same set. The groups shown in the figure only illustrate the grouping situation within a single window.

3.1 Complex mix encoding

Due to the imaging principles of the sparse array, the sidelobes away from the main lobe are elevated across all sections. These sidelobes often contain rich target features, rather than being simply regarded as noise to be removed. The far-field BP of the index (m,n) can be represented as:

$$\left|B(\boldsymbol{W},u_x,u_y)\right| = \left|\sum_{m=1}^{N}\sum_{n=1}^{N}\left[\omega_R\left(m,n\right)+j*\omega_I\left(m,n\right)\right]*\exp(-j\frac{2\pi f_0}{c}(u_x x_m+u_y y_m))\right| \tag{1}$$

where $\omega(m,n)$ indicate the weight coefficients of indices(m,n); $\boldsymbol{c}$ is the acoustical wave speed in water; $\boldsymbol{u}$ is the unit vector. The expression within the exponential function can be represented as:

$$-j\frac{2\pi f_0}{c}(u_x x_m+u_y y_m) = -j(\varphi_x(\alpha,m)+\varphi_y(\beta,n)) \tag{2}$$

and $\varphi_x(\alpha,m)$ and $\varphi_y(\beta,n)$ indicate the phase shift parameters in the horizontal (x-axis) and vertical (y-axis) directions, which can be represented as follows:

$$\varphi_x(\alpha,m) = \frac{2\pi dm\sin\alpha}{\lambda} \tag{3}$$

$$\varphi_y(\beta,n) = \frac{2\pi dn\sin\beta}{\lambda} \tag{4}$$

Considering the specificity of 3D sonar point clouds, a simple network Γ is used to perform complex decomposition on the original point cloud P based on prior knowledge. To ensure consistency, let$\|P\| = \sqrt{(P_R)^2+(P_I)^2}$. The stacked voxel feature encoding (VFE)[21] is used, where feature encoding is performed separately on the imaginary and real point clouds.

3.2 Group Mapping 3D transformer

Due to the sparsity of the sonar point cloud, most pillars are empty, which can lead to significant overhead when directly performing feature extraction. Additionally, the number of non-empty pillars within each sliding window of the same size varies, making it challenging to use traditional transformer architectures directly. Unlike traditional methods, when considering the sidelobes and noise in the 3D sonar point cloud, traditional methods cannot effectively extract sidelobe features. This paper proposes a center-based expansion token division method that efficiently processes the features after complex decoupling in the previous stage, achieving dynamic parallel 3D sparse backbones.

Specifically, after converting the original point cloud into pillars, we further divide it into multiple non-overlapping sliding windows of length $\boldsymbol{W}$ and width $\boldsymbol{L}$, and there are $\boldsymbol{K}$ non-empty pillars:

$$PL = \left\{p_i \,\|\, p_i = [(x_i,y_i);R_i;I_i;d_i]\right\}_{i=1}^{K} \tag{5}$$

where (x,y) are the coordinates of the sparse pillars, and $\boldsymbol{R}$ and $\boldsymbol{I}$ are the real and imaginary features of the complex encoding in the sparse pillars, both of dimension $\boldsymbol{s}$. Then, to ensure that the non-overlapping subsets are of the same size, the number of sets in the window is:

$$S = \begin{cases} \left\lceil \dfrac{K}{\delta} \right\rceil & if\, K\%\delta > 0 \\ \left\lceil \dfrac{K}{\delta} \right\rceil - 1 & otherwise \end{cases} \tag{6}$$

where $\lceil \cdot \rceil$ is ceiling function and $\boldsymbol{\delta}$ is a hyper-parameter that allocates the maximum number of non-empty pillars. Due to the sparsity of sonar point clouds, we adopt a sparse partitioning approach. For a given set of non-empty pillars $PL_i=\{p_1,p_2,...,p_n\}$ in the sliding window $\boldsymbol{W_i}$, the non-empty centroid $\boldsymbol{C_0}$ is taken as the first center of set:

$$set(i)\{p_j\}, ds(c_i, p_j) < R \tag{7}$$

where $ds(c,k)$ is the connected distance from the centroid $\boldsymbol{c}$ to $\boldsymbol{p}$, meaning that empty pillars cannot be traversed when calculating the distance. $\boldsymbol{R}$ is the expansion radius, those below $\boldsymbol{\delta}$ are completed using mask token. For the grouped real and imaginary parts of the set, they are composed of RT block and IT block through multi-head attention layers, including Position-wise Feed-Forward Networks(FFN)[20] and Layer Normalization.

After the 3D backbone, we concatenate the features of the real and imaginary parts, map them to the BEV, and perform detection using anchor-free[24] 2D backbone to obtain the final results.

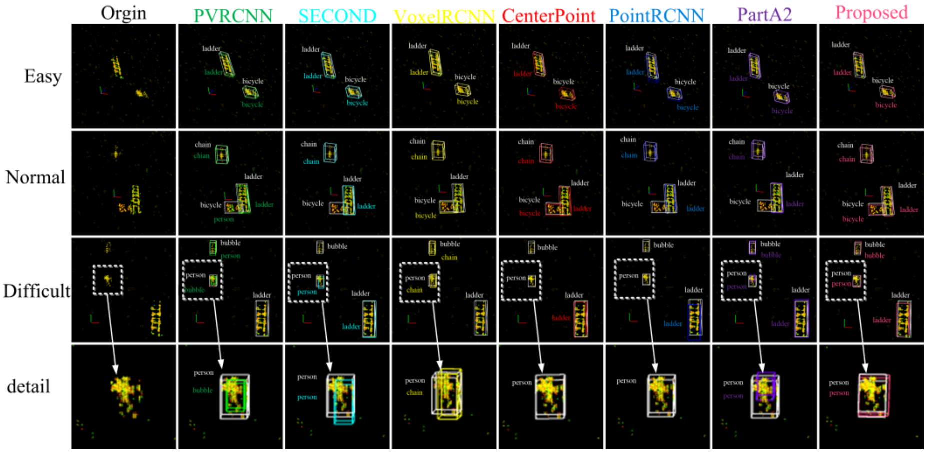

Fig. 3. Different methods for displaying 3D sonar point cloud images. Each row represents the comparison of different methods on the same point cloud, while each column shows the comparison of the same method on different point clouds. The boxes in different colors indicate different prediction results: white represents person, green represents bubble, blue represents chain, yellow represents ladder, and red represents bicycle.

4 Experiments

In this section, we provide quantitative and qualitative experimental results and analysis. Due to the current lack of publicly available 3D sonar point cloud datasets, we used our self-collected dataset as the training set and compared it with mainstream methods. To enhance the model's generalization ability, our model was pre-trained using the KITTI dataset. We divided the 7481 training samples into a training set, with 3712 samples in the train set and 3769 samples in the validation set. The test set contains 7518 samples. The point cloud is clipped within the range of (-8,8)m along the X-axis, (-8,8)m along the Y-axis, and (0,15)m along the Z-axis. For our model, the pillar size is (0.16,0.16,15)m.

In the process of group mapping within the model, we optimized the parameter details. By dividing the point cloud into imaginary and real parts, we observed that noise in the imaginary part is more dispersed, while targets in the real part are more concentrated. Therefore, we set the group radius for the imaginary part slightly smaller than that for the real part, with R_i=2, R_r=3.

Table 1 Performance comparison on 3D sonar dataset test-set with mean average precision(mAP)

Method	Person	Bubble	Chain	Ladder	Bicycle	mAP↑
VoxelRCNN[30]	18.6	60.8	80.4	50.4	88.9	59.8
SECOND[22]	5.6	1.8	63.7	36.6	57.6	33.0
PointRCNN[16]	26.3	47.3	69.0	5.5	26.0	34.8
PartA2[17]	46.0	59.3	48.2	48.8	87.5	58.0
DSVT[20]	57.7	73.2	**90.6**	**90.6**	90.8	80.6
Proposed	**62.5**	**76.7**	87.6	88.6	**94.8**	**82.3**

Table 1. shows the comparison of different mainstream methods. The current methods are primarily optimized for autonomous driving datasets, where the sizes of vehicles and pedestrians are relatively stable. However, in 3D sonar dataset, issues such as occlusion make it necessary to preset the sizes of anchors, which results in less effective outcomes. Consequently, the recognition rates for targets like persons and bubbles are relatively low.

Table 3 shows the 3d backbone with other methods and **Table 4** shows the 2d backbone with other methods. **Fig. 3** displays the visualization results of different methods. **Fig. 4** shows the detection performance vs speed of different methods, the size of the sphere is parameters of each methods.

Fig. 5 illustrates the different shapes of various targets in the 3D sonar dataset collected from Qiandao Lake. It is evident that bubbles and chains are quite similar in appearance, with chains exhibiting different forms in the water. Objects like bicycles, ladders, and squares maintain relatively stable shapes underwater, whereas divers are more challenging to identify due to various interferences and their smaller size.

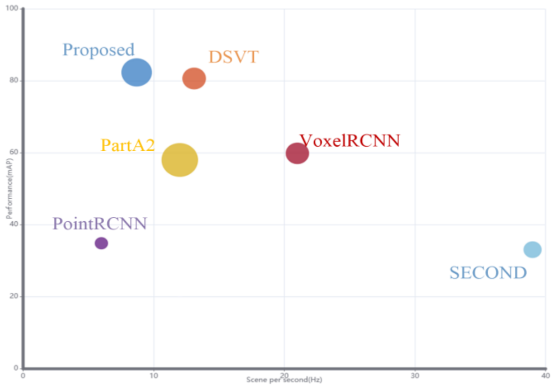

Fig. 4. Detection performance(mAP) vs speed(Hz) of different methods on 3d sonar datasets, the size of the sphere represents parameters of each models. All speeds are evaluated on an NVIDIA 2080ti GPU.

Table 2. Ablation Studies on Test-set

Method	Complex Decoupling	3D parallel Transformer	Result		
			Easy	Hard	mAP
(a)	✗	✗	77.7	54.6	68.5
(b)	✓	✗	84.2	64.8	74.5
(c)	✗	✓	86.3	65.2	75.8
(d)	✓	✓	90.3	69.6	82.3

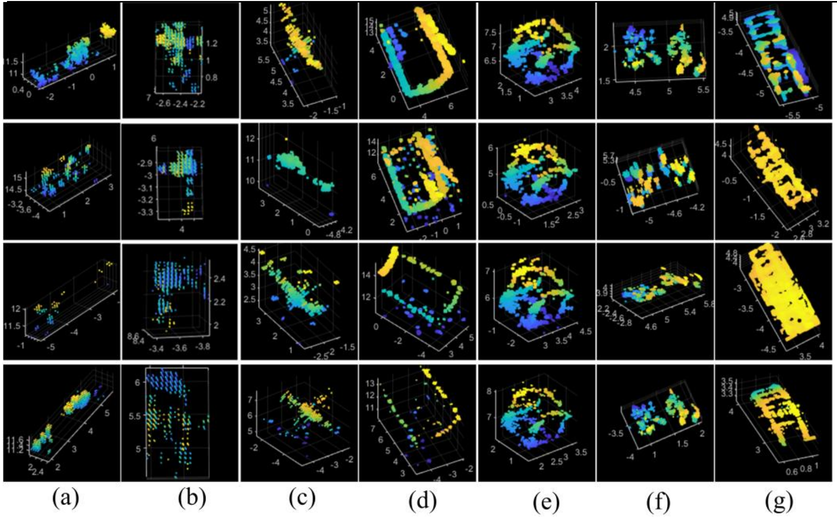

Fig. 5. Different categories and different shapes of targets within the annotation box after manual annotation, (a)Bubble, (b)Frogman, (c)Chain, (d)U-chain, (e)Cube, (f)Bicycle, (g)Ladder.

Table 3 Comparison with other methods. Only switch the 3D backbone while other components remain unchanged.

3D Backbone	Parameters	Frame rate	mAP↑
ResBackbone1x	95M	12.1	79.4
DSVT	92M	13.1	80.4
Proposed-single	69M	11.6	80.5
Propose-dual	123M	8.7	82.3

Table 4 Comparison with other methods. Only switch the 2D backbone while other components remain unchanged.

2D Backbone	Frame rate	Easy	Hard	mAP↑
CenterPoint-Pillar	26.2	71.3	63.2	68.4
CenterPoint-Voxel	23.1	76.6	65.8	71.2
Proposed-Pillar	10.2	88.5	69.9	79.2
Propose-Voxel	8.7	91.7	75.6	82.3

5 Conclusion

In this paper, we very first propose a 3D sonar object detection method based on dynamic complex encoding. Our design mainly includes complex decoupling encoding and a group mapping 3D transformer backbone. Through qualitative and quantitative experiments, the effectiveness of the method is demonstrated. To enhance the generalization ability, pre-training was conducted based on datasets such as KITTI. This work effectively solve challenge such as object occlusion and inter-object coupling in 3D sonar detection. In our practical experience, we have observed significant differences in results based on different objectives. Detection rates are higher for rigid objects such as wooden frames and bicycles, while the performance is poorer for objects that easily change shape, such as divers and fish. In future work, we will focus on enhancing generalization capability.

6 ACKNOWLEDGEMENT

The authors would like to thank the reviewers for the constructive comments which help to improve the quality and understandability of this paper. This work is supported by the National Science Foundation of China under Grant 62371421, by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-B2024003, by the “Vanguard Leading Goose + X” Science and Technology Programof Zhejiang Province under Grant 2025C01207, by the Leading Innovation Team of the Zhejiang Province under Grant 2021R01002.the Leading Innovation Team of the

Zhejiang Province under Grant 2021R01002, by the the Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020005.

References

1. Fl´avio P. Ribeiro and V´ıtor H. Nascimentoz: Fast transforms for acoustic imaging— part i: Theory. IEEE Transactions on Image Processing 20(8), 2229–2240 (2011)
2. Yu Wang, Chong Tang, Mingxue Cai, Jiye Yin, Shuo Wang, LongCheng, Rui Wang, and Min Tan.: Real-time underwater onboardvision sensing system for robotic gripping. IEEE Transactions onInstrumentation and Measurement 20, 1–11 (2021)
3. Avi Abu and Roee Diamant.: Enhanced fuzzy-based local information algorithm for sonar image segmentation. IEEE Transactions on Image Processing 29, 445–460 (2020)
4. Rizwan Khan, Atif Mehmood, Saeed Akbar, and Zhonglong Zheng.: Underwater image enhancement with an adaptive self supervised network. In 2023 IEEE International Conference on Multimedia and Expo, pp. 1355–1360, Brisbane(2023)
5. Biao Wu, Xinchen Ye, Fei Xue, and Rui Xu.: Learning data hallucination and reciprocal guidance for underwater depth estimation and color correction. In 2022 IEEE International Conference on Multimedia and Expo, pp. 1–6, Taipei(2022)
6. Chengcheng Ma, Weiliang Meng, Baoyuan Wu, Shibiao Xu, and Xiaopeng Zhang.: Towards effective adversarial attack against 3d point cloud classification. In 2021 IEEE International Conference on Multimedia and Expo, pp. 1–6, Shenzhen(2021)
7. Yi-Peng Liu, Qi Zhong, Ronghua Liang, Zhanqing Li, Haixia Wang, and Peng Chen.: Layer segmentation of oct fingerprints with an adaptive gaussian prior guided transformer. IEEE Transactions on Instrumentation and Measurement 71, 1–15 (2022)
8. R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 77–85, Honolulu(2017)
9. Keli Wen, Nan Zhang, Ge Li, and Wei Gao.: Mpvnn: Multi-resolution point-voxel non-parametric network for 3d point cloud processing. In 2024 IEEE International Conference on Multimedia and Expo, pp. 1–6, NiagraFalls(2024)
10. Pan Mu, Jing Fang, Haotian Qian, and Cong Bai.: Transmission and color-guided network for underwater image enhancement. In 2023 IEEE International Conference on Multimedia and Expo, pp.1337–1342, Brisbane (2023)
11. Muwei Jian, Qiang Qi, Junyu Dong, Yinlong Yin, Wenyin Zhang, and Kin-Man Lam.: The ouc-vision large-scale underwater image database. In 2017 IEEE International Conference on Multimedia and Expo, pp. 1297–1302, HongKong(2017)
12. Jing Hu, Xincheng Wang, Ziheng Liao, and Tingsong Xiao.: M-gcn:Multi-scale graph convolutional network for 3d point cloud classification. In 2023 IEEE International Conference on Multimedia and Expo, pp. 924–929, Brisbane(2023)
13. Zhenqiang Zhang, Chuantao Li, Jian Song, Jialiang Lv, Chunxiao Wang, Zhigang Zhao, and Jidong Huo.: Stui-net: Semi-supervised transformer for underwater information enhancement. In 2024 IEEE International Conference on Multimedia and Expo, pp. 1–6, NiagraFalls(2024)
14. Tianwei Yin, Xingyi Zhou, and Philipp Kr¨ahenb¨uhl.: Center-based 3d object detection and tracking. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11779–11788, Seattle(2024)
15. Qi Cai, Yingwei Pan, Ting Yao, and Tao Mei.: 3d cascade rcnn: High quality object detection in point clouds. IEEE Transactions on Image Processing 31. 5706–5719 (2022)

16. Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li.: Pointrcnn: 3d object proposal generation and detection from point cloud. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–779, Long Beach(2019)
17. Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(8), 2647–2664(2021)
18. Tian Zhou, Jikun Si, Luyao Wang, Chao Xu, and Xiaoyang Yu.: Automatic detection of underwater small targets using forward-looking sonar images. IEEE Transactions on Geoscience and Remote Sensing 60. 1–12 (2022)
19. Haram Choi, Jeongmin Lee, and Jihoon Yang.: N-gram in swin transformers for efficient lightweight image super-resolution. In 2023IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2071–2081, Vancouver(2023)
20. Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang.: Dsvt: Dynamic sparse voxel transformer with rotated sets. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13520–13529, Vancouver(2023)
21. Yin Zhou and Oncel Tuzel.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4490–4499, Salt Lake City(2018)
22. Yan Yan, Yuxing Mao, and Bo Li.: Second: Sparsely embedded convolutional detection. Sensors 18(10), (2018)
23. Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom.: Pointpillars: Fast encoders for object detection from point clouds. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12689–12697, Long Beach(2019)
24. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He.: centerpoint. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803, Salt Lake City(2018)
25. Yunqian He, Guihua Xia, Yongkang Luo, Li Su, Zhi Zhang, Wanyi Li, and Peng Wang.: Dvfenet: Dual-branch voxel feature extraction network for 3d object detection. Neurocomputing 459, 201–211 (2021)
26. Kiwoo Shin, Youngwook Paul Kwon, and Masayoshi Tomizuka.: Roar-net: A robust 3d object detection based on region approximation refinement. In 2019 IEEE Intelligent Vehicles Symposium, pp. 2510–2515, Paris(2019)
27. Minsung Sung, Jason Kim, Hyeonwoo Cho, Meungsuk Lee, and SonCheol Yu.: Underwater-sonar-image-based 3d point cloud reconstruction for high data utilization and object classification using a neural network. Electronics 9(11). (2020)
28. Trung Hoang, Kyle S. Dalton, Isaac D. Gerg, Thomas E. Blanford, Daniel C. Brown, and Vishal Monga.: Resonant scattering-inspired deep networks for munition detection in 3d sonar imagery. IEEE Transactions on Geoscience and Remote Sensing(61). 1–17(2023)
29. Heath Henley, Austin Berard, Evan Lapisky, and Matthew Zimmerman.: Deep learning in shallow water: Cnn-based 3d-fls target recognition. In OCEANS 2018 MTS/IEEE, pp. 1–7, Charleston(2018)
30. Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In 45th National Conference on Artificial Intelligence, Vancouver(2021)
31. Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In 2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10526–10535, Seattle(2020)