15th World Congress on Computational Mechanics (WCCM-XV)
8th Asian Pacific Congress on Computational Mechanics (APCOM-VIII)
Virtual Congress: 31 July – 5 August 2022
S. Koshizuka (Ed.)

OBJECT CLASSIFICATION AND SEGMENTATION BASED ON DEEP LEARNING USING UNDERWATER MAPPING DATA

HIROSHI OKAWA[1], SHIGEYUKI OMOTO[2], SHOTA YAGI[3], TAKASHI MIYAMOTO[4] and KAZUO KASHIYAMA[5]

[1] EJ Innovation Technology Center, Eight-Japan Engineering Consultants Inc.
5-33-11 Honcho, Nakano-ku, Tokyo 164-8601, Japan
ookawa-hi@ej-hds.co.jp

[2]EJ Innovation Technology Center, Eight-Japan Engineering Consultants Inc.
3-1-21 Tsushima-Kyomachi, Kita-ku, Okayama, Okayama 700-8617, Japan
oomoto-shi@ej-hds.co.jp

[3] Disaster Mitigation and Facility Maintenance Department, Eight-Japan Engineering Consultants Inc.
3-1-21 Tsushima-Kyomachi, Kita-ku, Okayama, Okayama 700-8617, Japan
yagi-sho@ej-hds.co.jp

[4] Department of Civil and Environmental Engineering, University of Yamanashi
4-3-11 Takeda, Kofu, Yamanashi 400-8511, JAPAN
tmiyamoto@yamanashi.ac.jp

[5] Department of Civil and Environmental Engineering, Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8521, JAPAN
kaz@civil.chuo-u.ac.jp

Key words: Underwater Mapping Data, AUV, ASV, Deep Learning, CNN, PointNet++

Abstract. *This paper presents a fast and accurate classification method for underwater objects using underwater mapping data obtained by a small Autonomous Underwater Vehicle (AUV) and autonomous surface vehicle (ASV). For the mapping data, in addition to underwater acoustic reflection intensity images, water depth data, point cloud data and backscattering reflection intensity data are employed. We propose the automatic classification and semantic segmentation method on deep learning using a convolutional neural network (CNN) and PointNet++. In order to verify the effectiveness of the present method, we applied it to the measured several underwater mapping data.*

1 INTRODUCTION

With the recent development of surveying and positioning technology and automated control robot technology, i.e., unmanned aerial vehicles (UAVs), land mapping technology has improved dramatically, besides advancements in underwater mapping technology. Recent developments in measuring instruments and underwater robots have made it possible to realize high-quality underwater mapping using high-resolution acoustic exploration equipment. The obtained underwater mapping data is represented by the strength of sound reflection. Therefore,

color information is not included. As a result, when performing object discrimination from the acquired data, this process depends on the interpretation by a professional technician, and it is difficult to make an immediate, automatic, and quick determination. Furthermore, there is a concern that there will be differences in the experience of workers and human error due to routine work.

In this paper, we introduce a method of using an unmanned robot such as an AUV or ASV equipped with various echo sounders in order to acquire underwater mapping data in a simple, safe, and high-definition manner; as well as automatically and quickly discriminate underwater objects via classification and semantic segmentation based on deep learning.

Underwater mapping data includes underwater acoustic reflection intensity images, water depth data, backscatter intensity data, or underwater point group data. For each, automatic classification and semantic segmentation are realized using convolutional neural network (CNN) and PointNet++[1]. In addition, to verify the effectiveness of this method, it was applied to the actually measured underwater mapping data.

2 UNDERWATER MAPPING DATA AND DATA ACQUISITION METHODS

2.1 Equipment used

Mapping data is defined as data to which various information is added based on geographical position coordinates; and underwater mapping data is data acquired underwater, such as images, topography, water quality, and time, and data with position coordinates added. As a method of acquiring underwater mapping data, it is common to equip a ship with an acoustic exploration device such as a side scan sonar or a multi-beam echo sounder for measurement.

In this study, an AUV was used to measure a wide range and a large depth, and an ASV was used to measure a narrow range with a relatively small depth with high accuracy; images were captured with various measuring devices installed. Using the AUV as a platform to mount an arbitrary sensor, necessary data can be acquired. By setting the course and depth, it is possible for the operator to automatically acquire data without waiting nearby. Among the sensors mounted onto the AUV used in this paper (Figure1), an interferometry echo sounder (EdgeTech; 2205AUV) was used, and it is possible to acquire three types of data (acoustic reflection intensity image / water depth data / backscatter intensity) simultaneously. Meanwhile, using an ASV as a platform to mount an arbitrary sensor, necessary data is acquired. By setting the course and depth, it is possible for the operator to automatically acquire data without waiting nearby. Among the sensors mounted on the ASV (Figure 2) used in this study, it is possible to acquire point cloud data using a multi-beam echo sounder (iWBMSh; manufactured by Norbit).

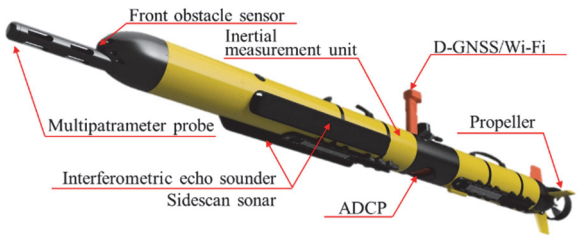

Figure 1: AUV and interferometric echo sounder

Figure 2: ASV and narrow multibeam echo sounder

2.2 Underwater acoustic reflection intensity image/water depth/backscattering intensity

Depth sounding with side-scan sonar is the main method for analyzing the bottom of the water. In this research, the echo sounder mounted onto the AUV emits a frequency of 10–2,000 kHz in a fan shape toward the bottom of the water; and the acoustic reflection intensity image / water depth data / backscatter intensity is acquired by the method of receiving the sound waves scattered and reflected at the bottom of the water. In water, the strength of the reflected sound wave differs depending on the geology of the bottom of the water. Therefore, by displaying the intensity in raster with shades, it is possible to capture the state of the bottom of the water as a surface like a black-and-white image.

The acoustic reflection intensity image is continuous data with a value in each cell, and the water depth data and backscatter intensity indicated by the green points in the Figure are discrete data with features in common plane coordinates. Figure 3 shows an example of acquired data displayed in raster for feature comparison.

2.3 Regarding the underwater point cloud

Multi-beam echosounder is another major method for comprehending the bottom of the water. In this method, a fan-shaped wave beam is oscillated from the sonar in the horizontal direction, and the received wave reflected to the bottom of the water and returned is received by the slit in the vertical direction. Multiple slits are lined up in the receiver, and many points can be measured with a single oscillation. In this study, the ASV was equipped with an echo sounder to acquire point cloud data, as shown in Figure. 4. Because the point cloud data in water does not have color information, only three-dimensional coordinate values are handled. A point cloud is a set of points, and each point can have information such as position information, color, and reflection intensity. Moreover, the situation inside the object is not reflected.

In addition, as a characteristic of point cloud data, invariants such as unordered and unclear adjacency as well as objects comprising point clouds have the characteristic of movement invariance such that the type of the object does not change even when processing such as movement or rotation of the point cloud is performed. Therefore, when using point cloud data as input data in the construction of a deep learning model, it is necessary to pay attention to these characteristics.

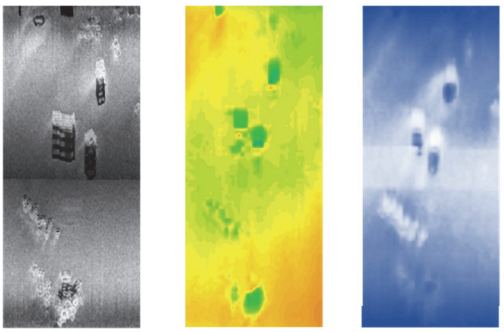

Figure 3: Example of data by interferometric echo sounder (Rasterized for comparison)

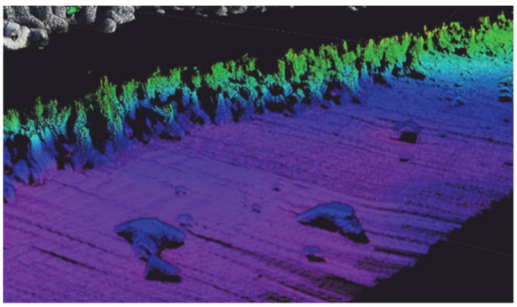

Figure 4: Example of data by narrow multibeam echo sounder (Colored according to depth)

3 OBJECT CLASSIFICATION VIA MULTIMODAL DEEP LEARNING USING UNDERWATER ACOUSTIC REFLECTION INTENSITY IMAGES

3.1 Object classification through multimodal deep learning

Using deep learning, features can be automatically extracted by letting a computer learn a feature extraction task based on human knowledge[3]. In addition, multimodal deep learning enables more accurate feature extraction by processing information from multiple modalities in an integrated manner[4].

In this study, after concatenating tensors in the input layer, a CNN was used as a feature extractor. Therefore, we decided to handle multiple modalities (Figure 5). When inputting data, after developing a grid of the discrete data by the nearest neighbor method based on the grid size of continuous data, the three data of acoustic reflection intensity image, water depth, and backscattering intensity were stacked in the channel direction to combine them as a tensor of height x width x 3ch. By performing these processes, it is possible to express the characteristics of the data based on the position coordinates, even if the data differs, and it is possible to obtain correlation between the data. More specifically, when dealing with data that has common geospatial coordinates, such as underwater mapping data, CNN processing can be performed without data position adjustment processing.

In this research, we aimed to improve the accuracy of image classification; and by performing multimodal deep learning that combines multiple types of underwater mapping data and a CNN, the data of arbitrary pixel size units were divided into 3 classes (artificial structures / rocks / others (gravel, etc.)). Figure 4 presents the model conFigureuration and structure proposed in this paper.

This model roughly constitutes an input layer, an intermediate layer, and an output layer. As a feature of this model, the acoustic reflection intensity image and the water depth data / backscatter intensity tensors are fused to the input data of each modality in the input layer. Next, after extracting the features of the data fused in the intermediate layer, the output layer is finally classified based on the features.

In addition, hyperparameters, which are values for suppressing the behavior of the algorithm, were set for each region. Please refer to the next chapter for the set values. By adjusting the value, improvement in model performance and suppression of overfitting can be expected.

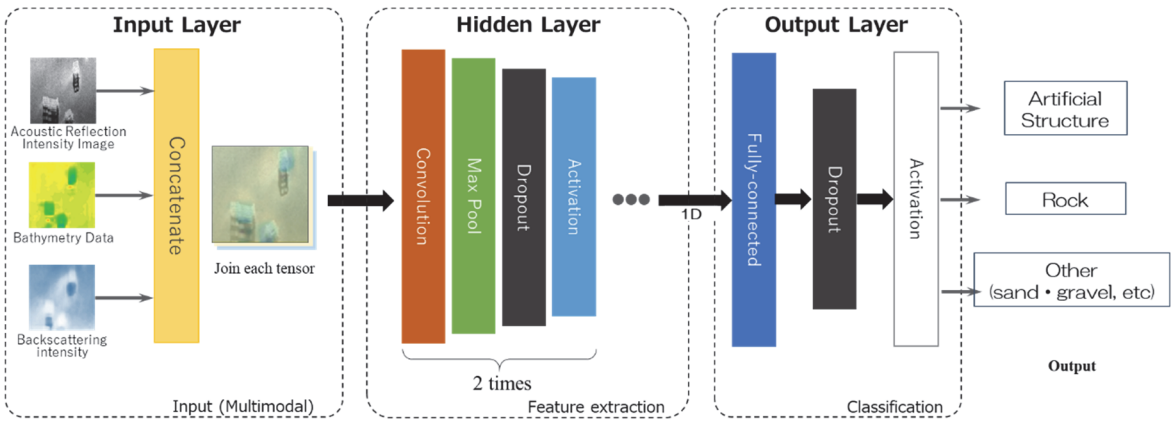

Figure 5: Convolutional Neural Network architecture

3.2 Application example

As the subject of this study, we focused on an area of sea with a depth of 20 m or less as a water area where other objects such as artificial structures, rocks, and gravel coexist. To obtain the underwater acoustic reflection intensity image, dual frequency simultaneous oscillation was set to 540 / 1,600 kHz, and the water depth data and backscatter intensity were set to the frequency of 500 kHz, and each data was acquired. The survey area was approximately 800 m × 600 m, and underwater mapping data was acquired by reciprocating over the sea area while maintaining a certain depth. The image segmentation size was 20 px × 20 px, approximately 10 cm per pixel. Annotation software was used to create the training dataset, as shown in Figure 6, and underwater acoustic reflection intensity was arbitrarily specified by a professional engineer according to the shape of the object in the image. The training data were also labeled in the water depth data and backscatter intensity data based on the position coordinates, and the images were divided. The number of divided images was adopted after considering and adjusting the effect on learning, as follows: 536 artificial structures, 712 rocks, 1,000 gravel, respectively. Half of these were used as training data (30% were validation data), and the other half (268 artificial structures / 356 rocks / 500 gravel) were used as test data. For comparison and verification, we will handle water depth data and underwater backscatter intensity data in addition to image classification using acoustic reflection intensity image data, and each hyperparameter is shown in Table-1. Table 2 details the cases and combinations, and Figure 7 shows the correct answer rate and the reproducibility rate for each class. Case I is synonymous with the so-called single-mode deep learning model. Cases II–IV are classified by a multimode deep learning model, and the correct answer rate, macro F-measure, and class reproducibility are used as indicators of correct answer accuracy. The result of Case IV can be confirmed as the best.

Table 1: CNN hyperparameter

Type	Convolution layer			Optimization Algorithm	Learning Rate	Number of Units in Fully Connected layer	Activation Function
	Number of Layers	Filter Size	Number of Filters				
Value	2	5	8	Adam	1e-5	20	Relu

Table 2: Evaluation Indicators by Data Combination

Number of Concatenation	Data Combination (Case Number)	Accuracy	F-measure
1	I.Acoustic reflection intensity images only	0.82	0.78
2	II.Acoustic reflection intensity images + Bathymetry data	0.92	0.91
	III.Acoustic reflection intensity images + Backscattering intensity data	0.84	0.82
3	IV. Acoustic reflection intensity images + Bathymetry data + Backscattering intensity data	0.94	0.93

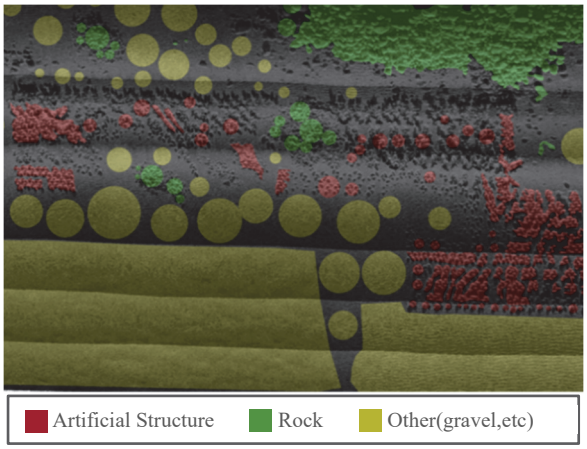

Figure 6: Example of acoustic reflection intensity image and training data

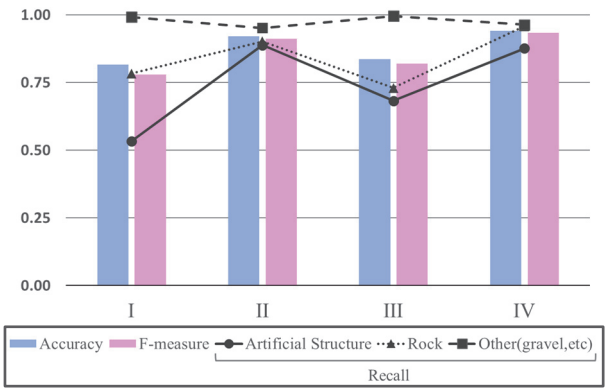

Figure 7: Accuracy/F-measure/Recall

4 OBJECT RECOGNITION BY DEEP LEARNING OF 3D POINT CLOUD USING UNDERWATER POINT CLOUD DATA

4.1 3D point cloud object recognition via deep learning

The 3D point cloud forms a 3D model with a large number of points containing 3D coordinate information. However, it is just a set of points, and these points have no order. In addition, the points are not always distributed at regular intervals, and the density of the points is not constant, so the adjacency between the points is unclear. A single point has no particular meaning, and a three-dimensional shape is expressed by multiple points in close proximity. Therefore, the positional relationship between adjacent points is very important. To perform deep learning in consideration of such properties, a 3D point cloud-based method for learning by directly inputting a 3D point cloud[5] is becoming mainstream instead of the above-mentioned CNN[6]. PointNet++ is adopted in this research as a typical method, and its structure is depicted in Figure 8. The part that extracts local features comprises three types of layers: sampling layer, grouping layer, and PointNet layer. In the sampling layer, the input 3D point cloud is sampled, in the grouping layer, the neighborhood point cloud centered on the sampled point is extracted, and in the PointNet layer, the feature quantity is calculated from the extracted neighborhood point cloud. Through these processes, it is possible to learn the local features of the input three-dimensional point cloud.

PointNet++ builds a network containing two abstract layers that obtain multiscale information according to the point density. The first is the sampling layer, which performs sampling in a metric space. Farthest point sampling (FPS) is used as the sampling method, and the representative points are selected from the input point cloud in a sense of equality. Next, in the second grouping layer, the point cloud not selected by FPS is assigned to the group of the closest representative point cloud among the representative point clouds selected by FPS. PointNet++ constitutes single scale grouping (SSG) that groups at one distance and multiscale grouping that considers multiple distances as grouping methods. However, in this study, we adopted SSG, which has a low calculation cost. By recursively applying these two layers and

PointNet[7], we could construct a deep learning framework that aggregates multiscale information and considers the local structure of the point cloud.

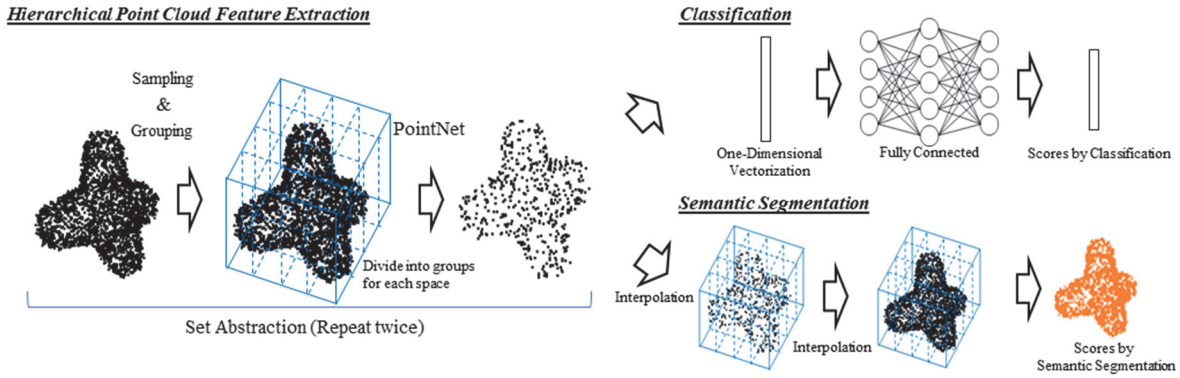

Figure 8: PointNet++ architecture and an example of its application to set classification.

4.2 Application example

As the subject of this study, we considered an area of sea area with a depth of 10 m or less as a water area dotted with wave-dissipating blocks. In addition, two types of wave-dissipating blocks were employed as detection and coloring targets. After acquiring the underwater point cloud data, the point cloud data of each block included in each fixed area was extracted using the point cloud processing software. As shown in Figure. 9, the extracted blocks are block A and block B, and the detection model and the semantic segmentation model were evaluated. The combination of the model used and the training /test data is shown in Table 3.

For the test data in the detection model, individual point cloud model data obtained from the data acquisition area was used. Whereas, the acquired area data was directly used for the test data in the semantic segmentation model.

In this study, the wave-dissipating block to be detected was an off-the-shelf product. Therefore, training data was created by realizing a CAD model based on the design drawing and converting it into a point cloud model. An example of the training data of each created block is provided on the left in Figure. 9. The total number of points that make up each model is approximately 3,000.

In addition to using the created individual point cloud model shown on the left in Figure. 9 for the training data in the detection model, we considered block data in which shadows and blind spots are missing from the water surface and devised a way to improve the detection accuracy by partially cutting off the original training data. The created point cloud data is shown on the right in Figure. 9. Using the same number of models in the processed point cloud model, the accuracy was compared with the unprocessed data. Meanwhile, for the training data in the semantic segmentation model, the area data was created using the individual block data with high accuracy in the detection model. Figure 10 presents an example of the created area data, and Table 4 shows the evaluation of the detection model. From the results of the detection model, it can be confirmed that the accuracy rate and classification accuracy were almost doubled using the machining data. In addition, the Mean-IoU value is as high as 0.996 in the results of semantic segmentation, which is at a level of practical application.

Table 3: Dataset for 3D point cloud deep learning

	Training data	Test data
Detection model	(1) 50 types of raw data (2) 50 types of processed data	Individual data from measurement area Block A: 115 pieces Block B: 54 pieces
Semantic segmentation model	20 types of area data with 24 pieces of processed block data on the same space	Region data from the measurement area Block A: 3 regions Block B: 1 region

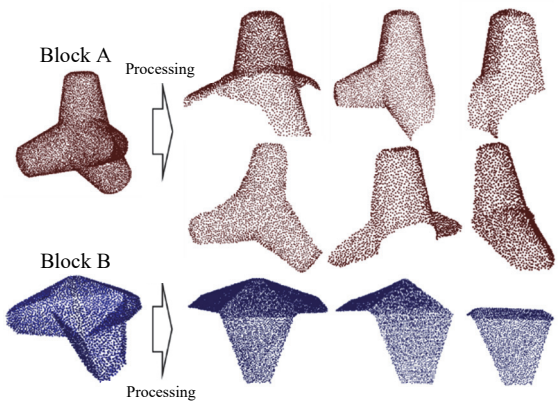

Figure 9: Point cloud model created by 3D CAD model

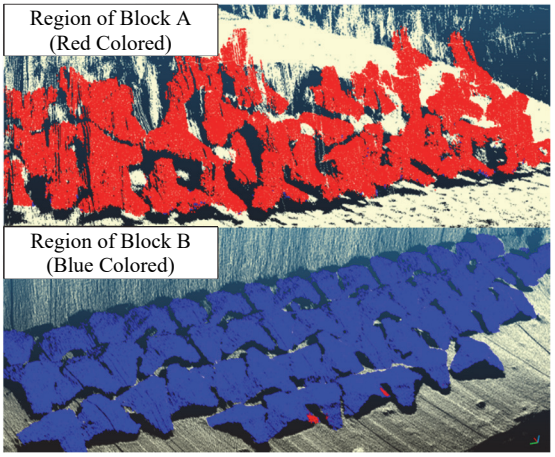

Figure 10: Object class segmentation of class "Block"

Table 4: Evaluation of classification models by confusion matrix

Raw data

		Actual		
		Block A	Block B	Recall
Predicted	Block A	15/115	100/115	0.13
	Block B	10/54	44/54	0.81
	Precision	0.60	0.31	F-measure 0.46

Processed data

		Actual		
		Block A	Block B	Recall
Predicted	Block A	93/115	22/115	0.81
	Block B	4/54	50/54	0.93
	Precision	0.96	0.69	F-measure 0.85

5 CONCLUSIONS

In this paper, we acquired mapping data using an unmanned robot such as an AUV or ASV, and classified and discriminated objects via deep learning. Automatic classification and semantic segmentation were realized using a CNN and PointNet++ depending on the acquisition method and data. In addition, to verify the effectiveness of this method, we applied it to actually measured underwater mapping data and obtained the following conclusions.

Combining all three types of underwater acoustic reflection intensity image, water depth data, and backscatter intensity data as input data into a three-layer tensor resulted in a dramatic

improvement in the reproducibility rate and overall correct answer rate in image classification, demonstrating the effectiveness of object classification by multimodal deep learning.

Moreover, the classification performance was dramatically improved by applying PointNet++ to the acquired point cloud data and adopting the point cloud data created from CAD data and the cut-processed data as training data.

As future tasks, we plan to increase the number of application cases, subdivide the classification class, and study classification for all underwater areas by transfer learning.

REFERENCES

[1] Qi, C.R., Yi, L., Su, H. and Guibas, L.J. : PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, In Advances in neural information processing systems, pp. 5099-5108, 2017.

[2] Krizhevsky, A., Sutskever, I., and Hinton, G.E..: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, pp.1097-1105, 2012.

[3] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A.Y.: Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689-696, 2011.

[4] Yulan, G., Hanyun, W., Qingyong, Hu., Hao, L., Li, L., and Mohammed B.: Deep learningfor 3d point clouds: A survey,IEEE transactions on pattern analysis and machine intelligence,2020.

[5] Zhou, Y., Tuzei, O.: VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection, Computer Vision and Pattern Recognition, pp. 4490-4499, 2018.

[6] Ji, S., Xu, W., Yang, M., and Yu, K.:3D Convolutional Neural Networks for HumanAction Recognition,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.35, pp.221-231, 2013.

[7] Qi, C.R., Su, H., Mo, K. and Guibas, L.J. : PointNet : Deep Leaning on Point Sets for 3D Classification and Segmentation, Computer Vision and Pattern Recognition, pp. 652-660, 2016.