ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888293

Active Visual Learning for Robots with Dueling Deep Q-Networks and Transformer Encoders

Hongliang Zeng
South China University of Technology
Guangzhou, China
scutzenghongl@gmail.com

Ping Zhang
South China University of Technology
Guangzhou, China
pzhang@scut.edu.cn

Fang Li
South China University of Technology
Guangzhou, China
cslifang@scut.edu.cn

Qinpeng Yi
South China University of Technology
Guangzhou, China
csqpyi@mail.scut.edu.cn

Jiahua Wang
South China University of Technology
Guangzhou, China
Jh_Wang_scut@outlook.com

Tingyu Ye
South China University of Technology
Guangzhou, China
cstingyuye@mail.scut.edu.cn

***Abstract*—Active vision learning aims to develop intelligent systems capable of actively exploring and understanding their surroundings to optimize detection performance. Although current research has begun to explore how reinforcement learning can drive robots to actively perceive their environment, it often overlooks the critical role of integrating historical contextual information. To address this, we propose an innovative approach to active vision learning, designed to enhance target detection performance in unknown environments. Specifically, this method combines reinforcement learning with deep neural network techniques and cleverly designs a sliding window mechanism to integrate observations over multiple steps into the state input. We employ PointNet++ for feature extraction and utilize a Transformer encoder module to process spatial contextual information, thereby constructing a comprehensive representation of the robot's environment. Additionally, our carefully designed reward mechanism encourages the robot to prioritize the exploration of diverse object categories. With these reward strategies, we achieve significant improvements in detection accuracy and sampling efficiency. We validated our approach on the real-world 3D dataset R3ED, and the results demonstrate that our method outperforms other baseline methods in terms of performance.**

***Index Terms*—Active vision learning, Robotic Control, Deep Reinforcement Learning.**

I. Introduction

Active vision learning [1]–[6] is an emerging field at the intersection of robotics and computer vision, focused on developing intelligent systems that actively explore and perceive their environments. By leveraging reinforcement learning (RL) [7]–[11] and deep neural networks [12]–[14], significant progress has been made in enabling robots to make informed decisions based on visual input. The ultimate goal is to allow robots to effectively adapt to new and dynamic environments, improving detection capabilities while reducing reliance on extensive data collection.

In recent years, 3D object detection [15]–[19] has been crucial for enabling robots to perceive and understand the spatial layout of their environments. Among these, DeepGCNs [20] utilizes the geometric relationships between 3D point clouds for accurate and robust object detection. VoteNet [21] learns to generate a set of 3D object proposals and then optimizes these proposals using a voting mechanism to estimate the final object pose. Another significant advancement is 3D-DETR [22], which employs a transformer-based architecture to directly predict 3D object bounding boxes from RGB images and associated point clouds. However, these algorithms heavily rely on sufficient and meticulously curated datasets for training, which limits their adaptability to unknown environments. When robots equipped with these detectors enter unfamiliar environments, their performance can significantly deteriorate due to the lack of prior knowledge.

Thus, it is crucial to develop effective motion strategies that allow robots to collect new training samples in novel environments. These strategies should enable robots to actively explore and perceive the new environment, gather diverse and representative data, and adapt their detection models to achieve reliable performance in previously unseen scenarios. Real 3D Embodied Dataset (R3ED) [23] introduces a real 3D embodied dataset to address the performance degradation issue associated with using synthetic data in real-world scenarios. Another study [24] models active object detection as a sequential action decision process and introduces a deep reinforcement learning framework. However, these methods do not sufficiently consider historical and spatial positional information, which may lead to an incomplete understanding of the environment.

To overcome this challenge, our approach introduces a state representation rich in contextual information as the input state for the DQN [7]. Specifically, we store the robot's historical movement trajectories in a buffer and use a sliding window mechanism to select the most recent observations for state encoding. To more effectively integrate observations across time frames, we employ a standard transformer encoder for the encoding task. Additionally, we have newly designed the action space and incorporated a dueling network architecture,

This work was supported by the Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016). *(Corresponding author: Ping Zhang.)*

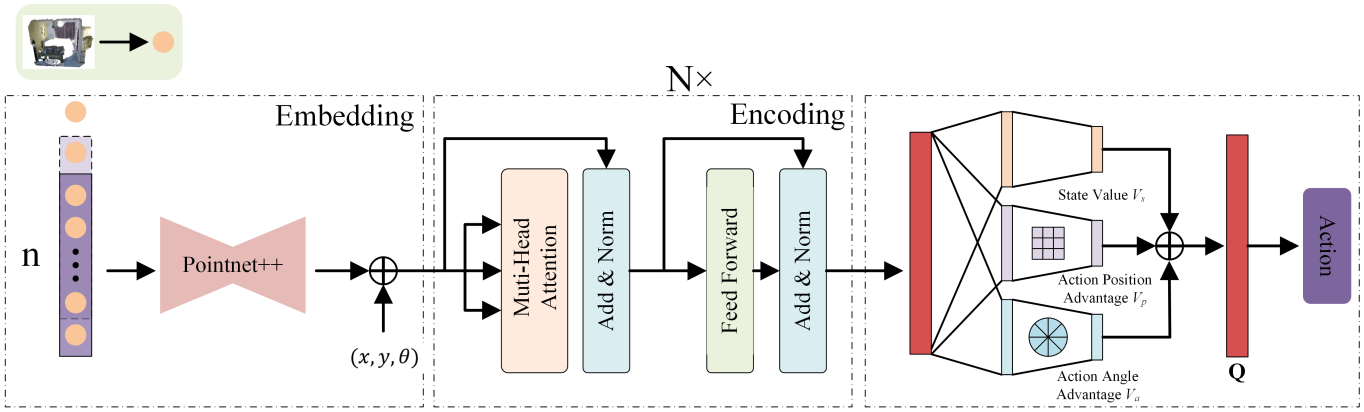

Fig. 1. The proposed active visual learning framework. The point cloud data within the sliding window are sequentially subjected to pointnet++ feature extraction followed by spatial location embedding operation. Then the features will be encoded by transformer encoder. Finally an action policy is generated by a dueling dqn network.

allowing the robot to independently assess the current state and potential action choices. We evaluated our method on the R3ED [23] collected in real-world scenarios and compared it with other benchmark methods. The experimental results demonstrate that our approach performs exceptionally well. Our contributions can be summarized as follows:

1) We propose a novel reinforcement learning method in active vision learning, leveraging a transformer encoder to enhance object detection in unknown and dynamic environments, empowering robots with active exploration and perception capabilities.
2) The proposed reward design enables robots to prioritize object category exploration and perform real-time evaluation of detection performance at the current location, achieving superior results on the R3ED dataset.

II. Background

We use a six-tuple $(S, A, R, p, \gamma, \pi)$ to describe the sequential action decision process for active visual learning of the robot, and they mean the following:

1) S is the set of states, where the robot can obtain state $s \in S$ at each time step.
2) A is the set of actions, and at each time step, the robot takes action $a \in A$ according to the current state s.
3) R is the reward, which is the sum of the rewards that the robot can receive for performing action a according to state s.
4) $p : S \times A \to S$ is the state transfer probability.
5) $\gamma \in [0, 1]$ is the discount factor.
6) $\pi : S \to A$ is an action policy that guides the robot to choose an action based on the current state.

The robot is equipped with a pre-trained detector to enter a new environment, and in the initial state, the robot acquires state s_0. In each subsequent time step, the robot selects an action $a_t = \pi(s_{t-1})$ based on the policy and the state s_{t-1} from the previous step. The objective of active vision learning is to find an optimal policy π^*, which recursively adjusts the state of the robot to obtain a better data collection trajectory.

III. Method

We propose a reinforcement learning method based on Dueling Deep Q-Networks [8]. As illustrated in Fig. 1, our approach involves processing individual point clouds within a sliding window using PointNet++ [25] for feature extraction. The resulting feature representations are then position-embedded to incorporate spatial information. Subsequently, the data stream is encoded using a transformer encoder module. Finally, the action policy is derived through the RL network.

A. State Representation

In active vision learning, the state representation is crucial for capturing relevant information about the robot's environment. While previous approaches often relied solely on the robot's current position, our method takes a more selective approach by considering a subset of historical point cloud frames along with the current frame. To achieve this, we establish a cache pool to store a collection of historical point cloud frames. The selected subset is then processed using PointNet++ [25] within a sliding window for feature extraction.

$$S_t = TE(F_{extract}(PC_{t-k:t}) \oplus PE(x, y, \theta)), \tag{1}$$

where $\oplus$ denotes vector concatenation, TE represents the Transformer Encoder, $PC_{t-k:t}$ denotes a sequence of point clouds within a sliding window of size k, $F_{extract}$ refers to feature extraction using PointNet++ [25], PE represents position embedding, and (x, y, θ) represents the current robot's position information.

By integrating PointNet++ for feature extraction and leveraging the Transformer encoder module with positional embedding, our state representation captures both local geometric details and global spatial dependencies. This comprehensive representation enables the robot to make informed decisions based on immediate visual input, historical context, and spatial relationships within the environment.

B. Action Representation

In our approach, we aim to enhance the robot's motion selection capabilities by dividing the action space into two components: position selection and camera angle adjustment. The position selection component focuses on choosing the robot's next position relative to its current location, allowing movement in eight directions: up, down, left, right, upper left, lower left, upper right, and lower right, as well as remaining in the current position. Meanwhile, the camera angle adjustment component enables the robot to modify its viewing angle in 45-degree intervals, providing eight possible angle options. Additionally, at time t, the robot maintains a feasible set $A_t \subseteq A$ of candidate actions to avoid collisions.

Inspired by the dueling DQN [8] architecture, we incorporate a similar design into our framework for action representation, as illustrated in Fig. 1. This architecture consists of three main components: state value V_s, action position advantage V_p, and action angle advantage V_a. Consequently, the output Q-value in reinforcement learning can be expressed as:

$$Q(s,a;\theta) = V_s(s;\theta_t,\theta_s) + (V_p(s,a_p;\theta_t,\theta_p) + (V_a(s,a_a;\theta_t,\theta_a) - \frac{1}{N_p}\sum_{a'_p \in A_p} V_p(s,a'_p;\theta_t,\theta_p)) - \frac{1}{N_a}\sum_{a'_a \in A_a} V_a(s,a'_a;\theta_t,\theta_a)), \tag{2}$$

where the parameters of the common encoding parts are denoted as θ_t, while the parameters θ_s, θ_p, and θ_a correspond to the three branches of the fully connected layers. N_p and N_a are the number of selected positions and angle actions, respectively.

C. Reward Function

Our reward function comprises three components. The first component focuses on the number of detectable objects in the current point cloud frame, where a higher number of detected objects results in a greater reward. It can be expressed as:

$$r_t^k = \lambda_k \times n_t, \tag{3}$$

where λ_k is an adjustable weighting factor, and n_t denotes the number of detectable objects in the point cloud collected at time step t.

The second component considers the entropy of the object classes observed in the historical collection of point cloud frames. We calculate the entropy by examining the set of all object classes and their respective frequencies in the historical frames. When a new point cloud frame is obtained at time step t, we update the set of object categories and calculate the change in entropy compared to the previous state as a reward:

$$r_t^e = \lambda_e \times (\mathcal{H}_t - \mathcal{H}_{t-1}), \tag{4}$$

where λ_e is an adjustable weighting factor, and $\mathcal{H}_t$ denotes the entropy of the object classes in the collected dataset at time step t. A higher entropy reward indicates the presence of a greater variety of object classes in the environment, reflecting higher environmental diversity. This encourages the robot to discover new objects during exploration, leading to a more comprehensive understanding of the structure and content of the environment.

TABLE I
THE DIVISION OF THE EXPERIMENTAL DATASET.

	pre-train scans				train scans		test scans
split1	Home_1	Home_2	Home_3	Home_4	Home_5	Home_6	Home_7
split2	Home_2	Home_3	Home_4	Home_5	Home_6	Home_7	Home_1
split3	Home_3	Home_4	Home_5	Home_6	Home_7	Home_1	Home_2
split4	Home_4	Home_5	Home_6	Home_7	Home_1	Home_2	Home_3

The third component, known as the performance difference reward, aims to assess the disparity between two detectors. We have trained two VoteNet [21] detectors: detector 1, which has no knowledge of the reinforcement learning training environment, and detector 2, which incorporates data from that environment during training. Our goal is to use policy learning to collect a small number of samples and enhance detector 1's performance to match that of detector 2 in the training environment.

At each time step t, both detectors are applied to the current point cloud frame, and the results are filtered based on a predefined classification score threshold. The remaining detection results for each class are then used to construct a cost matrix, where the elements represent the Intersection over Union (IoU) between corresponding bounding boxes from the two detectors. We employ a binary matching algorithm based on the cost matrix to determine the matches. The sum of IoU values from the matched pairs is computed to represent the performance difference between the detectors. The reward value of the third component can be expressed as:

$$r_t^p = -\lambda_p \times \frac{1}{n_t}\sum_c IOU(V_1, V_2) \tag{5}$$

where λ_p is an adjustable weighting factor, n_t is the number of objects in the point cloud frame at time step t, c represents the class of detected objects, and V_1 and V_2 are the two VoteNet [21] detectors, respectively. By considering the performance differences between detectors, we can gain insight into their effectiveness in a given environment. Higher IoU values indicate that the detection results of both detectors are very similar, suggesting a lower need for further exploration in that region. Conversely, lower IoU values indicate significant differences between the detection results of the two detectors, highlighting the importance of further exploration in that region.

Combining the above three components of the reward, the reward that can be obtained by the robot taking action a_t in a single time step t is:

$$r_t = r_t^k + r_t^e + r_t^p \tag{6}$$

IV. EXPERIMENTS

In this experiment, we aim to assess the efficacy of our proposed active vision learning approach in an unfamiliar environment. The robot is equipped with the pre-trained detector

TABLE II
PERFORMANCE COMPARISON OF ACTIVE VISUAL LEARNING POLICIES ON THE R3ED [23] DATASET.

Policy	mAP@0.25				mAP@0.5			
	split1	split2	split3	split4	split1	split2	split3	split4
Pre-train	19.08	18.91	20.43	19.57	8.65	8.21	9.03	7.96
Random	42.78	41.62	43.12	43.74	18.14	17.89	18.21	19.33
Unidirectional	49.07	48.19	49.64	48.69	23.07	22.85	23.51	24.17
Maximum distance	53.18	54.76	55.41	55.27	24.86	24.63	24.97	24.65
Semantic curiosity [1]	60.32	59.48	60.89	61.73	26.75	26.03	27.17	28.12
3D Divergency [23]	63.14	62.79	63.80	63.94	27.28	26.73	28.02	28.03
Proposed	**67.12**	**66.28**	**66.63**	**67.49**	**30.03**	**29.96**	**31.45**	**31.74**
improvement	*+3.98*	*+3.49*	*+2.83*	*+3.55*	*+2.75*	*+3.23*	*+3.43*	*+3.71*
w/o Contextual information	61.73	59.65	60.93	61.57	27.56	27.14	27.39	27.96
w/o Dueling Design	65.27	64.13	65.19	65.68	28.45	28.06	29.37	29.56

to develop a 100-step data acquisition strategy in the training scene and then evaluate its performance in the test scene. The task was trained on a computer with a 14-core Intel i9-10940X CPU and a 24GB NVIDIA GeForce RTX 3090 GPU.

A. Dataset

The experiments in this thesis use the Real 3D Embodied Dataset (R3ED) [23]. This dataset contains over 5,800 point clouds and 22,400 ground truth 3D bounding boxes. R3ED was collected from seven realistic indoor scenes using a dense sampling method that enables the simulation of robot actions and the acquisition of realistic data. Each room features more than 100 densely distributed sampling points, with each point capturing data from 8 different angles. We evaluate the performance of various active vision learning policies on the R3ED dataset using three different splits: split1, split2, and split3 (as described in Table I). We will train a VoteNet [21] detector in a pre-training scenario, then apply our active vision learning policy in a training scenario, and validate it in a test scenario. The metrics used for evaluation include mAP@0.25 and mAP@0.5.

B. Baseline Polices

To show the effectiveness of the proposed approach, we compared it with several benchmark strategies as follows:

1) Random Policy: The robot takes 100 random steps in a new environment.
2) Unidirectional Policy: The robot moves in one direction until it encounters an obstacle and then changes direction.
3) Maximum distance Policy: Robot moves maximum distance with 100 steps movement in new environment.
4) Semantic curiosity policy [1]: Training robot motion strategies based on the reward formula in [1]
5) 3D Divergency Policy [23]: Training robot motion strategies based on the reward formula in [23]

C. Performance Comparison

As shown in Table II, the results indicate that the pre-trained model has limited performance in novel environments, highlighting the need for active visual learning approaches. Next, we evaluate several active visual learning policies. Notably, our proposed policy outperforms all other policies across all splits. It achieves the highest mAP@0.25 scores of 67.12, 66.28, 66.63 and 67.49 for splits 1, 2, 3 and 4, respectively. Similarly, it achieves the highest mAP@0.5 scores of 30.03, 29.96, 31.45 and 31.74 for the respective splits.These results demonstrate the effectiveness of our proposed policy in improving object detection performance in unseen environments. The proposed policy leverages a combination of active exploration and perception strategies, enabling the robot to collect diverse and informative data to refine its detection model.

D. Ablation Study

We conducted two ablation studies. As shown in Table II, in the first study, we removed the sliding window and transformer encoder, using only the current viewpoint observation as input and ignoring contextual information. The experimental results revealed that, without the support of historical data, the robot tended to repeat its movements, leading to a significant decrease in performance. In another study, we replaced our proposed dueling network architecture with a traditional DQN [7] algorithm. In this case, although the performance in the test environment dropped by 2 mAP, it still outperformed the baseline methods, further emphasizing the importance of considering contextual information in this task.

V. CONCLUSION

This study presents a novel approach to active vision learning. By leveraging reinforcement learning and deep neural networks, our method endows robots with the ability to actively explore and perceive their environments, thereby enhancing 3D object detection performance in unknown environments. The integration of a transformer-based encoder and sliding window technique enables the robot to combine historical information with spatial cues, leading to more informed decision-making. Future work could focus on further optimizing the active vision learning process, expanding datasets, and exploring the integration of large language models to enhance the robot's ability to actively perceive and understand the world.

REFERENCES

[1] D. S. Chaplot, H. Jiang, S. Gupta, and A. Gupta, "Semantic curiosity for active visual learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 309–326.

[2] G. Chaudhary, L. Behera, and T. Sandhan, "Active perception system for enhanced visual signal recovery using deep reinforcement learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] S. K. Ramakrishnan, D. Jayaraman, and K. Grauman, "An exploration of embodied visual exploration," vol. 129. Springer, 2021, pp. 1616–1649.

[4] H. Zeng, P. Zhang, C. Wu, J. Wang, T. Ye, and F. Li, "Mars: Multimodal active robotic sensing for articulated characterization," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 1634–1642, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/181

[5] M. Mattamala, M. Ramezani, M. Camurri, and M. Fallon, "Learning camera performance models for active multi-camera visual teach and repeat," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 14 346–14 352.

[6] E. Safronov, N. Piga, M. Colledanchise, and L. Natale, "Active perception for ambiguous objects classification," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4437–4444.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[8] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1995–2003.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[10] H. Zeng, P. Zhang, F. Li, C. Lin, and J. Zhou, "Ahegc: Adaptive hindsight experience replay with goal-amended curiosity module for robot control," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[11] T. Ye, P. Zhang, H. Wang, H. Zeng, J. Wang, and T. Zeng, "Reinforcement learning-driven dual neighborhood structure artificial bee colony algorithm for continuous optimization problem," *Applied Soft Computing*, p. 112601, 2024.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Z. Zhang, B. Sun, H. Yang, and Q. Huang, "H3dnet: 3d object detection using hybrid geometric primitives," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 311–329.

[16] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu, and J. Wang, "Mlcvnet: Multi-level context votenet for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 447–10 456.

[17] D. Rukhovich, A. Vorontsova, and A. Konushin, "Fcaf3d: fully convolutional anchor-free 3d object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Springer, 2022, pp. 477–493.

[18] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3d object detection via transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2949–2958.

[19] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-temporal self-supervised representation learning for 3d point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6535–6545.

[20] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?" in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9267–9276.

[21] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[22] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2906–2917.

[23] Q. Zhao, L. Zhang, L. Wu, H. Qiao, and Z. Liu, "A real 3d embodied dataset for robotic active visual learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6646–6652, 2022.

[24] X. Han, H. Liu, F. Sun, and X. Zhang, "Active object detection with multistep action prediction using deep q-network," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3723–3731, 2019.

[25] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.