Theory and Practice of Science and Technology
2025, VOL. 6, NO. 5, 68-73
DOI: 10.47297/taposatWSP2633-456915.20250605

Research on UAV Perception Systems for Dynamic Indoor Environments: PointNet++ Optimization with Hierarchical Contextual Attention

Yuhan Jiang*

Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, 999077, China

ABSTRACT

This research addresses the challenges of dynamic perception systems for UAVs in indoor environments by proposing an optimization method for PointNet++ that integrates hierarchical contextual attention mechanisms. To address the issue of imprecise boundary feature extraction in traditional LiDAR point cloud processing for dynamic object detection and tracking, we designed a Hierarchical Contextual Attention Feature Propagation (HCAFP) module. This module establishes weighted connections between high-level semantic features and low-level geometric features, significantly enhancing the model's perception capabilities for object part boundaries. Experimental results show that the enhanced model achieved an instance average IoU of 86.7% on the ShapeNetCore dataset, an improvement of 1.3 percentage points over the original PointNet++ MSG architecture. Additionally, we proposed a series of model optimization strategies, including model compression and quantization, lightweight model structure design, and hierarchical sparse attention mechanisms, enabling the improved model to be efficiently deployed on UAV platforms with limited computational resources while maintaining competitive performance (85.3% instance average IoU). This research provides a new technical approach for UAV autonomous navigation and dynamic object perception in indoor environments, with significant implications for UAV applications in complex dynamic indoor settings.

KEYWORDS

3D point clouds; Object Tracking, Part segmentation, Hierarchical contextual attention

1 Introduction

1.1 Research Background

Autonomous mobile robots refer to mechanical devices that can autonomously locate, plan paths, and move in an environment through their own perception, decision-making, and control systems[1]. Relying on various sensors and embedded systems, these robots perceive their surroundings and use built-in algorithms for path planning and obstacle avoidance, achieving safe and efficient mobility.

As typical mobile robots, quadrotor UAVs are rapidly integrating into human society due to their low cost and high mobility. They have been widely applied in agricultural monitoring, warehouse logistics, disaster rescue, and other scenario[2]. However, the dynamic nature and complexity of modern environments require robots to have more advanced autonomous perception capabilities, such as UAV autonomous following and human-robot collaboration scenarios. In indoor scenarios, UAVs cannot rely on GPS. Indoor environments pose unique challenges to perception systems due to their complex spatial layouts, dynamic objects (humans, pets, movable furniture), and confined spaces. Such environments contain dense dynamic targets, such as personnel and equipment, with diverse movement patterns (such as sudden stops and turns), making autonomous target tracking and trajectory prediction challenging for UAVs[3].

Traditional perception technologies have certain limitations in capturing rapidly changing environmental features and cannot comprehensively assess surrounding dynamic objects and changing conditions. For example, visual sensors can easily produce motion blur in low light conditions, lack of texture, or rapid movement, making it difficult to accurately estimate the three-dimensional position of targets. Infrared or TOF sensors are susceptible to environmental temperature or reflection interference, have low resolution, and struggle to distinguish dense targets[4-6]. Traditional computer vision methods often perform poorly in these scenarios due to changing lighting conditions, occlusions, and the need for depth information. LiDAR (Light Detection and Ranging) technology has become the main sensing modality for such applications due to its ability to generate high-precision 3D point clouds under various lighting conditions. LiDAR-based systems solve many of these challenges by providing direct 3D spatial information, but they also introduce new computational demands for point cloud processing. Real-time processing of these unstructured point clouds poses significant challenges for platforms with limited computational resources, such as small UAVs.

The PointNet++ architecture with Multi-Scale Grouping (MSG) has demonstrated excellent capabilities in processing unordered point sets.[7] However, the standard implementation has limitations in boundary feature extraction, leading to poor performance in segmentation tasks. This shortcoming is particularly critical for indoor autonomous navigation, as precise target boundary detection is essential for collision avoidance and interaction planning.

*** Corresponding Author:** jiangyuhan819@@163.com

This research is dedicated to meeting the demands of UAVs for dynamic target detection in rapidly changing and diverse indoor scenarios, deeply investigating the key issues of indoor quadrotor UAV dynamic target tracking and trajectory prediction. We significantly improve the capability of boundary region feature extraction while maintaining computational efficiency suitable for edge computing platforms such as UAVs by integrating hierarchical contextual attention mechanisms into the PointNet++ MSG architecture. This enhancement is particularly important for applications requiring fine understanding of object parts and structures in dynamic indoor environments, providing key technical support for UAVs to achieve safer and more precise indoor autonomous navigation and interaction.

1.2 Research Overview

UAV applications in indoor environments place strict requirements on perception systems, including real-time processing capability, high-precision environmental perception, and low-power operation. Wu et al. (2020) explored a series of lightweight neural network architectures for UAV platform computational limitations[8], while Tanaka et al. (2020) focused on model compression techniques to enable efficient operation of deep learning algorithms on resource-limited edge devices[9].

For the special challenges of indoor UAV navigation, multiple studies have proposed LiDAR-based solutions. Holmberg and his team (2022) proposed a LiDAR-based indoor precise positioning and navigation method that integrates inertial data, low-drift LiDAR odometry, plane feature extraction, and loop detection in a graph optimization-based framework. They particularly focused on lateral positioning accuracy when vehicles pass through narrow openings and conducted experimental evaluations using high-resolution LiDAR, comparing with state-of-the-art methods such as LIO-SAM and Cartographer[10]. This technology can be applied to small autonomous vehicles operating in unknown indoor environments, providing an effective solution for precise navigation under GPS-limited conditions. Ren et al. (2025) developed a safety-assured high-speed navigation system for micro aerial vehicles (MAVs). The system adopts a LiDAR-based sensing solution, achieving fully onboard perception and computational capabilities, enabling UAVs to fly safely at speeds exceeding 20 meters per second in unknown complex environments. Their proposed dual-trajectory strategy strikes a balance between ensuring safety and increasing speed, allowing UAVs to intelligently plan paths and avoid obstacles. This technology has important value for time-critical applications such as search and rescue and disaster relief, achieving autonomous navigation performance close to bird flight capabilities[6].

However, existing research still faces significant challenges in balancing model complexity and performance, especially for tasks requiring fine part segmentation and precise boundary detection. Traditional methods struggle to achieve accurate feature extraction while maintaining high computational efficiency, while complex models are difficult to run in real-time on resource-constrained platforms such as UAVs.

2 Methodology

2.1 Related Work

2.1.1 PointNet++ Algorithm

PointNet++ represents a key advancement in point cloud processing, effectively extending the capabilities of the original PointNet through a hierarchical feature learning framework. This architecture is specifically designed to process unordered point set data, capable of capturing multi-scale features from local details to global structures[7]. The PointNet++ architecture consists of five key modules working collaboratively: the sampling layer uses farthest point sampling algorithm to select representative point subsets, ensuring uniform coverage of the original point cloud geometric structure; the grouping layer constructs local neighborhoods around each sampled point, capturing local geometric features; the PointNet layer serves as a local feature extractor, transforming raw coordinates into abstract feature representations; the set abstraction module integrates the aforementioned components, building a multi-level abstraction process, achieving progressive transformation from low-level geometry to high-level semantics; the feature propagation module propagates features from downsampled points back to the original point cloud through interpolation. Additionally, the Multi-Scale Grouping (MSG) variant of PointNet++ effectively solves the problem of point cloud density variation by applying grouping operations of different radii to each center point, simultaneously capturing local structures at multiple spatial scales, enabling the network to utilize large-scale features in sparse point areas while preserving fine structural information in dense point areas.

2.1.2 Hierarchical Attention Networks

This research proposes to introduce hierarchical contextual attention mechanisms into the PointNet++ architecture, addressing issues such as imprecise boundary feature extraction, information bottlenecks, and limited context integration capabilities by establishing feature propagation channels between different levels. This mechanism contains four core concepts: hierarchical feature integration breaks information silos in traditional feature propagation; context weighting mechanism introduces dynamic assessment based on feature correlation; enhanced focus on target part boundary features improves segmentation accuracy; adaptive feature propagation mechanism uses feature similarity as a weight factor to make propagation more precise. These innovations bring significant advantages: enhanced boundary representation capability, improved ability to capture long-distance dependencies, efficient utilization of computational resources, and effective integration of geometric features and semantic features.

2.2 Model Optimization

2.2.1 Quantization-aware Architecture Design

We implemented a quantization-aware design for our model, systematically integrating quantization support modules to optimize model performance and computational efficiency. Specifically, we introduced QuantStub and DeQuantStub modules in the network architecture, enabling the model to inherently support 8-bit integer quantization. Unlike traditional PointNet++, this optimized design adopts an explicit layer fusion strategy, unifying convolution layers, batch normalization layers, and activation functions to build an efficient inference computation pipeline. This method, based on the quantized neural network acceleration theory proposed by Jacob et al. (2018) [11], not only significantly reduces memory access overhead but also optimizes cache utilization on edge devices.

2.2.2 Sparse Hierarchical Attention Mechanism and Feature Propagation Enhancement

The most significant improvement in this research is the enhancement in attention mechanisms and feature propagation, primarily reflected in innovative designs at several key levels. We proposed the Hierarchical Attention Feature Propagation module, which systematically replaces the original PointNet Feature Propagation structure in PointNet++. This enhanced module integrates three complementary attention mechanisms: spatial attention, channel attention, and context fusion, achieving adaptive extraction and propagation of point cloud features through a query-key-value architecture.

Notably, we introduced a sparse attention mechanism equipped with learnable temperature parameters, significantly enhancing the model's perception capability for point cloud spatial structures. To adapt to resource-constrained environments, this research implemented a series of lightweight efficiency optimization strategies: adopting grouped convolution techniques (groups=4) to reduce parameter count; implementing attention dimension reduction; and introducing sparse computation mechanisms, effectively reducing computational complexity by thresholding and ignoring redundant attention weights. Additionally, we integrated a channel attention module inspired by Squeeze-and-Excitation, which achieves dynamic reweighting of feature channels through adaptive pooling and nonlinear transformations. Tests show that these designs improve inference efficiency and accuracy while maintaining the model's representation capability, providing a feasible solution for point cloud processing in computationally resource-constrained scenarios.

2.2.3 Training Strategy Optimization

Based on the knowledge distillation theory proposed by Hinton et al. (2015)[12], we constructed a distillation framework specifically applicable to point cloud segmentation tasks. This framework achieves model compression and effective performance transfer through systematic design. First, we implemented a soft target injection mechanism, generating soft labels from the teacher model through softmax operations modulated by high-temperature parameters, providing the student model with richer inter-class similarity information than hard labels. Second, we designed a balanced loss function, introducing an α parameter to dynamically adjust the weight ratio between hard label loss and distillation loss, enabling the student model to both maintain accuracy for the original task and effectively absorb the implicit knowledge of the teacher model. Additionally, through the regulation of temperature parameters, we optimized the smoothness of soft labels, enhancing the efficiency and effectiveness of knowledge transfer.

We also integrated an adaptive early stopping mechanism based on validation set performance, achieving dynamic optimization of the training process through systematic monitoring of the Intersection over Union (IoU) metric on the validation set. This mechanism constructs stopping criteria using configurable patience value parameters and minimum gain thresholds, automatically terminating the training process when performance improvement falls below the preset threshold for multiple consecutive evaluation rounds, preventing model overfitting phenomena and saving computational resources and training time.

3 Experimental Evaluation

3.1 Dataset

3.1.1 Data

This research uses the *ShapeNetCore* dataset for model evaluation, which is one of the standard benchmarks for point cloud segmentation tasks.

This dataset contains 51, 300 3D models across 55 categories, covering various everyday items such as furniture, vehicles, and electronic devices. Due to its high quality and fine-grained annotations, it is widely used in computer vision and graphics for 3D shape classification, retrieval, reconstruction, and other tasks. We follow the standard division method, splitting the dataset into training (80%), validation (10%), and test (10%) sets. Each point cloud is sampled to N= 2048 points, including 3D coordinate information.

3.1.2 Evaluation Metrics

To comprehensively evaluate model performance, we adopt the following metrics:

a) Part mean Intersection over Union (*mIoU*)**:** Calculate the *IoU* for each part, then average over all parts. For part *p* of category *c*, *IoU* is calculated as:

$$\mathrm{IoU}_{c,p} = \frac{|P_{c,p} \cap G_{c,p}|}{|P_{c,p} \cup G_{c,p}|}$$

where $P_{c,p}$ represents the set of points predicted as part *p* of category *c*, and $G_{c,p}$ represents the set of points with true labels as part *p* of category *c*.

b) Instance average *IoU* (Inst. avg IoU): First calculate the average *IoU* of all parts for each instance, then average over all instances:

$$\text{Inst. avg IoU} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{|P_i|} \sum_{p \in P_i} \mathrm{IoU}_{c_i,p}$$

where $\mathcal{D}$ is the test set，P_i is the set of parts for the i-th instance，c_i is the category of the i-th instance。

c) Category average *IoU* (Cat. avg IoU): First calculate the average *IoU* of all instances for each category, then average over all categories:

$$\text{Cat. avg IoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \frac{1}{|P_i|} \sum_{p \in P_i} \mathrm{IoU}_{c,p}$$

where $\mathcal{C}$ is the set of all categories, and $\mathcal{D}_c$ is the set of all instances of category *c*.

3.1.3 Experimental Configuration

Table 3.1 Hardware and Software Environment

Category	Configuration
Hardware	GPU: NVIDIA L20 48GB
	CPU: 20 vCPU Intel(R) Xeon(R) Platinum 8457C
	RAM: 100GB
Software	PyTorch 2.0.1
	CUDA 11.8
	Python 3.9.16

Table 3.2 Training Configuration

Parameter	Setting
Batch Size	16
Optimizer	Adam (β_1=0.9, β_2=0.999)
Initial Learning Rate	0.001, decays by a factor of 0.5 every 20 epochs
Number of Epochs	Maximum 250, with early stopping (patience=7)
Data Augmentation	Random scaling (±0.2), random translation (±0.2)

3.2 Experimental Results and Analysis

3.2.1 Segmentation Accuracy Comparison

The experimental results shown in Table 3.4 demonstrate the effectiveness of the proposed Hierarchical Contextual Attention (HCA) mechanism in point cloud part segmentation tasks. Quantitative analysis shows that the HCA-enhanced version significantly outperforms the original PointNet++ (MSG version) architecture on multiple evaluation metrics. Specifically, the HCA variant achieved an instance average IoU of 86.7%, an improvement of 1.3 percentage points over the baseline model's 85.4%. Similarly, in the category average IoU evaluation, the HCA model reached 83.5%, exceeding the original architecture's baseline of 82.6% by 0.9 percentage points.

Table 3.3 Model Configuration and Performance Comparison

Feature	Original PointNet++ (Msg)	Improved PointNet++ (HCA)	Lightweight PointNet++ (HCA-Light)
Set Abstraction Layer	Multi-Scale Grouping	Same as original version	Lightweight design (halved sampling points, reduced feature channels)
Feature Propagation Layer	Standard PointNet feature propagation layer	Hierarchical Attention Feature Propagation	HierarchicalAttentionFeaturePropagation
Attention Configuration	Not applicable	Attention dimensions set to 192, 128, and 64	Sparse attention computation (threshold=0.1)
Knowledge Distillation	Not applicable	Not applicable	Applied (temperature=3.0, α=0.4)
Precision Setting	Single precision (FP32)	Single precision (FP32)	Mixed precision (FP16/FP32)
Training Time	12 hours	14 hours	7 hours

Table 3.4 shows the segmentation accuracy comparison of different models on the ShapeNetCore test set.

Table 3.4 The segmentation accuracy (%) of different models on the ShapeNetCore test set

Model	Inst. avg IoU	Cat. avg IoU	AirPlane	Bag	Cap	Car	Chair	Earphone	Guitar	Knife
Original PointNet++ (Msg version)	85.4	82.6	83.5	81.1	89.4	79.1	90.7	71.6	91.3	87.0
Improved PointNet++ (HCA)	86.7	83.5	84.2	82.3	90.1	80.2	90.8	73.1	92.1	88.2
Lightweight PointNet++ (HCA-Light)	85.3	82.2	83.7	81.5	89.8	79.3	89.9	72.0	91.7	87.4

Table 3.4 The segmentation accuracy (%) of different models on the ShapeNetCore test set (Continued)

Model	Inst. avg IoU	Cat. avg IoU	Lamp	Laptop	Motorbike	Mug	Pistol	Rocket	Skateboard	Table
Original PointNet++ (Msg)	85.4	82.6	84.2	95.6	73.6	95.3	81.5	58.9	76.4	82.5
Improved PointNet++ (HCA)	86.7	83.5	82.7	96.2	75.8	95.9	83.9	61.3	78.2	84.1
Lightweight PointNet++ (Msg version)	85.3	82.2	81.5	95.8	74.5	95.5	82.8	59.4	76.9	83.6

Further analysis of performance across individual categories shows consistent improvements in most object classes. This enhancement is particularly evident in categories with complex geometric structures, such as guitars (92.1% versus 91.3%), mugs (83.9% versus 81.5%), and motorcycles (75.8% versus 73.6%). This pattern suggests that the hierarchical attention mechanism effectively captures complex local-global relationships in point clouds, achieving more precise boundary delineation in segmentation tasks. The performance improvement on objects with complex part arrangements validates our hypothesis that context-aware feature propagation enhances the model's ability to discern subtle geometric transitions between adjacent parts.

Notably, the lightweight variant (HCA-Light), despite significantly reduced parameter count and computational complexity, maintained competitive performance (85.3% instance average IoU). From a practical deployment perspective, this finding is particularly important as it indicates that the proposed optimization strategies—including sparse attention computation, dimensionality reduction, and knowledge distillation—effectively preserved the model's representational capabilities while significantly improving computational efficiency. The minimal performance degradation of the lightweight model (only 0.1 percentage points lower than the original architecture) highlights the effectiveness of our proposed architectural modifications and training methods.

4 Conclusion

This research introduces an innovative hierarchical contextual attention mechanism for the PointNet++ MSG architecture, significantly improving feature extraction capabilities for object part boundaries. Our method demonstrates significant performance improvements on the ShapeNetCore benchmark dataset, achieving an instance average IoU of 86.7%, an improvement of 1.3 percentage points over the original architecture. This progress is particularly evident in object categories with complex geometric structures, indicating that the proposed method effectively enhances the model's perception capabilities for detail-rich boundary regions.

Simultaneously, we developed model compression and lightweight structure design strategies, successfully deploying the model on UAV platforms with limited computational resources. Notably, our lightweight variant, despite a parameter reduction of approximately 50%, maintained an instance average IoU of 85.3%, only 0.1 percentage points lower than the original architecture, while significantly reducing computational requirements.

The enhanced feature representation capability enables the model to more precisely identify object part boundaries and maintain semantic consistency, which is crucial for UAV navigation applications requiring fine understanding of 3D scenes. Improved boundary feature extraction translates to more accurate dynamic object detection and tracking capabilities, providing a solid foundation for autonomous navigation of UAVs in complex indoor environments.

Despite the encouraging results of this research, we have also identified several limitations and challenges, including memory overhead introduced by attention mechanisms, increased training complexity, insufficient explainability, and limited temporal processing capabilities. These limitations point to important directions for future research, including developing temporal attention mechanisms, exploring multimodal fusion, introducing uncertainty estimation, designing self-supervised pretraining strategies, and developing hardware-aware architecture search methods.

Overall, this research represents an important step toward providing robust real-time 3D perception for autonomous UAVs in indoor dynamic environments. The proposed hierarchical contextual attention-enhanced PointNet++ model not only theoretically advances point cloud processing technology but also provides a feasible solution for UAV perception systems in practical applications, with the potential to bring positive impacts to fields such as indoor autonomous navigation, human-robot interaction, and security monitoring.

Acknowledgements

I would like to thank Professor Lu Peng for his guidance, and I would like to express my gratitude to my family and friends for their support and companionship.

References

[1] KATONA K, NEAMAH H A, KORONDI P. Obstacle Avoidance and Path Planning Methods for Autonomous Navigation of Mobile Robot[J]. Sensors (14248220), 2024, 24(11).

[2] FAN, BANGKUI, LI, et al. Review on the Technological Development and Application of UAV Systems[J]. Chinese Journal of Electronics, 2020, v.29(02): 5-13.

[3] DU W, BELTRAME G. LiDAR-based Real-Time Object Detection and Tracking in Dynamic Environments[J]. 2024.

[4] DE CARVALHO M V L, SIMONI R, YOSHIOKA L. Autonomous Navigation and Collision Avoidance for Mobile Robots: Classification and Review [J]. 2024.

[5] KLANAR G, SEDER M, BLAI S. Advanced Sensors Technologies Applied in Mobile Robot[J]. Sensors (14248220), 2023, 23(6).

[6] REN Y, ZHU F, LU G, et al. Safety-assured high-speed navigation for MAVs[J]. Science Robotics, 2025, 10(98).

[7] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[J]. 2017.

[8] FAHAD SAFDAR M, SHARIF H, REHMAN F, et al. A Review of Recent Advances in Deep Learning for Object Detection[C/OL]//2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS). 2022: 1-6. DOI: 10.1109/ ICONICS56716.2022.10100486.

[9] TANAKA H, KUNIN D, YAMINS D L K, et al. Pruning neural networks without any data by iteratively conserving synaptic flow[J]. 2020.

[10] HOLMBERG M, KARLSSON O, TULLDAHL M. Lidar Positioning for Indoor Precision Navigation[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022: 358-367. DOI:10.1109/CVPRW56347.2022.00051.

[11] JACOB B, KLIGYS S, CHEN B, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[12] HINTON G E, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J/OL]. ArXiv, 2015, abs/1503.02531. https://api. semanticscholar.org/CorpusID:7200347.