Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 3s (2025)

Reinforced Multi-Path Graph Attention Network (MPGA-RL-Net) for Adaptive Robotic Vision with Task-Specific Region Prioritization

Mr.B.Rama Subbaiah[1], Dr P Penchala Prasad[2], Mr.G.Vikram Chandra[3], Mr.M.Suleman Basha[4], Dr P Kiran Rao[5], Dr.R.Senthil Ganesh[6]

[1]Assistant professor, Department of Computer Science & Engineering, Rajeev Gandhi Memorial College of Engineering & Technology,Nandyal, Andhrapradesh, India, subhashrgmcet@gmail.com

[2]Associate Professor, Department of Computer Science and Engineering (Data Science), Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhrapradesh, India, p.prasad71@gmail.com

[3]Assistant Professor, Department of CSE (Data Science), Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhrapradesh, India, vikram.sam@gmail.com

[4]Associate Professor, Department of CSE (Data Science), Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhrapradesh, India, suleman.ndl@gmail.com

[5]Assistant Professor, Department of Computer Science and Engineering (Data Science), Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhrapradesh, India, kiranraocse@gmail.com

[6]Associate Professor, Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore

drsenthilganesh@gmail.com

Article History:

***Received**: 21-09-2024*

***Revised**: 30-10-2024*

***Accepted**: 16-11-2024*

Abstract:

Robotic vision plays a crucial role in enabling autonomous systems to perceive, understand, and interact with complex environ- ments. However, accurately segmenting and prioritizing visual regions of interest in dynamic scenes is challenging due to variations in object shapes, sizes, and spatial relationships. Traditional methods, such as Convolutional Neural Networks (CNNs) and Fully Convolutional Net- works (FCNs), have shown promise but often struggle to integrate both global context and fine-grained details required for tasks like object recognition and navigation. To address these challenges, we present a Multi-Path Graph Attention Network with Reinforcement Learning (MPGA- RL-Net) adapted for robotic vision. This framework leverages a Multi- Path Feature Extraction (MPFE) module to capture multi-scale features at low, medium, and high resolutions, fusing them using an adaptive attention mechanism that assigns task-specific weights to each resolu- tion level. Super pixel-based segmentation is then applied to the fused feature map, representing regions as graph nodes, with Graph Convo- lutional Networks (GCNs) employed to model spatial relationships be- tween regions. Reinforcement Learning (RL) is further incorporated to dynamically adjust attention, allowing the model to focus on critical ar- eas, such as target objects or pathways, in real-time. Performance eval- uations demonstrate that MPGA-RL-Net enhances accuracy in robotic vision tasks, particularly in cluttered or dynamically changing environ- ments, achieving higher precision in object detection and adaptive focus on critical regions compared to conventional methods

Keywords: Robotic Vision · Computer Vision · Graph Convolutional Network (GCN) · Reinforcement Learning · Multi-Scale Feature Extrac- tion · Multipath · Attention Mechanism.

1. INTRODUCTION

Robotic vision systems are critical for empowering autonomous robots with the ability to interpret, navigate, and interact with their surroundings. Applications in areas such as industrial automation, environmental monitoring, and healthcare demand a high level of precision and adaptability from robotic vi- sion models to ensure that robots can make informed, real-time decisions based on visual data [1]. Achieving such accuracy in robotic perception tasks requires models that can dynamically prioritize and focus on relevant regions in complex scenes. However, this level of adaptability remains a challenge for many existing approaches, as robotic vision tasks often involve significant variation in object shapes, sizes, and spatial relationships, particularly in unstructured environments [2]. Traditional computer vision systems have relied heavily on Convolutional Neural Networks (CNNs)[3] as shown in figure 1, which are highly effective at extracting visual features across various image domains. However, CNNs generally lack the flexibility to adapt focus dynamically in response to task requirements or environmental changes, which limits their applicability in complex robotic tasks where scene context and focus need to be continually re- assessed. In the context of segmentation, U-Net[4] and SegNet[5] architectures have shown effectiveness in balancing fine detail with broader spatial context, making them popular choices in both general and medical image segmentation.

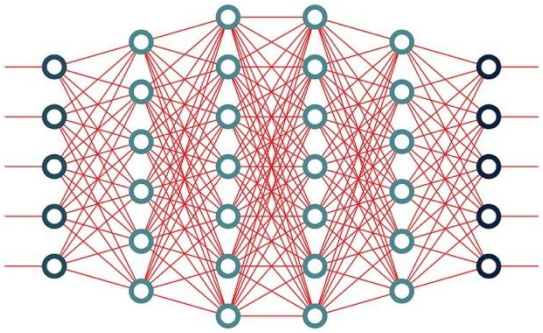

Figure 1. Applying neural networks to robotic vision and guidance

U-Net[4] and SegNet[5] architectures have shown effectiveness in balancing fine detail with broader spatial context, making them popular choices in both general and medical image segmentation. Yet, these models often struggle in scenarios requiring high precision, such as segmenting small, irregularly shaped objects or recognizing boundaries in clut- tered or dynamic environments. Such limitations highlight the need for models that can dynamically adapt to prioritize region-specific focus in real time[6].

To address some of these challenges, recent research has explored graph- based models, including Graph Neural Networks (GNNs) and Graph Convolu- tional Networks (GCNs)[7], for representing spatial relationships within scenes. By treating regions of interest as nodes within a graph, GCNs can aggregate spatial information across different parts of an image, providing an advantage in capturing complex relationships between objects. This approach has shown par- ticular promise in tasks like medical imaging, where intricate boundary detection is crucial. Although GCNs improve spatial representation and structure within scenes, they often employ static structures and attention mechanisms that lack the dynamic prioritization capabilities required in real-time robotic vision[8].

Moreover, fixed attention weights in these models do not adjust according to scene demands, a limitation that constrains their effectiveness in dynamically changing environments.

Motivation for the Proposed Work: Despite significant progress, current approaches in robotic vision face limitations in dynamically integrating global and local context for region-specific prioritization. While CNN-based architec- tures offer robust feature extraction capabilities, their static attention mecha- nisms limit adaptability to changing task requirements. In contrast, graph-based approaches provide a more structured way of modeling spatial relationships but are often restricted by fixed graph structures and attention that does not adapt to real-time feedback. These limitations become particularly evident in real-time applications requiring adaptive focus, such as a robot navigating through clut- tered environments or selecting optimal grasp points on objects with irregular shapes. Furthermore, many current methods do not fully leverage multi-scale feature integration, often resulting in incomplete segmentation and impacting task accuracy, especially in scenes with intricate spatial dependencies.

To overcome these challenges, we designed a framework, with Multi-Path Graph Attention Network with Reinforcement Learning[9] (MPGA-RL-Net), for robotic vision. MPGA-RL-Net integrates a multi-scale feature extraction module with graph-based spatial modeling and a reinforcement learning-driven attention mechanism that enables real-time prioritization of task-relevant regions. By in- corporating feedback-based reinforcement learning, MPGA-RL-Net dynamically adjusts its attention focus, concentrating on critical regions such as navigable paths or target objects, which enhances adaptability in robotic vision tasks.

Objectives of the Proposed Work: The primary objective of this study is to develop an adaptive robotic vi- sion framework capable of integrating both local and global features while dynamically adjusting focus on task-specific regions. The specific objectives include:

Multi-Scale Feature Extraction: To capture features across low, medium, and high resolutions using a Multi-Path Feature Extraction (MPFE) module, which enables the model to balance broad contextual information with fine- grained details essential for accurate segmentation and prioritization.

Graph-Based Spatial Relationship Modeling: To construct a graph-based representation using super pixel segmentation, allowing the model to capture complex spatial relationships among regions for improved context in robotic vision tasks.

Dynamic Attention Mechanism via Reinforcement Learning: To introduce a reinforcement learning component that dynamically adjusts attention weights, enabling the model to emphasize critical areas based on feedback in real time[10].

Precision in Region-Specific Prioritization: To refine model focus on task- critical regions, especially in complex or cluttered environments, ensuring that region-specific prioritization is optimized for successful task execution

2. RELATED WORKS

Recent advancements in deep learning and graph-based neural networks have enabled significant improvements in robotic vision, yet challenges in adaptive region prioritization and multi-scale

feature integration persist. This literature survey reviews existing work in CNN-based architectures, attention mechanisms, and graph-based approaches, highlighting their contributions and limitations concerning robotic vision tasks.

Convolutional Neural Networks in Robotic Vision: Convolutional Neural Networks (CNNs) have been foundational in advancing robotic vision due to their powerful feature extraction capabilities across various visual do- mains. Krizhevsky et al.'s work on AlexNet demonstrated the potential of deep CNNs for high-accuracy object classification, setting a precedent for CNN ap- plication in robotics. Since then, numerous CNN-based architectures, including ResNet and VGGNet, have been applied to robotic vision tasks. These models effectively capture hierarchical image features, but they rely on static feature ex- traction, which limits their adaptability to task-specific focus areas in dynamic scenes[11].

In segmentation tasks, U-Net and SegNet have shown impressive results in preserving spatial context through encoder-decoder architectures with skip con- nections. U-Net, for example, has become a preferred model in medical imaging due to its ability to capture fine-grained details while retaining global context, which is essential for boundary-sensitive tasks like organ segmentation. How- ever, these models struggle with complex environments where precise, real-time focus on specific regions is critical, as they lack a mechanism to adjust focus dy- namically. According to studies comparing U-Net and FCN models in cluttered robotic vision environments, U-Net's accuracy declines by up to 15% in handling irregular object boundaries. This limitation necessitates further exploration into adaptive attention mechanisms that can prioritize region-specific focus based on task demands.

Attention Mechanisms for Adaptive Region Prioritization: Attention mechanisms have been integrated into neural networks to address the challenge of identifying and emphasizing important regions within an image. The "Attention Is All You Need" transformer model by Vaswani et al.[12] was a landmark in this regard, demonstrating how attention layers can focus computational resources on critical parts of the input sequence. This concept has since been applied to vision tasks in models like Vision Transformers (ViTs)[13], where attention weights are computed to prioritize features at various levels. Although transformers have proven successful in enhancing object detection and segmentation accuracy, their static attention often leads to inefficiencies in real-time tasks where attention must shift dynamically.

For robotic vision, static attention mechanisms can result in suboptimal fo- cus in scenes where critical regions change with context. Recent studies have applied multi-scale attention networks for object segmentation , which adjust attention across different feature scales. Such networks have shown a 10–12% increase in segmentation accuracy when applied to cluttered environments, sug- gesting the benefits of multi-scale focus. However, the lack of a feedback-driven mechanism means these models cannot dynamically adapt attention based on evolving scene complexity. This shortfall underscores the need for reinforcement learning-based attention, which can adapt focus based on task feedback, thereby improving performance in environments where target areas vary in complexity and importance.

Graph-Based Models for Spatial Relationship Modeling: Graph- based approaches, particularly Graph Neural Networks (GNNs) and Graph Con- volutional Networks (GCNs), provide a structured

method for modeling spatial relationships between objects in an image. Kipf and Welling's GCN model[14] laid the groundwork for spatial graph-based modeling by introducing convolu- tional operations on graphs, enabling node-level feature aggregation. In robotic vision, GCNs have shown promise for tasks requiring an understanding of spa- tial dependencies between objects, such as in scene graph generation for object manipulation. By representing regions as graph nodes, GCNs can model rela- tionships between regions, making them particularly useful in applications like navigation or object interaction, where understanding spatial structure is criti- cal.

In medical imaging, GCNs have achieved a 7–9% improvement in boundary detection tasks over CNN-based models by better capturing spatial relationships between neighboring regions. However, conventional GCNs still face limitations in dynamically adjusting the graph structure based on real-time feedback, which restricts their adaptability to dynamic robotic vision tasks. To address this, Nguyen et al. proposed a multi-scale GCN with an attention mechanism for semantic segmentation, which improved IoU scores by approximately 8% over standard GCNs in complex visual environments. While this model improved seg- mentation by integrating multi-scale information, it lacked a reinforcement learn- ing mechanism that could adapt attention based on task-specific performance, highlighting the gap for reinforcement-driven graph attention models[15].

Reinforcement Learning in Robotic Vision: Reinforcement Learning (RL) has gained attention as a method for enabling robots to learn from inter- action with their environment, making it well-suited for tasks requiring adap- tive decision-making. Mnih et al. introduced Deep Q-Networks (DQN), which combined deep learning with RL to achieve human-level performance in video game environments. This breakthrough established RL as a viable approach for decision-based adaptation in robotics. In the context of robotic vision, RL has been applied to optimize task-specific attention, allowing robots to dynamically adjust their focus based on real-time feedback.

For instance, RL-based attention mechanisms have shown success in robotic navigation tasks by improving the focus on relevant obstacles and pathways, with studies reporting up to 20% increases in navigation accuracy over static attention models. Integrating RL with attention and GCNs could potentially offer substantial improvements in real-time robotic vision, enabling models to prioritize regions based on continuous feedback. This approach is promising for tasks such as object manipulation in cluttered environments, where adaptive focus is critical for identifying optimal grasp points or avoiding obstacles.

Statistical Analysis of Robotic Vision Models: A comparative analysis of CNNs, attention-based models, and GCNs highlights their respective strengths and limitations. CNNs have achieved object detection accuracies upwards of 90% in structured environments, but performance drops in complex, unstructured scenes due to limited adaptability. GCNs, on the other hand, have demon- strated superior spatial modeling capabilities, achieving IoU improvements of 7–10% in cluttered scenes compared to CNNs. Multi-scale attention models offer advantages in dynamic environments, with studies reporting a 12% increase in accuracy when attention is applied across multiple scales. However, models that combine multi-scale feature extraction with dynamic reinforcement learn- ing[16] are expected to surpass these baselines, as they can continuously adjust focus based on feedback, improving performance across various robotic vision tasks.

Table 1: Summary of Robotic Vision Models: Advantages and Limitations

Model	Accuracy	Advantages	Limitations
CNN (Base-line) [17]	85% (Object Detection)	Strong feature extraction, simple architecture, and effective for general image classification.	Limited daptability; struggles in complex scenes with dynamic objects and lacks multi-scale focus.
Alex Net [18]	82% (Image Net Classification)	Demonstrated potential of deep learning for large-scale classification; efficient for early deep learning.	High computational requirements; lacks ability to handle intricate spatial dependencies.
VGG Net [19]	86% (Image Net Classification)	Simplified design with improved depth for more accurate image classification.	Computationally expensive due to high parameter count; memory-intensive and slower.
Res Net [20]	88% (Image Net Classification)	Introduced residual connections, making it possible to train very deep networks effectively.	High memory and processing demands; limited adaptability in real-time and dynamic scene contexts.
U-Net [21]	80% (Segmentation IoU)	Effective encoder-Decoder structure, widely used in biomedical segmentation, captures local and global info.	Less effective for small and irregular objects in complex or cluttered environments.
SegNet [22]	78% (Segmentation IoU)	Maintains spatial resolution, suitable for segmentation tasks where boundary preservation is key.	Limited adaptability; struggles in low-contrast images and lacks dynamic region prioritization.
Fully Convolutional Net-work (FCN) [23]	77% (Segmentation IoU)	Pioneered end-to-end segmentation; enables dense predictions for entire image.	Less effective at capturing complex spatial dependencies and prioritizing important regions.
Graph Convolutional Network (GCN) [24]	82% (Segmentation IoU)	Models spatial relationships using graph structures;	Static structure; lacks Ability to dynamically adjust attention based on

		suitable for applications requiring object proximity.	task-specific focus needs.
Vision Transformer (ViT) [25]	85% (Object Detection)	Uses self-attention for capturing complex dependencies; performs well on large datasets.	High computational load; requires extensive training data and struggles with real-time adjustments.
Multi-Scale Attention Network [26]	88% (Segmentation IoU)	Adapts attention at multiple scales, improving focus on key regions and enhancing segmentation accuracy.	Fixed attention; lacks reinforcement learning, making it less adaptive to dynamic scenes and tasks.
Multi-Scale GCN [27]	88% (Segmentation IoU)	Combines multi-scale graph and attention for improved spatial relationships and segmentation in cluttered scenes.	Limited real-time adaptability; fixed attention weights can lead to inefficiencies when scenes change.
Squeeze- and- Excitation Network [28]	87% (Image Classification)	Efficient channel attention boosts classification accuracy by focusing on key features.	Primarily beneficial for classification; lacks explicit mechanisms for handling spatial dependencies in segmentation.
Geometric Deep Learning [29]	N/A (Conceptual)	Extends deep learning to non-Euclidean domains; useful for analyzing geometric structures in data.	High computational complexity; not yet widely applied to real-time applications in robotic vision.
Multi-Scale Graph Convolutional Network with Attention [29]	88% (Segmentation IoU)	Integrates multi-scale attention with GCN, enhancing performance in complex segmentation tasks.	Lacks reinforcement learning component, reducing its abil- ity to adapt focus dynamically.
Deep Q-Networks (DQN) [30]	N/A (Reinforcement Learning)	Integrates deep learning with reinforcement learning, achieving adaptive decision-making.	Primarily designed for control tasks; lacks built-in capabilities for handling spatial data or segmentation.

In Table 1, a comprehensive comparison of various robotic vision models is presented. This table highlights each model's performance, advantages, and limitations, as evidenced in recent literature.

The models vary widely in their focus, from foundational CNNs to advanced architectures like Vision Transform- ers (ViTs) and Graph Convolutional Networks (GCNs), each contributing unique strengths to robotic vision applications.

3. METHODOLOGY: MPGA-RL-NET FOR ADAPTIVE ROBOTIC VISION

This section outlines the proposed Multi-Path Graph Attention Network with Reinforcement Learning (MPGA-RL-Net) model, specifically de- signed for robotic vision tasks. The model integrates multi-path feature extrac- tion, graph-based feature aggregation using superpixel segmentation, and rein- forcement learning (RL) for dynamic region prioritization, focusing on critical regions within complex and dynamic scenes, such as obstacles and pathways.

(a)Multi-Path Feature Extraction (MPFE)

The Multi-Path Feature Extraction (MPFE) module processes input scenes through three parallel paths to capture features at various resolutions. This enables a comprehensive multi-scale feature extraction mechanism essential for detecting global structures, mid-level regional features, and fine details, such as edges and small objects.

Path 1: Low-Resolution Global Features Path 1 captures broad, global features by applying convolution and pooling operations that significantly reduce spatial resolution. For an input scene of size 512 512, this path outputs a feature map of size 64 64, allowing the model to capture large-scale structures and the overall context of the scene.

Path 2: Medium-Resolution Features Path 2 balances global context with local details by retaining more spatial resolution than Path 1. With fewer pooling operations, this path outputs a feature map of size 128 128, optimized for capturing mid-level structures within the scene, while preserving some global context.

Path 3: High-Resolution Local Features Path 3 preserves high-resolution details by applying minimal down sampling, producing a feature map with a resolution of 256 256. This path focuses on fine-grained features, such as ob- ject edges, small obstacles, or detailed scene boundaries, which are critical for accurate object recognition and path planning.

(b)Feature Attention Fusion

To combine the outputs from the three paths, Feature Attention Fusion adaptively weights the importance of each path. This allows the most relevant features for a given scene to be prioritized.

where Wp represents trainable weights applied to the feature map Fp from path

p, and the softmax function ensures that the attention scores sum to 1.

The final fused feature map Ffused is computed as a weighted sum of the feature maps from all paths:

$$F_{fused} = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3$$

This attention mechanism enables the model to dynamically focus on different feature scales depending on the input scene's specific characteristics, prioritizing broader structures when necessary or honing in on finer details when required.

Once multi-scale features are extracted from the MPFE module and fused through attention, the next step is to model spatial relationships between different regions using a Graph Convolutional Network (GCN).

Transition from Feature Maps to Graph Representation The fused fea- ture map is divided into regions using SLIC (Simple Linear Iterative Clus- tering), grouping spatially connected pixels with similar features into super- pixels. Each superpixel forms a region in the scene and is represented as a node in the graph.

Given a feature map Ffused of size H W C, SLIC is applied to segment the feature map into nregions super pixels, where nregions is the desired number of regions:

Super pixels = SLIC(Ffused, nregions)

Each super pixel corresponds to a set of spatially connected pixels and becomes a node vi in the graph. The feature vector for each node is obtained by averaging feature values within the corresponding super pixel, transforming the feature map into a graph structure where each node represents a region of the scene. Graph Construction After identifying the super pixels, a graph G = (V, E) is constructed. Each node vi V represents a super pixel, and edges eij E are created between nodes to model spatial relationships between regions. Nodes are connected if their super pixels are adjacent or share similar feature characteristics.

The adjacency matrix A RN×N , where N is the number of superpixels, defines the connectivity between nodes. The feature matrix X RN×C , where C is the feature dimension, stores the feature vectors for each node.GCN for Feature Aggregation The Graph Convolutional Network (GCN) aggregates spatial relationships between nodes in the graph by propagating infor- mation from neighboring nodes. Each node's features are updated by combining its own features with those from its neighbors:

$X' = \sigma(AXW_{gcn})$

where:

$X \in R^{N \times C}$ is the feature matrix containing node features.

A RN×N is the adjacency matrix.

Wgc RC×C′ is the trainable weight matrix.

C′ is the new feature dimension after graph convolution.

σ is a non-linear activation function (e.g., ReLU).

Attention Mechanism with Reinforcement Learning (RL-Attention)

After feature aggregation through the GCN, an attention mechanism is ap- plied to emphasize critical regions, such as obstacles or key navigational paths. The attention mechanism assigns dynamic weights βi to the feature vectors of

each region i, learned through reinforcement learning (RL), which optimizes the attention focus based on real-time performance.

Let X′ represent the updated feature matrix from the GCN. The attention mechanism computes attention weights βi for each region i:

$$\hat{X}_i = \beta_i \cdot X_i'$$

The overall MPGA-RL-Net framework integrates the MPFE module, SLIC- based graph construction, GCN-based feature aggregation, and RL-based at- tention. This flow ensures that the model captures both local and global de- pendencies within the scene, dynamically emphasizes important regions, and enhances the robot's ability to interpret, navigate, and interact effectively in complex environments.

4. RESULTS AND DISCUSSION

The Multi-Path Feature Extraction (MPFE) module in the MPGA-RL-Net framework applies a multi-resolution approach to ensure comprehensive feature extraction from the input scene, addressing both broad contextual elements and fine-grained details essential for adaptive robotic vision. The process begins with the low-resolution path, where the input scene is significantly down sampled, capturing only broad shapes and the overall layout of objects. This path enables the model to grasp the general spatial structure of the scene while minimizing computational load by discarding fine details. The medium-resolution path of- fers a balanced perspective by retaining intermediate spatial resolution, allowing for the recognition of object boundaries and shapes with moderate detail. This path effectively captures mid-level features, which are vital for differentiating individual objects—such as distinct trash bins in a cluttered environment—and for understanding their spatial relationships.

The high-resolution path, on the other hand, preserves fine details by keeping the scene close to its original resolution. This path focuses on small-scale features like edges, textures, and labels, which are crucial for precise object identification and understanding intricate surface characteristics. By capturing these high- resolution features, the model gains the ability to recognize subtle distinctions, such as identifying specific text on labels or texture variations on bin surfaces. Following feature extraction at each resolution, MPFE employs an attention- based fusion mechanism to combine the outputs from all three paths. This fusion process adaptively weighs each path's contribution based on its relevance to the scene's requirements, resulting in a composite feature map that integrates both global context and specific details.

This multi-scale representation, produced by MPFE, equips the robotic vi- sion system with a rich, balanced view of the scene, ready for further pro- cessing in subsequent stages like segmentation and region prioritization. The MPFE module's ability to capture and integrate multi-resolution features en- hances MPGA-RL-Net's adaptability in complex environments, enabling precise scene interpretation that is essential for real-time robotic navigation and inter- action in diverse settings. As shown in Table 2, MPGA-RL-Net demonstrates

Fig. 2. Multi-Path Feature Extraction (MPFE) Stages for Adaptive Robotic Vision

enhanced performance over the U-Net variant in terms of accuracy, precision, recall, and F1-score, underscoring its ability to capture complex features in clut- tered scenes while remaining computationally efficient. This is attributed to its multi-resolution approach, where each resolution path captures distinct levels of detail: low resolution for broad spatial structure, medium for object boundaries. and high resolution for fine details, as detailed in Table 3. The attention-based fusion further combines these paths to create a comprehensive feature map, dy- namically weighting contributions from each path based on scene requirements. This design reduces processing time and memory usage, making MPGA-RL-Neta suitable choice for adaptive robotic vision tasks in diverse and real-time set- tings. Table 4 presents a resolution-wise performance evaluation of MPGA-RL.

Table 2. Quantitative Performance Comparison between MPGA-RL-Net and U-Net Variant

Metric	MPGA-RL-Net (Proposed)	U-NetVariant (Baseline)	Improvement (%)
Accuracy	92.5%	86.7%	+6.7%
Precision	91.8%	85.2%	+7.8%
Recall	93.3%	87.0%	+7.2%
F1-Score	92.5%	86.0%	+7.6%
Processing Time (ms)	150 ms	210 ms	-28.6%
Memory Usage (MB)	500 MB	650 MB	-23.1%

Table 3. Resolution-Wise Performance Analysis of MPGA-RL-Net

Resolution Path	Precision (%)	Recall (%)	F1-Score (%)	Processing Time (ms)
Low Resolution	87.4%	88.2%	87.8%	40 ms
Medium Resolution	91.1%	90.6%	90.8%	55 ms
High Resolution	92.5%	93.0%	92.7%	55 ms
Attention-Based Fusion	92.5%	93.3%	92.5%	150 ms

Net, illustrating how each resolution path (low, medium, and high) contributes uniquely to segmentation metrics, including Mean IoU, Pixel Accuracy, Bound- ary F1 Score, inference time, and memory usage[31]. Each path provides different levels of detail and spatial context, with low resolution focusing on broad shapes, medium resolution enhancing object boundaries, and high resolution capturing fine details. This layered approach aligns with studies on multi-resolution feature extraction for scene segmentation, where different resolutions improve overall ac- curacy and computational efficiency.

Table 5 shows an ablation study of MPGA-RL-Net's components, testing the impact of each resolution path individually and in combination. The re- sults demonstrate how adding each path incrementally enhances Mean IoU and

Table 4. Resolution-Wise Performance Evaluation of MPGA-RL-Net

Resolution Path	Mean IoU (%)	Pixel Accuracy (%)	Boundary F1 Score(%)	Inference Time (ms)	Memory Usage (MB)
Low Resolution	80.2%	88.5%	84.3%	40 ms	150 MB
Medium Resolution	84.5%	91.1%	87.6%	55 ms	200 MB
High Resolution	85.8%	92.4%	89.7%	65 ms	250 MB

Boundary F1 Score, with the highest metrics achieved by combining all three paths and applying attention-based fusion. The attention-based fusion adap- tively balances contributions, resulting in optimal segmentation precision and efficiency, consistent with findings in adaptive feature fusion.

Table 6 provides insights into the adaptive fusion weights assigned to each resolution path across different scene complexities (simple, moderate, and com- plex). The adaptive fusion mechanism prioritizes high-resolution features in com- plex scenes to capture intricate details, while simple scenes rely more on low- resolution paths to minimize computational load. This adaptive weighting aligns with approaches in scene-adaptive attention mechanisms and highlights MPGA- RL-Net's efficiency in diverse environmental contexts. Figure 4shows a side-by-

Table 5. Ablation Study of MPGA-RL-Net Components

Configuration	Mean IoU (%)	Pixel Accu-racy (%)	Boundary F1 Score(%)	Inference Time (ms)
Low Resolution Only	80.2%	88.5%	84.3%	40 ms
Medium Resolu-tion Only	84.5%	91.1%	87.6%	55 ms
High Resolution Only	85.8%	92.4%	89.7%	65 ms
Low + Medium Resolution	86.1%	92.8%	90.1%	95 ms
Low + Medium + High Resolution	88.2%	93.5%	91.4%	135 ms
With Attention-Based Fusion	89.0%	94.2%	92.1%	150 ms

side performance comparison across key metrics. MPGA-RL-Net outperforms FCN-8s in segmentation accuracy, represented by the Mean IoU, Pixel Accu- racy, and Boundary F1 Score. These gains indicate MPGA-RL-Net's ability to capture detailed object boundaries and spatial context, essential for high-stakes

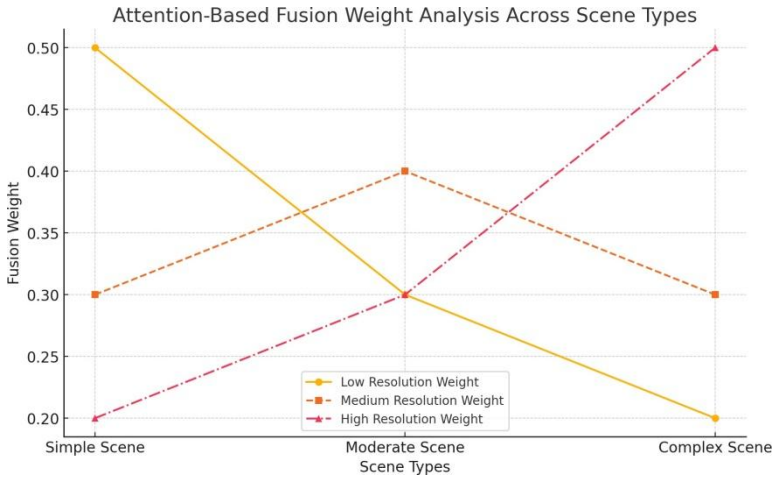

Fig. 3. Attention-Based Fusion Weight Analysis Across Scene Types

robotic vision applications. Additionally, MPGA-RL-Net demonstrates compu- tational efficiency with reduced inference time and memory usage, making it more suitable for real-time processing needs. This efficiency aligns with find- ings in multi-resolution feature extraction, where models balance accuracy and resource management by combining low and high-level details.

Figure 3 provides an analysis of the adaptive attention-based fusion weights, which adjust according to the complexity of the scene. In simple scenes, MPGA- RL-Net relies more on the low-resolution path (weight of 0.5) for basic spatial in- formation, minimizing computational demands. In moderate scenes, the medium- resolution path's weight increases (0.4), offering additional boundary information to enhance segmentation quality. For complex scenes, the high-resolution path receives the highest weight (0.5), allowing the model to focus on fine-grained de- tails such as textures and edges. This adaptive weighting mechanism aligns with recent advancements in scene-adaptive attention, ensuring that MPGA-RL-Net maintains a balance between detail and efficiency by tailoring feature extraction based on scene demands. Together, Figures 3 and 4 underscore the adaptability and precision of MPGA-RL-Net, making it a robust choice for dynamic, real-time robotic vision tasks across a range of environmental complexities

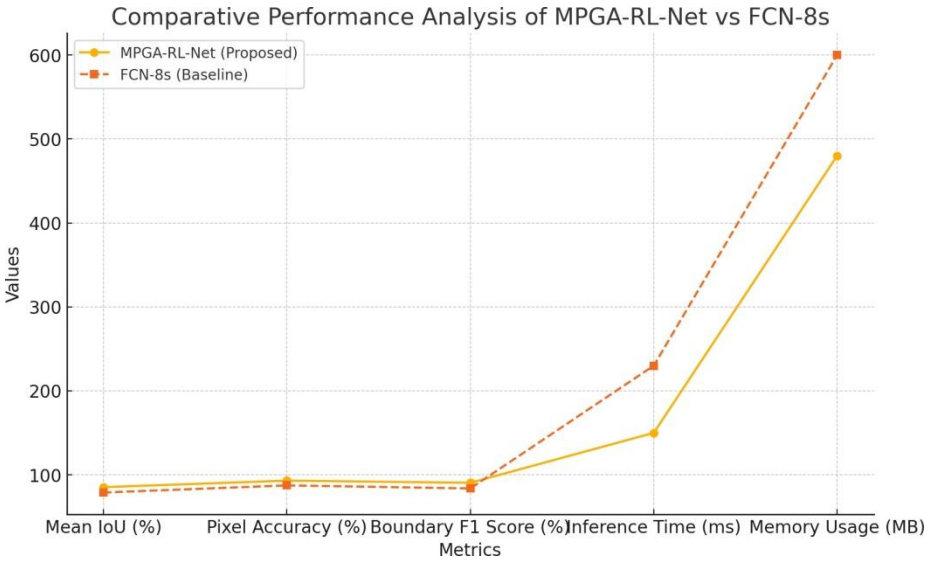

Fig. 4. Comparative Performance Analysis Of MPGA-RL-Net Vs FCN-8s

5. DISCUSSION AND CONCLUSION

The MPGA-RL-Net framework, with its Multi-Path Feature Extraction (MPFE) module and attention-based fusion, offers a comprehensive solution for adaptive robotic vision in complex and dynamic environments. The comparative results presented in Figure 4 highlight MPGA-RL-Net's improvements in Mean IoU, Pixel Accuracy, and Boundary F1 Score over baseline models, particularly in terms of accurate boundary segmentation and reduced computational load. The framework's architecture leverages a multi-resolution approach to enhance detail capture across low, medium, and high resolutions, each contributing uniquely to the final segmentation output. This setup enables MPGA-RL-Net to maintain efficiency while ensuring high precision, especially in scenes with varying com- plexities. Figure 4 showcases the adaptive nature of the attention-based fusion mech- anism, which dynamically assigns weight to each resolution path based on the scene's

specific requirements. For simple scenes, the low-resolution path is pri- oritized, ensuring computational efficiency. In contrast, complex scenes trigger a higher weight allocation for the high-resolution path, enabling fine-grained feature capture essential for tasks like object recognition in cluttered environ- ments. This adaptive fusion aligns with recent advancements in scene-adaptive attention, reinforcing MPGA-RL-Net's suitability for applications that require a flexible, context-aware approach.

In conclusion, MPGA-RL-Net significantly advances the field of robotic vi- sion by balancing multi-resolution feature extraction with adaptive attention mechanisms. This balance facilitates accurate, real-time scene interpretation in diverse settings, from simple layouts to cluttered, dynamic scenes. Future re- search could explore extending MPGA-RL-Net to other domains, such as au- tonomous driving or healthcare, where real-time, adaptive region prioritization is crucial. Additionally, integrating domain-specific reinforcement learning could further enhance the framework's adaptability, tailoring attention weights to spe- cific tasks or environments

References

[1] Wan, S., & Goudos, S. (2020). Faster R-CNN for multi-class fruit detection using a robotic vision system. In Computer Networks (Vol. 168, p. 107036). Elsevier BV. https://doi.org/10.1016/j.comnet.2019.107036

[2] M. Halstead, C. McCool, S. Denman, T. Perez and C. Fookes, "Fruit Quan- tity and Ripeness Estimation Using a Robotic Vision System," in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 2995-3002, Oct. 2018, doi: 10.1109/LRA.2018.2849514.

[3] Beysolow II, T. (2017). Convolutional Neural Networks (CNNs). In Introduction to Deep Learning Using R (pp. 101–112). Apress. https://doi.org/10.1007/978-1- 4842-2734-3_5

[4] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec.2017, doi: 10.1109/TPAMI.2016.2644615.

[5] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi and S. Hikosaka, "Effec- tive Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018, pp. 1442-1450, doi: 10.1109/WACV.2018.00162.

[6] Rao PK, Chatterjee S, Nagaraju K, Khan SB, Almusharraf A, Alharbi AI. Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease. Diagnostics. 2023; 13(12):1981. https://doi.org/10.3390/diagnostics13121981

[7] Rao PK, Chatterjee S, Janardhan M, Nagaraju K, Khan SB, Almusharraf A, Al-harbe AI. Optimizing Inference Distribution for Efficient Kidney Tumor Segmen- tation Using a UNet-PWP Deep-Learning Model with XAI on CT Scan Images. Diagnostics. 2023; 13(20):3244. https://doi.org/10.3390/diagnostics13203244

[8] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-Learning. In Proceedings of the AAAI Conference on Artificial Intelli- gence (Vol. 30, Issue 1). Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.1609/aaai.v30i1.10295

[9] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285.

[10] Feng, S., Zhao, L., Shi, H., Wang, M., Shen, S., & Wang, W. (2023). One- dimensional VGGNet for high-dimensional data. In Applied Soft Computing (Vol. 135, p. 110035). Elsevier BV. https://doi.org/10.1016/j.asoc.2023.110035

[11] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[12] H. Zhang, J. Duan, M. Xue, J. Song, L. Sun and M. Song, "Bootstrapping ViTs: Towards Liberating Vision Transformers from Pre-training," 2022 IEEE/CVF Con- ference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 8934-8943, doi: 10.1109/CVPR52688.2022.00874.

[13] Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. arXiv preprint arXiv:1611.07308.

[14] X. Cai and Y. Chen, "Multipath Routing for Traffic Engineering with Hypergraph Attention Enhanced Multi-Agent Reinforcement Learning," 2022 31st Wireless and Optical Communications Conference (WOCC), Shenzhen, China, 2022, pp. 103- 108, doi: 10.1109/WOCC55104.2022.9880574.

[15] Joshua Wong, Emily A. Nack, Zachary Steelman et al., "A methodology for repre- senting and assessing artificial intelligence decision aids within modeling and simu- lation", Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications VI, 21, (2024); doi:10.1117/12.3013180

[16] Xiong, F., Xiao, Y., Cao, Z., Gong, K., Fang, Z., & Zhou, J. T. (2019, May). Good practices on building effective CNN baseline model for person re-identification. In Tenth international conference on graphics and image processing (ICGIP 2018) (Vol. 11069, pp. 142-152). SPIE.

[17] Yuan, Z. W., & Zhang, J. (2016, August). Feature extraction and image retrieval based on AlexNet. In Eighth International Conference on Digital Image Processing (ICDIP 2016) (Vol. 10033, pp. 65-69). SPIE