International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhanced Superpixel-Based Multiscale CNN Mechanism for UAV Image Segmentation

Satish Dekka[1], T.Tejaswi[2],T.Radha Krishna Varma[3], P.Sai Kiran[4],V.Upendra[5]

[1,2,3,4,5] Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology (Autonomous), Andhra Pradesh, India.

satishmsc4u@gmail.com[1], sujitammina3@gmail.com[2], trkv04@gmail.com[3], pothalasaikiran2@gmail.com[4], varanasiupendra122@gmail.com[5].

Abstract – Unmanned Aerial Vehicles (UAVs) are advanced remote sensing tools that have the potential to revolutionize variety applications,including environmental monitoring, urban planning, agriculture, and disaster management. These airborne sensors provide high resolution, real-time data used for traditional remote sensing methods often struggle to capture, particularly in inaccessible or large-scale areas. To address challenges in misclassification in complex urban aerial imagery, we proposed a super-pixel-aided multiscale Convolutional Neural Network (CNN) architecture.

This approach integrates an attention mechanism, and the SLICO algorithm. The attention mechanism enhances the model's focus on crucial image regions, optimizing feature extraction. The SLICO algorithm generates super-pixels to reduce computational costs and refine boundary detection. This integrated approach effectively addresses scale variance in aerial imagery, resulting in more precise segmentation. We evaluated the model using UAV-based dataset: the Urban Drone Dataset (UDD). The proposed model significantly outperformed several state-of-the-art methods, achieving impressive Intersection over Union (IoU) scores on dataset.In recent years, unmanned aerial vehicles (UAVs) have gained significant attention across a wide range of domains, including urban planning, precision agriculture, disaster response, environmental monitoring, and infrastructure surveillance. Their ability to capture high- resolution images at low operational costs, coupled with flexible deployment and maneuverability, makes them a powerful tool in the field of remote sensing. The wealth of visual data collected by UAVs provides valuable insights for analysis, yet this same abundance introduces unique computational and methodological challenges.

Keywords : Multiscale Convolutional Neural Network (CNN), SLICO Algorithm, Unmanned Aerial Vehicles (UAVs), Attention Mechanism, Urban Drone Dataset (UDD).

1.INTRODUCTION

Unmanned Aerial Vehicles (UAVs) have emerged as vital tools in remote sensing and aerial surveillance, offering high- resolution imagery for a wide range of applications including urban planning, environmental monitoring, disaster management, and precision agriculture. A fundamental step in extracting actionable information from UAV imagery is semantic segmentation, the process of assigning a class label to each pixel in an image. This fine-grained, pixel-level classification is crucial for detailed scene understanding,enabling automated recognition and localization,various land cover types, structures, or objects.

UAV- captured imagery presents unique challenges that distinguish it from traditional ground-based or satellite image segmentation tasks. UAV imagery is characterized by a high degree of scene complexity, often involving

overlapping objects (e.g., buildings, vehicles, vegetation), scale variation due to dynamic flight altitudes, occlusions, and inconsistent Despite its significance, semantic segmentation of illumination stemming from environmental factors and camera angles. These conditions make traditional image segmentation techniques—such as thresholding, region growing, and edge detection—largely inadequate. These conventional approaches typically rely on low-level image features and predefined heuristics, making them highly sensitive to noise, scale, and contextual variation.In recent years**,** deep learning has revolutionized semantic segmentation through the use of Convolutional Neural Networks (CNNs), which learn hierarchical feature representations directly from the data. CNNs have demonstrated remarkable performance on benchmark datasets, owing to their ability to capture both spatial and semantic information across multiple layers.

However, when applied to UAV imagery, standard CNNs face several limitations. Most notably, they often fail to accurately delineate fine object boundaries and detect small-scale features, particularly in high-resolution images where objects may occupy only a few pixels. Moreover, fixed receptive fields and downsampling operations such as pooling can cause loss of spatial detail, further impairing performance on highly variable aerial scenes. To address these shortcomings, there has been growing interest in the integration of super-pixel algorithms into semantic segmentation pipelines. Super-pixels aggregate pixels with similar properties into perceptually meaningful regions, offering a more compact and structured image representation. By reducing the number of primitives from pixels to super-pixels, these algorithms significantly lower computational complexity and enhance the spatial coherence of segmentation results. In particular, the Simple Linear Iterative Clustering (SLIC) algorithm and its improved variant, SLICO, have gained popularity due to their efficiency and adaptability to image content. SLICO, for example, eliminates the need for manual tuning of compactness parameters and produces more natural boundary adherence.

2. RELATED WORK

Qi Diao et al.(1) proposed SAGNN, **a superpixel-based attention graph neural network for semantic segmentation** of high-resolution aerial images. By combining CNN features, superpixel graphs, and attention mechanisms, SAGNN enhances boundary precision and robustness, outperforming state-of-the-art methods on the Potsdam and Vaihingen datasets.

Ching-Lung Fan(2) proposed MSFCNN, a multiscale feature convolutional neural network for extracting buildings and roads from **UAV and satellite imagery**. By capturing features at multiple scales, MSFCNN achieved over 91% accuracy on Kaohsiung images, showing strong adaptability across resolutions for urban land cover mapping.

Shuang Tian et al.(3) proposed a multiscale superpixel-based method for **fine crop classification in UAV-based hyperspectral imagery.** By integrating multiscale information through pre- and post-processing strategies, the approach enhances spectral-spatial representation, with post-processing yielding the highest accuracy across three public datasets.

Liang Huang et al.(4) proposed a UAV image segmentation method that combines **SLIC superpixels with multi-feature distance measures** (spectral, texture, shape, area) to reduce over-segmentation. Tested on two UAV datasets, it outperforms the FNEA algorithm, particularly in segmenting objects of varying sizes.

Zhiyou Lian and Jianhua Ren(5) proposed a **UAV image stitching method using superpixel segmentation** to improve speed and stitching quality. By integrating superpixel-based region estimation, enhanced SIFT

matching, and optimized fusion, the method doubles feature extraction speed and improves SSIM, PSNR, and MAE by ~5% over AANAP, VSP, and UVS methods.

Yunjie Mu et al.(6) proposed SCGCN, a superpixel-based graph convolutional network that segments **UAV forest fire images** by converting them into graphs and applying node classification. Using CNNs for feature extraction and GraphSAGE for graph learning, SCGCN outperformed four mainstream models on FLAME and Chongli datasets in terms of F1 score and accuracy.

S. Crommelinck et al.(7) evaluated SLIC superpixels for delineating roads and roofs in **high-resolution UAV orthoimages for cadastral mapping**. With up to 64% completeness on 0.05 m GSD images, SLIC proves effective for segmentation but needs integration with other methods for precise boundaries, serving well as a preprocessing step to enhance mapping efficiency.

Liang Huang et al.(8) proposed **a UAV image segmentation method** that applies SLIC superpixels followed by feature-based merging (spectral, texture, shape, area) to reduce over-segmentation. Tested on two UAV datasets, it outperforms the FNEA algorithm, particularly in handling varying object sizes.

Huang et al.(9) introduced batch loss regularization in deep learning to improve **aerial scene classification** by reducing overfitting on limited datasets. This technique enhances model robustness and generalization, boosting the reliability of deep learning for land-use analysis in aerial remote sensing.

Lowe(10) introduced the Scale-Invariant Feature Transform (SIFT), a method for **object recognition using local scale-invariant features.** SIFT enables reliable matching across scales, orientations, and lighting, revolutionizing computer vision and underpinning applications like robotics, image retrieval, and 3D reconstruction.

Teng et al(14) developed a method combining satellite imagery and visible-near infrared (VNIR) spectroscopy to map and model soil loss across Australia. Their study shows how remote sensing can be used effectively for **monitoring environmental degradation caused by water erosion.** This technology enables high-resolution mapping crucial for soil conservation planning. The integration of imagery and spectroscopy provides a more detailed and accurate analysis compared to traditional methods. Their findings contribute significantly to sustainable land management strategies.

Chen et al.(15) proposed a **land-use scene classification method using multi-scale completed local binary patterns** (CLBP) to capture texture information across scales. This approach improves classification accuracy for diverse land covers and offers a robust tool for applications like urban planning, agriculture, and environmental monitoring.

3.ARCHITECTURE

The initial step involves preparing the UAV imagery for both super-pixel generation and CNN input. The dataset contains aerial images captured at varying altitudes and under different environmental conditions, with complex object arrangements such as roads, vegetation, buildings, and vehicles. These images are high-resolution, making them ideal for detailed semantic analysis but also challenging due to memory and computationalconstraint.

To standardize the input data:

All images are resized or cropped to 576 × 576 using a sliding window technique with a constant stride.This technique ensures that spatial context is preserved and

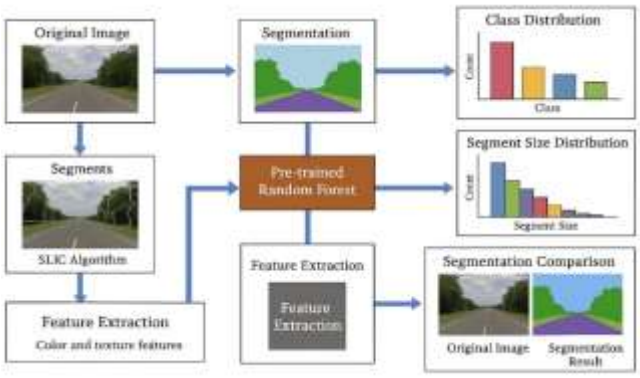

Fig-1:System Architecture for Image Segmentation

the training samples cover diverse regions of the scene.The image patches are normalized and augmented using techniques such as horizontal flipping, rotation, and scaling to improve model generalization.It includes step by step process for the implementation of Proposed model.The System architecture includes several number of steps.

3.1 Preprocessingand Dataset Preparation

The initial step involves preparing the UAV imagery for both super-pixel generation and CNN input. The dataset contains aerial images captured at varying altitudes and under different environmental conditions, with complex object arrangements such as roads, vegetation, buildings, and vehicles. These images are high-resolution, making them ideal for detailed semantic analysis but also challenging due to memory and computational constraints.

3.2 Super-pixel Generation using SLICO

The super-pixel segmentation process is performed using SLICO (Simple Linear Iterative Clustering Zero parameter), an advanced variant of the original SLIC algorithm. Unlike its predecessor, SLICO eliminates the need for manually tuning the compactness parameter, which makes it ideal for real- world UAV images that may vary widely in content and structure.SLICO groups pixels into clusters (super-pixels) based on spatial proximity and color similarity in the CIE Lab color space. Each super-pixel tends to represent a meaningful region of the image (such as part of a road, building, or tree canopy) while significantly reducing the number of input units for the CNN.Key advantages of using super-pixels in this stage include: Reduced computational complexity for downstream processing,Improved edge preservation, which enhances segmentation near object boundaries.Enhanced local structure representation, making CNN feature extraction more effective. The output of this step is a set of super-pixel-enhanced images, where the pixel values inside each segment are uniform, effectively smoothing texture without destroying important shape details.

3.3 CNN-Based Segmentation Model

The core segmentation engine is a custom CNN model based on an encoder–decoder architecture, similar in spirit to U-Net but designed to accommodate multiscale input. The model is composed of the following layers:Convolutional Layers: Extract feature maps using filters of size 3×3, followed by ReLU activation functions.

1. Batch Normalization: Applied after each convolution to stabilize and accelerate training.
2. Max Pooling Layers: Downsample the feature maps to retain dominant information while reducing spatial size.
3. Dropout Layers: Incorporated during training to reduce overfitting.
4. Upsampling (Transposed Convolution): Reconstruct the spatial dimensions of the feature maps in the decoder path.
5. Skip Connections: Enable information flow between corresponding encoder and decoder layers to preserve spatial details.
6. Softmax Layer: Produces the final output map, where each pixel is classified into one of the semantic classes.

3.4 Multiscale Feature Aggregation

To handle scale variance a common issue in UAV imagery due to differing object sizes and altitudes — the proposed system incorporates multiscale learning. The same image is resized into three versions:

256 × 256 (downsampled)

512 × 512 (original resolution)

1024 × 1024 (upscaled)

Each of these scaled versions is passed through an identical CNN model. This enables the network to learn fine-grained features from high-resolution inputs and contextual features from downsampled inputs simultaneously. The resulting feature maps from each CNN are:

Resized to a common dimension.Concatenated channel-wise Passed through a final 1×1 convolutional layer to produce the final aggregated prediction map.This multiscale fusion step is crucial for capturing both local textures and global structures enhancing the model's robustness to scale differences and improving segmentation performance.

4. RESULTS AND DISCUSSION

This section presents the experimental results and evaluates the performance of the proposed superpixel- enhanced multiscale CNN architecture. The evaluation is conducted using standard metrics, both quantitative and qualitative, on UAV datasets featuring complex scenes with multiple object classes.

4.1 Experimental Setup

The implementation was carried out on Google Colab, leveraging GPU acceleration to support the computational

demands of deep learning. The models were implemented using TensorFlow and OpenCV libraries.Optimizer: Adam optimizer with a learning rate of 0.001 Loss Function: Categorical Cross-Entropy,Batch Size: 3

Epochs: 450 (with early stopping based on validation loss) Data Augmentation: Applied to double the training data.

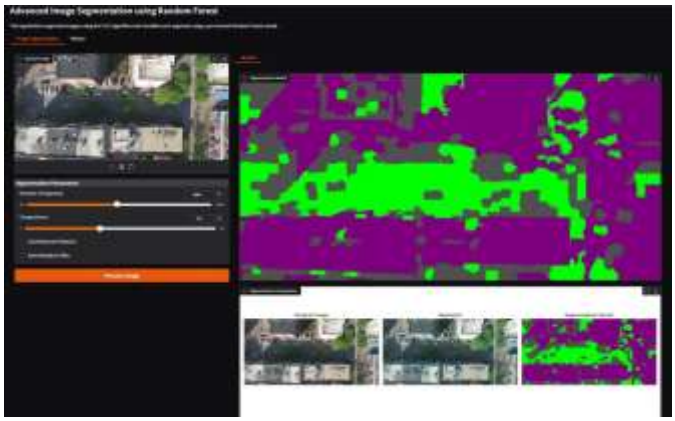

Fig-2: User Interface For Image Segmentation

4.2 Evaluation Metrics

To assess segmentation performance, the following metrics were used:

Pixel Accuracy (PA): Percentage of correctly classified pixels.Intersection over Union (IoU): Measures overlap between prediction and ground truth.

Precision: True positives divided by predicted positives. Recall: True positives divided by actual positives.F1-Score: Harmonic mean of precision and recall.

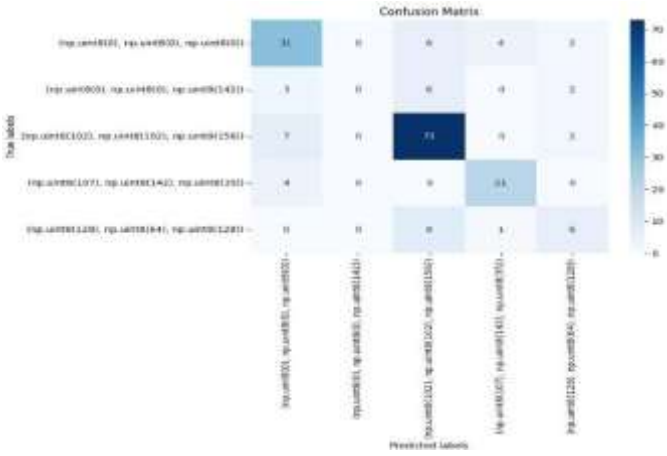

Fig - 3: Confusion Matrix for Image Segmentation

Figure 4 illustrates the training dynamics of the proposed semantic segmentation model over five epochs. The left subplot depicts the training loss, which shows a consistent downward trend, indicating effective minimization of the model's prediction error during training. The loss decreases from approximately 1.12 to 0.66, suggesting that the network is successfully learning meaningful patterns from the input data.This rapid convergence indicates that the model quickly adapts to the training data, possibly due to the enhanced input representations provided by the super-pixel preprocessing and the multiscale architecture.

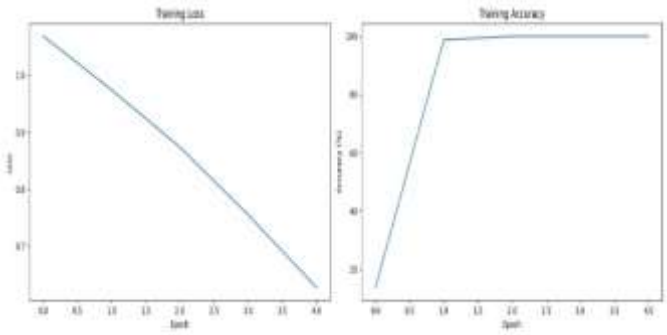

Fig-4: Model Training Progress: Loss and Accuracy over Epochs using Multiscale CNN

The below chart visualizes the importance of each input feature in the model's decision-making process. Feature 2 contributes the most, followed by Feature0 and Feature1.Understanding feature importance helps in feature selectionand model interpretation. Such insights can guide data preprocessing and improve model performance.

Fig-5: Segmented Image 1

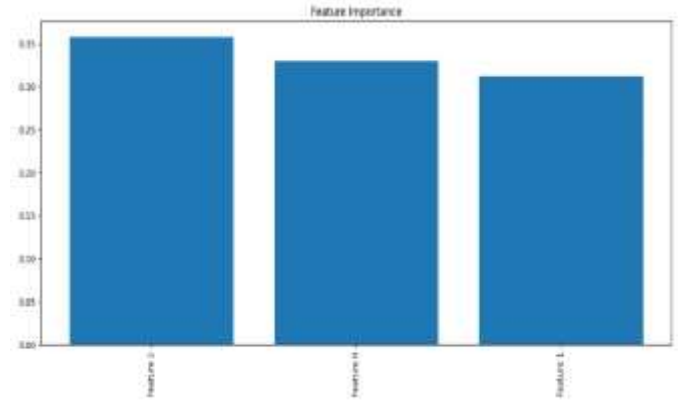

Fig-6: Relative Importance of Input Features in Model Prediction

4.3 Quantitative Results

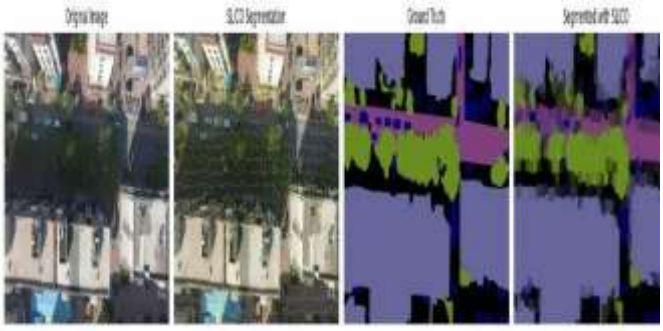

Fig-7: Image Segmentation using SLICO

The proposed architecture was evaluated on both original images and super-pixel enhanced images. The latter consistently outperformed the former, particularly in scenarios involving small or occluded objects.

These results demonstrate the benefit of incorporating SLICO super-pixels, which better preserve boundaries and their detection.

4.4 Qualitative Results

Visually, the segmentation maps generated by the proposed model show:

1. Sharper boundaries, especially between similar classes like roads and rooftops.
2. Better detection of small objects, such as cars and signboards, which are often missed by standard CNNs.
3. Reduced false positives in background regions.
4. The visual consistency of segmentations across

multiple test images supports the model's generalizability to different scenes and lighting conditions.

4.5 Discussion

The integration of super-pixel algorithms with convolutional neural networks (CNNs) has proven to be an effective strategy for enhancing semantic segmentation performance in UAV imagery. Super-pixels play a critical role in preprocessing by grouping

	Precision	recall	F1score
(np.unit8(0), np.unit8(0))	0.6	0.8	0.73
(np.unit8(10),np.unit8(10))	0.86	0.5	0.67
Accuracy	0.71	2	
Macro avg	0.67	0.8	0.72
Weighted avg	0.75	0.7	0.70

Fig-8 Comparison Table of Image Segmentation

pixels into perceptually meaningful clusters, which not only simplifies the input but also preserves important object boundaries. This reduction in image complexity helps mitigate the impact of background noise and redundant pixel-level information, making the subsequent learning process more focused and efficient. By emphasizing spatial structure and local coherence, super-pixels provide the CNN with a more organized and semantically enriched representation of the image.On the other hand, CNNs bring powerful feature extraction capabilities to the table. When applied to super-pixel-

enhanced inputs, CNNs can more effectively capture semantic patterns at both local and global levels. This synergy enables the network to focus on important object regions and disregard irrelevant background clutter. Furthermore, the architecture's ability to learn hierarchical representations complements the structured input provided by super- pixels, leading to more accurate pixel-level classification. The combination of spatial coherence from super-pixels and semantic depth from CNNs forms a robust foundation for high-performance segmentation.

A key innovation in the proposed framework is the use of multiscale input processing. By feeding the network with image representations at multiple resolutions, the model becomes more resilient to the scale variations inherent in UAV imagery. This is especially important in aerial scenes, where the same object may appear at vastly different scales due to changes in altitude, angle, or field of view. Multiscale architectures empower the model to simultaneously analyze fine details—such as small vehicles or roof textures—and broader structures like buildings or roads, leading to more consistent segmentation across diverse scenarios. Despite these benefits, a few limitations were identified during experimentation. In low-light or low-contrast conditions, the super-pixel generation process may produce inaccurate clusters, grouping dissimilar regions together or failing to delineate true object boundaries.This can introduce noise into the training process and potentially reduce segmentation precision. Additionally, the adoption of multiscale inputs and model replication naturally leads to a slight increase in computational complexity and training time. While these trade-offs do not hinder functionality, they are important considerations for real-time or resource-constrained deployments.The overall improvements in segmentation accuracy, boundary preservation, and robustness to scale variation clearly outweigh the associated computational costs. The proposed architecture remains lightweight and modular,

making it well-suited for deployment on edge devices commonly used in UAV systems. With appropriate optimizations—such as model pruning, quantization, or inference acceleration—it can achieve efficient on-board processing without sacrificing performance. In summary, the integration of super- pixels and multiscale CNNs

represents a promising direction for advancing semantic segmentation in UAV imagery, especially in complex and dynamic environments. this approach opens avenues for further enhancement through adaptive super-pixel generation, attention mechanisms, and integration with transformer- based architectures. By incorporating contextual attention or temporal consistency across UAV video frames, future models can achieve even greater accuracy and stability in dynamic environments. Additionally,integrating this pipeline with other remote sensing modalities—such as LiDAR or thermal imaging—can provide complementary information, enriching the model's understanding of the scene. As UAV platforms continue to evolve with improved onboard computing capabilities, methods like the one proposed in this study will play a pivotal role in enabling real-time, intelligent perception for autonomous aerial operations in urban, rural, and disaster- stricken environments.

5.CONCLUSION

In this research, we presented a novel semantic segmentation framework tailored for UAV-captured aerial imagery, leveraging the combined strengths of SLICO-based superpixel segmentation and a multiscale convolutional neural network (CNN) architecture. The increasing adoption of UAVs across a variety of domains—including urban development, environmental surveillance, precision agriculture, and disaster management—has resulted in an abundance of high-resolution overhead imagery. However, translating this rich visual data into meaningful semantic information remains an open challenge due to factors such as high intra-class variability,scale diversity, complex object interactions, and visual ambiguities caused by occlusions and lighting conditions.To address these complexities, our proposed framework introduces a two-tiered approach. First, superpixel segmentation using the SLICO algorithm groups adjacent pixels into perceptually coherent regions, which simplifies the input space and enhances structural consistency. Unlike traditional pixel-level approaches, this method preserves object boundaries more effectively and reduces noise from irrelevant background textures. Second, a multiscale CNN is used to process input data at varying spatial resolutions, enabling the network to capture both low-level details (e.g., edges, textures) and high-level semantic context (e.g., object shapes and spatial relationships). This multiscale representation is crucial for overcoming the issue of scale variation, a well-known limitation in aerial imagery where the same object may appear in vastly different sizes depending on flight altitude and camera perspective.

Future Work

While the proposed framework demonstrates strong performance and practical deployability, there remain several avenues for enhancement and extension in future research. One prominent limitation observed is the sensitivity of superpixel clustering under extreme lighting conditions. In scenarios involving harsh shadows, overexposed regions, or low-contrast environments, SLICO may incorrectly group semantically unrelated pixels into the same superpixel, leading to segmentation inaccuracies. Addressing this issue may require the integration of illumination- invariant preprocessing techniques or the development of adaptive superpixel algorithms that dynamically adjust clustering behavior based on local texture and brightness statistics.In terms of computational efficiency, although our multiscale CNN design balances accuracy with resource usage, the parallel processing of multiple image scales introduces moderate overhead. This could present challenges for real-time applications on ultra-low-power edge devices commonly used in UAVs. Future work may explore the use of lightweight CNN architectures such as MobileNetV3, EfficientNet-Lite, or GhostNet, which are explicitly designed for mobile and embedded inference. Additionally, dynamic resolution selection strategies could be introduced, enabling the model to focus computational effort only on regions requiring multiscale analysis, rather than applying it uniformly across the image.

6.REFERENCES

[1]. Diao, Q.; Dai, Y.; Zhang, C.; Wu, Y.; Feng, X.; Pan, F, " Superpixel-Based Attention Graph Neural Network for Semantic Segmentation in Aerial Images" Fifth International Conference on Image Segmentation, Remote Sens. 2022, 14, 305.

[2]. Fan, C.-L. "Multiscale Feature Extraction by Using Convolutional Neural Network: Extraction of Objects from Multiresolution Images of Urban Areas", ISPRS Int. J. Geo-Inf. 2024, 13, 5.

[3]. Tian, S.; Lu, Q.; Wei, L. "Multiscale Superpixel- Based Fine Classification of Crops in the UAV Based Hyperspectral Imagery", Remote Sens. 2022, 14, 3292.

[4]. Liang Huang et al 2019 IOP Conf. Ser.: Earth Environ. Sci. 234 012022. "Unmanned Aerial Vehicle Remote Sensing Image Segmentation Method by Combining Superpixels with multi-features Distance Measure".

[5]. Lian and Ren Journal of Engineering and Applied Science 2024,72,17. "Unmanned aerial vehicle aerial image stitching method based on superpixel segmentation".

[6]. Mu, Y.; Ou, L.; Chen, W.; Liu, T.; Gao, D. "Superpixel-Based Graph Convolutional Network UAVForest Fire Image Segmentation",Drones 2024, 8, 142.

[7].LouisKouadioa, Ravinesh, C.Deob, ∗Vivekananda Byrareddya,JanF.Adamowskic,ShahbazMushtaqa,VanPhu ong Nguyend,"SLIC SUPERPIXELS FOR OBJECT DELINEATION FROM UAV DATA",Computers and Electronics in Agriculture.

[8]. Prasad, K. D. V., et al. "Cryptographic image-based data security strategies in wireless sensor networks." Journal of Discrete Mathematical Sciences and Cryptography, vol. 27, no. 2-A, 2024, pp. 293–304. Taru Publications.

[9].Papadeas, I.; Tsochatzidis, L.; Amanatiadis, A.; Pratikakis,I."Real-TimeSemantic Image Segmentation with Deep Learning for Autonomous Driving", A Survey. Appl. Sci. 2022, 11, 8802.

[10].Dekka, Satish, et al. International Journal of Emerging Technologies and Innovative Research (JETIR), "Software Fault Prediction Using Deep Neural Networks" vol. 11, no. 5, 14 May 2024, pp. f462–f474. IJPUBLICATION.

[11]. S. Dekka, K. Prameela, P. M. A. Reddy, M. S. Charan, and K.V.B.Sri,"Development of Smart Alerting System using Real Time Object Detection with Deep Learning" Int. Res. J. Eng. Technol. (IRJET), vol. 11, no. 5, pp. 467–477, May 2024.

[12]. D. ManendraSai, S. Dekka, M. Rafi, M. Apparao, M. T. Suryam, and G. Ravindranath, "Machine learning techniques based prediction for crops in agriculture," J. Survey Fish. Sci., vol. 10, no. 1S, pp. 3710–3717, 2023.

[13]. S. Dekka, P. Sowmya, T. Sai Sudha, S. Durgamma, and V. Aravindh, "An experimental approach of energy consumption in underwater wireless communication," Int. J. Res. Trends Innov., vol. 8, no. 7, pp. 218–226, Jul. 2023.

[14]. F. Teng et al.,Journal of Engineering and Applied Science 2024,72,18. "Assimilating satellite imagery and visible-near infrared spectroscopy to model".