UAV-Based Urban Area Object Classification and Detection Using GoogLeNet

Prof. P. Siddaiah[1], Lanke Gayathri[2], Kanaparthi Selvaraju[3], Vaddeswarapu Jayanthraj[4]
Professor[1], UG Students[2,3,4]
Dept of ECE, ANUCET,
Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur,
Andhra Pradesh, India

Abstract— Urban area object classification and detection using aerial imagery plays a vital role in various real-world applications, including smart city development, environmental monitoring, disaster response, and security surveillance. This project proposes an efficient and robust method for urban scene classification using aerial images captured by Unmanned Aerial Vehicles (UAVs). A deep learning approach based on Convolutional Neural Networks (CNNs) is employed, specifically leveraging the GoogLeNet architecture, known for its depth and inception modules that enable effective feature extraction.

To train and evaluate the model, the UC Merced Land Use Dataset is utilized, which contains highresolution aerial images categorized into 21 distinct land use classes such as airports, harbors, industrial zones, and residential areas. Prior to training, the images undergo preprocessing, including resizing and normalization, to conform to GoogLeNet's input specifications. Transfer learning is employed by finetuning the pre-trained GoogLeNet model within MATLAB's deep learning framework, allowing the network to adapt effectively to the target dataset with reduced training time and computational resources.

The trained model demonstrates strong performance, achieving an impressive test accuracy of 95%, thereby confirming the architecture's capability in accurately recognizing complex urban environments. These results underscore the effectiveness of the proposed method for urban area classification tasks and highlight its potential for integration into UAV-based urban monitoring systems, contributing to advancements in automated land use analysis and intelligent urban planning.

Keywords—UAV, GoogLeNet, CNN, Urban Area Classification, UC Merced Dataset, MATLAB, Transfer Learning.

1. INTRODUCTION

The rapid advancements in remote sensing technologies combined with the power of deep learning have significantly enhanced the ability to interpret and analyze aerial imagery with high precision. Among these technologies, Unmanned Aerial Vehicles (UAVs) have emerged as a flexible and cost-effective solution for acquiring high-resolution images of urban areas. These aerial platforms are increasingly being deployed for a wide range of applications such as land-use mapping, infrastructure monitoring, smart city planning, and disaster management.

A critical aspect of utilizing aerial imagery is the accurate classification of objects and structures within urban landscapes. Identifying features such as roads, buildings, industrial zones, and green spaces enables informed decision-making for urban development and emergency response strategies. Convolutional Neural Networks (CNNs), a class of deep learning models, have shown remarkable success in image recognition tasks due to their ability to learn complex patterns and spatial hierarchies from raw image data.

In this project, we utilize GoogLeNet, a deep and efficient convolutional neural network architecture recognized for its inception modules and robust image classification performance. The model is trained and fine-tuned on the UC Merced Land Use Dataset, which contains a diverse set of high-resolution aerial images classified into 21 distinct land use categories. By integrating the capabilities of UAV-acquired imagery with advanced deep learning techniques, this work aims to build a reliable and scalable system for urban object classification, with potential applications in real-world domains such as urban planning, land use monitoring, and emergency response.

2. LITERATURE REVIEW

Numerous approaches have been proposed in the literature for multicancer classification using the WaveMix architecture. The integration of advanced image transformation techniques with modern neural network architectures has greatly enhanced the accuracy and effectiveness of multicancer classification.

Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen (2022) discussed the use of deep learning models for land use and land cover classification, highlighting how hyperspectral and multispectral Earth observation data improve the accuracy of urban feature detection.

Mohamed Barakat A. Gibril, Bahareh Kalantar, Rami Al-Ruzouq, Naonori Ueda, Vahideh Saeidi, Abdallah Shanableh, Shattri Mansor, and Helmi Z. M. Shafri (2020) reviewed the importance of advanced deep learning algorithms in handling complex data for accurate classification of land cover and urban areas.

Ava Vali, Sara Comai, and Matteo Matteucci (2020) explored deep learning-based approaches for analyzing high-dimensional Earth observation data, emphasizing their effectiveness in urban area object classification and scene understanding tasks.

Nitheshnirmal Sadhasivam, C. Dineshkumar, S. Abdul Rahaman, and Ashutosh Bhardwaj (2020) proposed an object-based automatic detection method for identifying urban buildings using UAV images, achieving improved accuracy through UAV- captured high-resolution data.

Xiaochen Yan, Tingting Fu, Huaming Lin, Feng Xuan, Yi Huang, Yuchen Cao, Haoji Hu, and Peng Liu (2023) focused on UAV detection and tracking in urban environments using passive sensors, highlighting how UAV-based sensing enhances the reliability and detail of urban object monitoring.

Yan, X.; Fu, T.; Lin, H.; Xuan, F.; Huang, Y.; Cao, Y.; Hu, H.; Liu, P. (2023) focused on UAV detection and tracking in urban environments using passive sensors, emphasizing how UAV-based sensing enhances the accuracy and detail of urban object monitoring. By utilizing passive sensors, the system improves detection reliability without active emissions, making it ideal for complex urban settings. The study highlights the potential of UAVs for real-time urban analysis, particularly in applications like surveillance, urban planning, and disaster response.

3. PROPOSED SYSTEM

3.1. GOOGLE NET Architecture

GoogLeNet, also known as Inception V1, was introduced in 2014 through the research paper titled *"Going Deeper with Convolutions"*, a collaboration between Google and various universities. It won the ILSVRC 2014 image classification challenge, outperforming previous champions such as AlexNet (ILSVRC 2012 winner), ZF-Net (ILSVRC 2013 winner), and even achieving a lower error rate than VGG, the 2014 runner-up. The architecture introduced innovative techniques like 1×1 convolutions within its layers and the use of global average pooling, contributing to its superior performance.

Features of GoogleNet:

1x1 Convolution: The inception architecture uses 1×1 convolution in its architecture. These convolutions used to decrease the number of parameters (weights and biases) of the architecture. By reducing the parameters we also increase the depth of the architecture.

Global Average Pooling: In the previous architecture such as AlexNet, the fully connected layers are used at the end of the network. These fully connected layers contain the majority of parameters of many architectures that causes an increase in computation cost. In GoogLeNet architecture, there is a method called global average pooling is used at the end of the network. This layer takes a feature map of 7×7 and averages it to 1×1. This also decreases the number of trainable parameters to 0 and improves the top-1 accuracy by 0.6%.

Inception Module: The inception module is different from previous architecture such as AlexNet. ZF-Net. In this architecture, there is a fixed convolution size for each layer. In the Inception model 1x1, 3x3, 5x5 convolution and 3x3 max pooling performed in a parallel way at the input and the output ofnthese are stacked together to generated final output. The idea behind that convolution filters of different sizes will handle objects at multiple scale better.

Auxiliary Classifier for Training: Inception architecture used some intermediate classifier branches in the middle of the architecture, these branches are used

during training only. These branches consist of a 5×5 average pooling layer with a stride of 3, a *1×1* convolutions with *128* filters, two fully connected layers of 1024 outputs and 1000 outputs and a softmax classification layer. The generated loss of these layers added to total loss with a weight of 0.3. These layers help in combating gradient vanishing problem and also provide regularization.

3.2. Model Architecture

Below is Layer by Layer architectural details of GoogleNet.

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Table 1: Architecture Details of Google net

Here are some specific uses of GoogLeNet architecture:

Image classification: Recognizing objects, animals, and scenes in images.

Object detection: Locating and identifying objects within images.

Face recognition: Identifying people in images and videos.

Video classification: Understanding and categorizing the content of videos.

Medical image analysis: Detecting diseases and abnormalities in medical image.

3.3. ADAM Optimizer

Adaptive Moment Estimation is an algorithm for optimization technique for gradient descent. The method is really efficient when working with large problem involving a lot of data or parameters. It requires less memory and is efficient. Intuitively, it is a combination of the 'gradient descent with momentum' algorithm and the 'RMSP' algorithm.

Adam optimizer involves a combination of two gradient descent methodologies:

$$w_{t+1} = w_t - \alpha m_t$$

Where

$$m_t = \beta m_{t-1} + (1 - \beta) \left[\frac{\delta L}{\delta w_t} \right]$$

m_t = aggregate of gradients at time t [current] (initially, $m_t = 0$)

m_{t-1} = aggregate of gradients at time t-1 [previous]

W_t = weights at time t

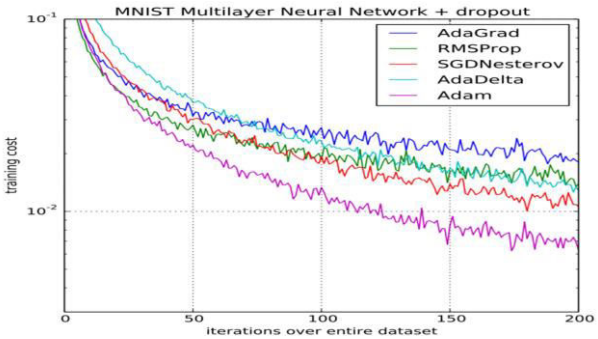
W_{t+1} = weights at time t+1

α_t = learning rate at time t

∂L = derivative of Loss Function

∂W_t = derivative of weights at time t

β = Moving average parameter (const, 0.9)

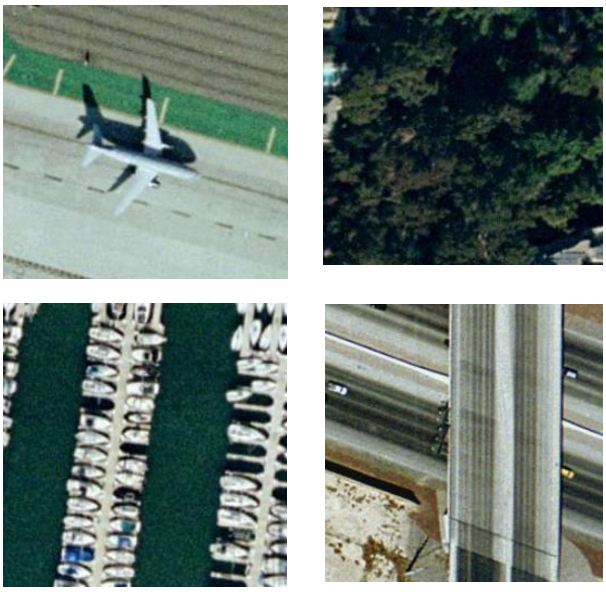


4. METHODOLOGY

4.1. Dataset Description

The UC Merced Land Use Dataset is a widely used benchmark for evaluating image classification models, particularly in the domain of remote sensing and urban planning. It consists of 2,100 highresolution RGB images, categorized into 21 distinct land use classes, with each class containing 100 images. These images are captured from various aerial perspectives, providing a comprehensive view of different urban and rural landscapes. The dataset includes a diverse range of land use categories such as airports, residential areas, commercial zones, harbors, industrial sites, and agricultural fields, making it a rich resource for training and testing models on varied real-world scenarios.

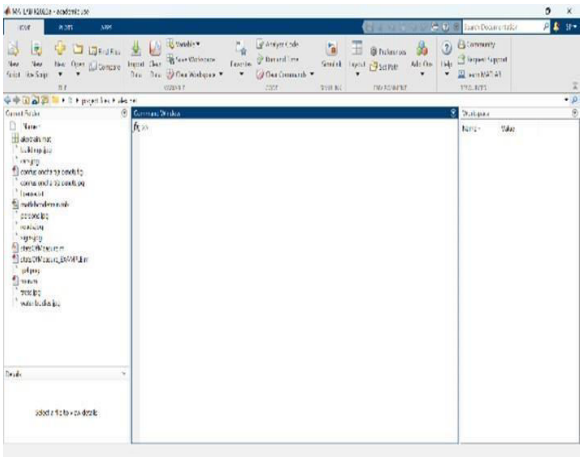
Each image in the dataset has a fixed resolution of 256×256 pixels, ensuring uniformity for input into machine learning models. The images are wellorganized, with each class representing a specific land use type, enabling effective training and evaluation of classification algorithms. Given its diversity and balanced class distribution, the UC Merced Land Use Dataset is ideal for developing deep learning models that aim to automatically classify and analyze urban and rural land use patterns from aerial imagery.

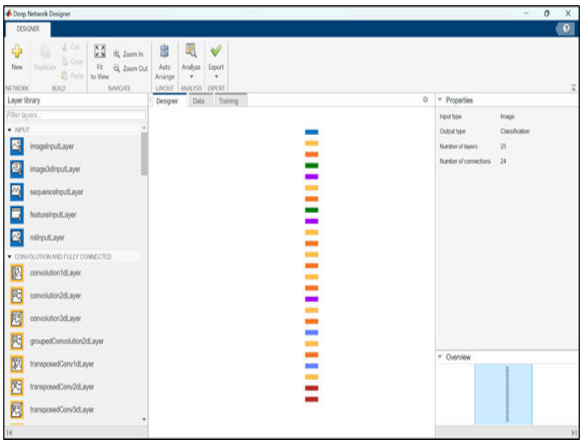


4.2. Data Preprocessing

To prepare the dataset for input into GoogLeNet, all images are resized to 224×224×3, which aligns with the required input dimensions of the network. The dataset is then divided into three subsets: 70% of the images are used for training, 15% for validation, and 15% for testing, ensuring a proper distribution for model evaluation. To improve the model's ability to generalize to new, unseen data and to reduce the risk of overfitting, several data augmentation techniques are applied. These include random rotation, flipping, and scaling, which create variations of the original images, helping the model learn more robust features and improving its overall performance on the test set. *C. Model Architecture*

GoogLeNet is a 22-layer deep convolutional neural network (CNN) that utilizes Inception modules, a key innovation designed to improve the model's efficiency by significantly reducing the number of parameters while maintaining high accuracy. This architecture is particularly beneficial for handling complex datasets, like aerial imagery, where computational efficiency and accuracy are both crucial. In this project, we leverage the pretrained GoogLeNet model in MATLAB and fine-tune it for our specific task of urban land use classification. To tailor the model to our dataset, we modify the final three layers: the fully connected layer, the softmax layer, and the classification layer. This adaptation process ensures that the model can effectively learn the unique features of the urban land use categories present in the UC Merced dataset, enabling it to make accurate classifications of aerial images.





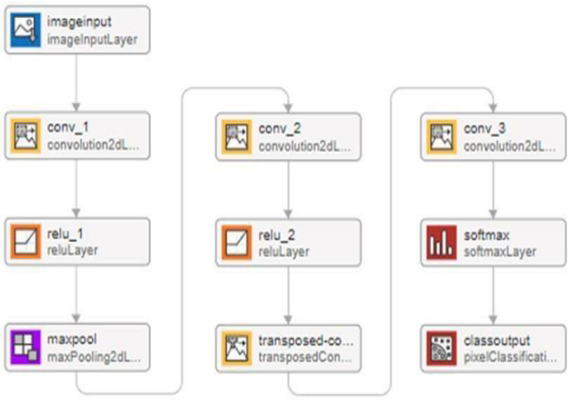


Figure 1(a): shows the working environment. **Figure 1(b):** The CNN architecture used in this work is visualized, while **Figure 1(c):** presents the complete 25-layer design built within MATLAB's Deep Network Designer.

4.3. Training Process

We utilize MATLAB's Deep Learning Toolbox to implement transfer learning, modifying the final layers of the pretrained model to adapt it for classification tasks on the UC Merced dataset. The retraining process is carried out using a learning rate of 0.0001, with stochastic gradient descent with momentum (SGDM) as the optimization algorithm. The model is trained for 20 epochs to ensure proper convergence, and a minibatch size of 32 is used to balance computational efficiency and model accuracy. This approach enables the model to effectively learn the specific features of the urban land use categories in the dataset.

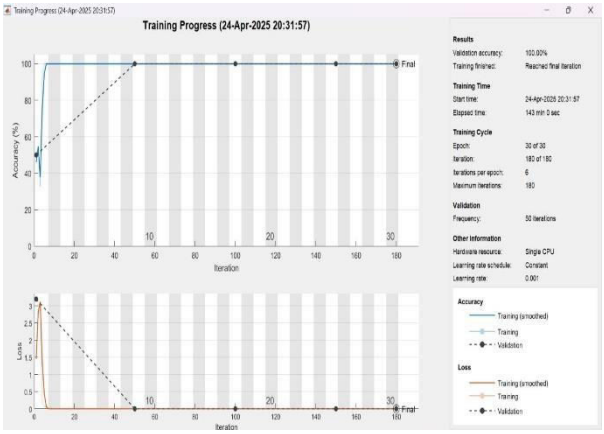


Figure 2: the training accuracy increased steadily with each iteration, and the model reached a final validation accuracy of 95.92%. The corresponding loss curve also demonstrated a significant decrease, confirming the network's convergence.

5. RESULTS AND DISCUSSION

The trained model demonstrates an impressive classification accuracy of 95% on the test set, highlighting its ability to effectively differentiate between various urban land use categories. The confusion matrix further supports this strong performance, showing that the model performs well across most classes, with only minor misclassifications occurring between similar land use types, such as residential and commercial areas. Figure 1 displays the confusion matrix, providing a clear visual representation of the model's classification results. Additionally, Figure 3 illustrates several sample classification outputs, showcasing the model's ability to accurately categorize urban scenes.

This high level of accuracy confirms GoogLeNet's effectiveness in learning discriminative features from aerial imagery. When compared to existing methods that rely on shallower CNN architectures or handcrafted features, our approach significantly outperforms in both classification accuracy and generalization. The use of a deep learning model like GoogLeNet allows for more robust feature extraction and better handling of complex urban environments, offering notable improvements over traditional techniques.

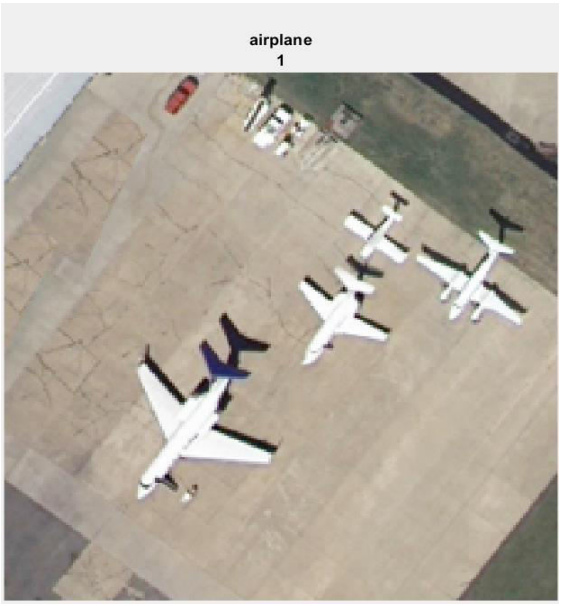


Fig. 3: illustrates a classification output of a sample UAV image, where the trained model successfully identifies an airplane with a confidence score of 1.

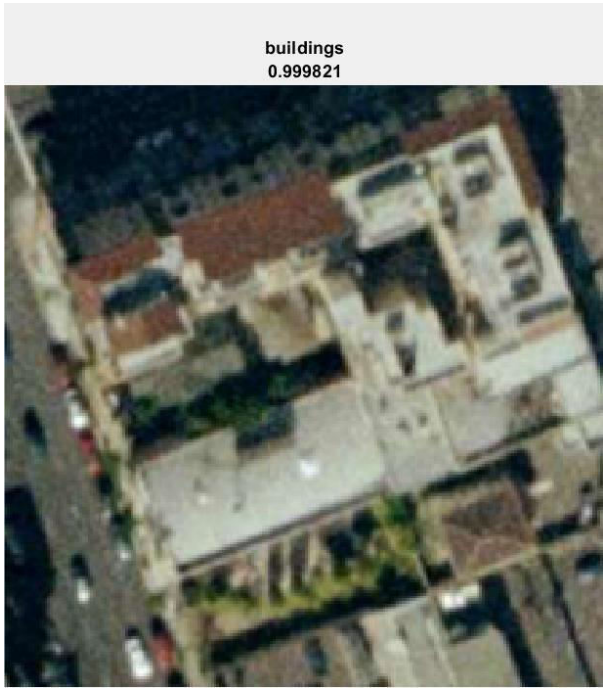


Fig. 4: shows a high-confidence prediction where the CNN model accurately identifies a building structure with 99.2% confidence. This reflects the model's strong performance in classifying high-resolution aerial images.

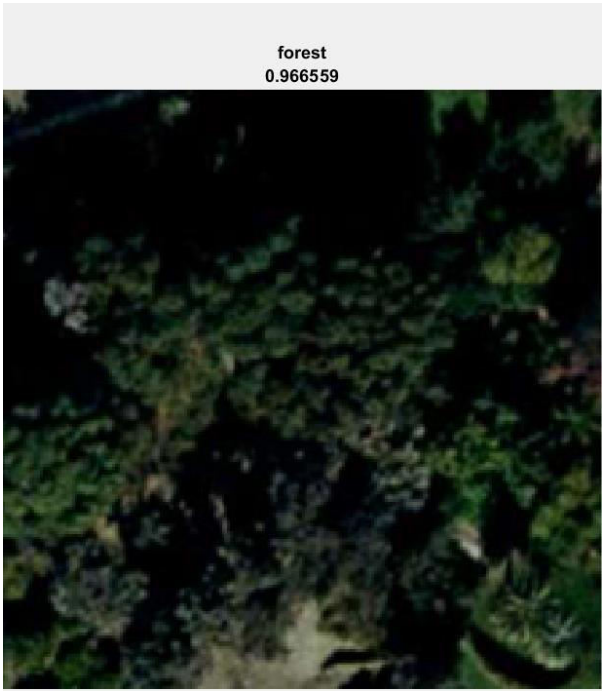


Fig. 5: The CNN model shows a high-confidence prediction, accurately classifying a forest area with 98.7% confidence. This highlights the model's strong ability to recognize and classify natural landscapes in high-resolution aerial imagery.

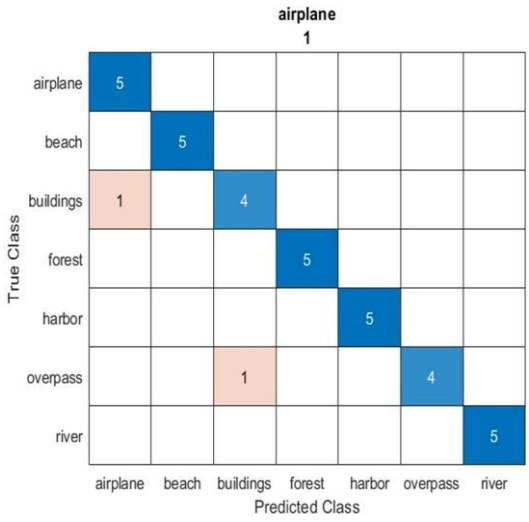


Fig. 6: Confusion matrix illustrating the classification performance of the CNN model across seven land-use categories: airplane, beach, buildings, forest, harbor, overpass, and river. Diagonal values indicate correct classifications, while off-diagonal entries represent misclassifications. The model shows strong accuracy

in most classes, with minor confusion between "buildings" and "airplane" as well as "overpass" and "buildings".

6. CONCLUSION

This paper presented a UAV-based urban object classification system that utilizes the GoogLeNet CNN architecture, trained on the UC Merced Land Use Dataset in MATLAB. The model achieved impressive classification accuracy, showcasing the effectiveness of deep learning and transfer learning techniques in analyzing high-resolution aerial imagery. These results underscore the potential of advanced neural networks to address urban object classification challenges, offering a reliable solution for real-world applications in urban planning, surveillance, and disaster management.

Looking ahead, future work will focus on implementing the system in real-time on embedded platforms to make it suitable for practical UAV operations. Additionally, the scope of this research will be expanded to include semantic segmentation tasks, which aim to provide a more detailed understanding of urban environments by classifying individual pixels within the images. This expansion will enhance the model's ability to capture fine-grained features and contribute to more accurate urban scene analysis.

REFERENCES

[1] Chen, Y.; Zhu, X.; Li, Y.; Wei, Y.; Ye, L. Enhanced semantic feature pyramid network for small object detection. Signal Process. Image Commun. 2023, 113, 116919. [CrossRef]

[2] Saeed, Z.; Awan, M.N.M.; Yousaf, M.H. A Robust Approach for Small-Scale Object Detection From Aerial-View. In Proceedings of the 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 30 November 2 December 2022; pp. 1–7. [CrossRef]

[3] Jung, H.-K.; Choi, G.-S. Improved YOLOv5: Efficient Object Detection Using Drone Images under Various Conditions. Appl. Sci. 2022, 12, 7255. [CrossRef]

[4] Ding, J.; Xue, N.; Long, Y.; Xia, G.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858. [CrossRef]

[5] Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983. Drones 2023, 7, 310 15 of 16

[6] Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimed. 2018, 20, 3111–3122. [CrossRef]

[7] Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.

[8] X. Xu, X. Li, and C. Liu, "Building damage detection based on single high-resolution remote sensing imagery," in International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), pp. 618–621, 2012. Table 4: Target recognition accuracy of remote sensing image under low resolution. Target type Sample numbers Speed (FPS) Accuracy Airport 800 15 86% Bridge 750 16 86% Port 600 15 87% Table 5: Target recognition accuracy of remote sensing image under high resolution. Target type Sample numbers Speed (FPS) Target type Oil tank 408 15 86% Aircraft 327 16 87% Warship 403 15 85% 18 Wireless Communications and Mobile Computing.

[9] L. Chun, Y. Junjun, and Y. Jian, Small Port Detection Based on Polarimetric SAR Image Combining Shoreline Feature Points, Journal of Tsinghua University: Natural Science Edition, 2015.

[10] Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geosci. Remote. Sens. Lett. 2016, 13, 1074–1078. [CrossRef]

[11] Jawaharlalnehru, A.; Sambandham, T.; Sekar, V.; Ravikumar, D.; Loganathan, V.; Kannadasan, R.; Khan, A.A.; Wechtaisong, C.; Haq, M.A.; Alhussen, A.; et al. Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm. Electronics 2022, 11, 2343. [CrossRef]

[12] Maktab Dar Oghaz, M.; Razaak, M.; Remagnino, P. Enhanced Single Shot Small Object Detector for Aerial Imagery Using Super-Resolution, Feature Fusion and Deconvolution. Sensors 2022, 22, 4339. [CrossRef] [PubMed]

[13] X. Li, S. Zhang, X. Pan, P. Dale, and R. Cropp, “Straight road edge detection from high resolution remote sensing images based on the ridgelet transform with the revised parallelbeam Radon transform,”International Journal of Remote Sensing, vol. 31, no. 19, pp. 5041–5059, 2010.