Journal of Advanced College of Engineering and Management

A Review on Aerial Image Classification Techniques for Urban Planning

[1]Aayushma Pant, [2]Mission Shrestha, [5*]Prashant Acharya, [3]Sakshi Poudyal, [5]Sritu Yadav, [4]Chiranjibi Katuwal, [6]Sandesh Lamsal

[1]*Department of Electronics & Computer Engineering, Tribhuvan University*
[2]*Department of Computer Science & Engineering, Kathmandu University*
[3]*Department of Computer Science & Engineering, Pokhara University*
[4]*Department of Political Science, Tribhuvan University*
[5]*Kathmandu Model Secondary School*
[6]*Department of Civil Engineering, Tribhuvan University*
[*]*Corresponding Author: [*]prashantacharya.business@gmail.com*

DOI: 10.3126/jacem.v11i1.84534

Abstract
Aerial image classification is important in urban planning by providing detailed insights into land usage, environmental impact, and land cover analysis. This review paper explores the different deep-learning models applied to aerial images, focusing on their performance, and architecture and its application. Several models, including convolutional neural networks (CNNs), relation-enhanced multiscale networks, and attention-based architectures, are evaluated using high-resolution datasets such as the ISPRS Potsdam and UC Merced. The study compares the accuracy, feature extraction techniques, and classification results across these models, comprehensively analyzing their applicability to urban land cover and land use mapping. Furthermore, this paper examines the challenges associated with aerial image classification, including handling occlusions, spatial resolutions, and model generalization. The findings highlight the effectiveness of deep learning techniques in addressing complex urban planning tasks and offer recommendations for future research. This review serves as a comprehensive guide for researchers and practitioners seeking to enhance aerial image classification methods in urban planning.

Keywords*— Image classification, Computer vision, Neural Networks, Urban Planning, CNN*

1. INTRODUCTION

The rapid, unplanned urbanization over the past six decades (Singh, 2024, p.19) [1] poses significant challenges for modern cities, such as traffic congestion, waste management, and infrastructure development (Lempesis, 2023, p.1) [2]. These challenges are exacerbated by outdated planning approaches and traditional individualistic behaviors (European Commission, 2017; Rode & Hoffman, 2015) [3]. Therefore, digitization and transformation in urban planning methods are crucial to improving the quality of life.

Planning urban regions with a focus on different land uses, residential areas, and vegetation zones can result in systematic and well-developed cities. Various studies have incorporated computer vision, robotics, and remote sensing methods to estimate statistical parameters from a bird's-eye view for urban management (Fyleris, 2022, p.2) [4]. By utilizing images captured by drones, satellites, and aircraft, we can analyze land use, and monitor environmental change by using various image classification algorithms. However, integrating aerial image

classification into urban planning has its challenges. While the technology offers numerous benefits, such as enhanced data insights and improved decision-making, issues like data quality, scalability, and integration with existing urban planning systems remain significant obstacles. This review paper aims to explore the various techniques and machine learning models used for land cover classification focused on aerial images.

A. Motivation

The motivation behind this review is to help address challenges caused by rapid urbanization and the limitations of traditional urban planning methods.

1. Address Urban challenges:
   This review paper aims to explore various deep learning approaches used in aerial images that can tackle urban issues like the management of residential areas, environment monitoring etc.
2. Technological Advancement:
   The review is also motivated by the potential of computer vision and Machine learning models to revolutionize urban planning.

B. Objectives

1. Evaluate Model Performance:
   The review will assess the effectiveness of various models across different image resolutions, determining the best approaches for urban planning applications.
2. Compare Classical and Deep Learning Techniques:
   The objective includes a detailed comparison of traditional methods versus deep learning approaches, highlighting their respective strengths and limitations.
3. Analyze Challenges and Solutions:
   The review will focus on identifying and addressing challenges such as data quality, scalability, and occlusions, providing strategies to overcome these obstacles.

2. BACKGROUND

A. Background Of The Study

Aerial image classification has emerged as a powerful tool for urban planners, especially when it comes to obtaining valuable insights that can inform decision-making. Different image classification models perform differently across image resolutions in urban planning contexts. For example, the MobileViT model achieved better metrics at high resolutions leading to improved multi-label classification of urban green spaces, which is necessary for studying urban growth and land use change (Lin et al., 2024) [5]. Moreover, advanced techniques like neural networks, Markov random fields, and generative adversarial networks are also explored for their potential in this field.

Urban planning tasks such as data collection, problem identification, public participation, plan implementation and decision-making can be supported by CV techniques that extract useful insights from images captured by satellites, drones and on the ground (Marasinghe et al., 2023). In addition, these methods combine

multi-scale imagery from different viewpoints to enable precise 3D reconstructions of complex urban environments (Rumpler et al., 2017) [6]. As a result, recent studies have hinted at the potential of advanced deep learning techniques in large-scale land use classification and detection of urban patterns using satellite images, especially through CNNs (Albert et al., 2017) [7]. By measuring various aspects of cities such as built environment, natural features, human interactions and infrastructure they can help identify errors or opportunities for improvement (Ibrahim et al., 2020) [8]. Therefore, it is possible to use this approach to gain valuable information from street-level imagery that highlights attributes like greenery presence, and pavement materials used in construction facades among other things (Liu & Sevtsuk, 2024) [9]. LiDAR and close-range imagery could be employed to estimate building usable floor area which would enhance property-related information for investment analysis, and fiscal projection purposes besides planning functions in some cases (Janowski et al., 2021) [10]. Nonetheless, there are still challenges like ambiguous attribute definitions and standardized measurement methods that constrain their application.

There are issues related to image identification that CV applications in urban planning face. These include data quality such as unclear attribute definitions and lack of standardization measurement methods (Liu & Sevtsuk, 2024) [11], scalability when working on large-scale spatiotemporal analyses of satellite imagery (Tekouabou et al., 2022) [12], occlusions occurring in very-high-resolution images with complex backgrounds which make it difficult to extract building information accurately (Majd et al., 2019) [13]. Additionally, the object-based appropriate segmentation scale parameters determination is a problem. Recent research has addressed these challenges by focusing on data quality, scalability, and occlusions. Satellite image classification is challenged by data availability, quality, and distribution issues that may be handled using scalable deep learning methods (Tehsin et al., 2023) [14]. Airborne hyperspectral imaging helps urban land use classification to be more precise about mixed pixel problems as well as the same spectral signatures. To address occlusions in crosswalk detection Zhang et al. (2024) [15] have compared aerial view, street-view and dual-perspective methods with the latter two proving more effective for heavily occluded areas. Arndt and Lunga (2021) [16] present a scalable deep-learning approach which classifies urban structural units across multiple cities surpassing traditional approaches.

Adopting CV for urban planning faces its own set of challenges such as limited data and consideration of constraints on this technology to avoid potential pitfalls. Moreover, remote sensing applications in aerial data analysis for urban planners provide the best opportunities to make better decisions concerning the future of cities rather than relying on outmoded GIS or manual planning methods.

B. Dataset Comparison

The data sources have extended from public domain databases such as Eurostat to paid and regional databases data that have some trade-offs between repeatability and context.

We have studied different datasets as shown in the table below:

Table I: Datasets Comparison

Datasets	Description
WHU building (WHU dataset and China typical building dataset (CHN dataset))	Building datasets from China.
Eurostat dataset	10 classes, 27000 labelled images, Sentinel-2 satellite data.
Vienna and Potsdam datasets	High-resolution aerial imagery with varying resolutions.
GID remote sensing images	Five GID datasets, 150 GF-2 satellite images.
ISPRS Vaihingen, Shanghai datasets	High-resolution aerial imagery.
Ziyuan-3 (ZY-3 Satellite Image, Sentinel-1A SAR Image, Auxiliary Datasets)	Multi-source data for urban analysis.
GaoFen-5 (GF-5 Hyperspectral Satellite Imagery, Landsat 8 Multispectral Data)	Hyperspectral and multispectral satellite data.
Airborne LiDAR 2015, Aerial Images 2014	LiDAR and aerial image data for urban modeling.
UAV images with multiple spectral bands, DSM data	Unmanned aerial vehicle (UAV) imagery with various spectral bands and digital surface models (DSM).

3. METHODOLOGIES

A. Model Architecture

1. Convolutional Neural Networks (CNNs):

a. BGC-Net: BGC-Net, or Building Graph Convolutional Network, is a new model introduced by Wenzhuo Zhang to enhance the extraction of buildings from high-resolution aerial images. The core modules of BGC-NET include two main components:

- Atrous Attention Pyramid (AAP) Module: This module combines the attention mechanism with atrous convolution. It is designed to improve the model's ability to extract buildings of various sizes by fusing multi-scale features effectively. The attention mechanism helps the model focus on relevant parts of the image, enhancing the extraction process.
- Dual Graph Convolutional (DGN) Module: This module is built on the principles of Graph Convolutional Networks (GCNs), this module adds long-range contextual information to the segmentation process. It significantly improves the accuracy of object edge segmentation, which is crucial for delineating building boundaries accurately.

BGC-Net integrates advanced deep-learning techniques to enhance the extraction of buildings from aerial imagery. It addresses common challenges faced by traditional methods and demonstrates superior performance in various evaluations.

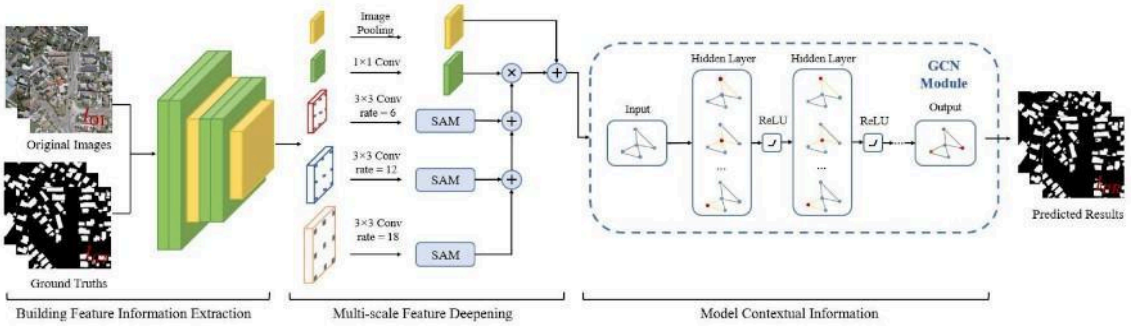

Fig.1: BCG-Net Architecture

b. U-Net: The U-Net model is a convolutional neural network architecture specifically designed for semantic segmentation tasks, which is crucial for processing remote sensing images. In this paper by (Xiaoling Xie. etl, 2022) [17]. the U-Net is improved by incorporating pooling index up sampling and dimension superposition, allowing it to effectively extract both high-level abstract features and low-level detail features, thereby reducing the loss of edge information during deconvolution. The architecture follows an encoder-decoder structure, where the encoder captures semantic information and the decoder generates the segmentation map, facilitating end-to-end training with fewer remote sensing training images and faster segmentation efficiency. The improved U-Net model is utilized for the classification of land use types in remote sensing images, enabling dynamic monitoring of urban and rural planning.

c. DeepLabv3, ResNet50: (Fyleris, T.etl 2022) [18] utilized the DeepLabv3 model with a ResNet-50 backbone for urban change detection from aerial images. The ResNet50 architecture serves as the backbone for DeepLabv3. ResNet-50 is a convolutional neural network that includes 50 layers and is known for its residual learning framework, which helps in training deeper networks by mitigating the vanishing gradient problem. The training process conducted using the GluonCV MXNet framework, involved transfer learning, where the model was initialized with weights pre-trained on the ImageNet dataset.

d. Various CNNs (AlexNet, ResNet50, DenseNet, etc.): Antonio Rangel did the experimentation of different convolution neural networks for determining the lang cover aerial image classification. He compared alexNet, Resnet and Dense Net, MobileNet V3 for land cover classification. He also compared with transformer-based model ViT, MaxVision Transformer on EuroSat datasets.

   AlexNet, ResNet-50, and DenseNet are three influential deep-learning architectures. AlexNet, introduced by Krizhevsky et al. in 2012, revolutionized image classification by utilizing deep convolutional layers, ReLU activations, and dropout to significantly reduce overfitting, achieving groundbreaking performance in the ImageNet competition. It consists of eight layers, five convolutional layers followed by three fully connected layers.

ResNet-50, introduced by He et al. in 2015, is a 50-layer deep convolutional neural network that addresses the vanishing gradient problem. It introduces the concept of residual learning, where identity shortcut connections (or skip connections) are used to allow gradients to flow more effectively through the network. This enables the training of much deeper networks without the degradation problem, achieving superior performance on complex tasks like image classification, object detection, and more.

DenseNet, proposed by Huang et al. in 2017 [19], connects each layer to every other layer in a feed-forward fashion, where each layer receives inputs from all previous layers and passes its feature maps to all subsequent layers. It mitigates the vanishing gradient problem and promotes feature propagation, resulting in more efficient networks with fewer parameters compared to traditional architectures.

2. Transformer-Based Models
    a. ViT, Swin Transformer, MaxViT:

Vision Transformer (ViT) is a novel deep learning architecture introduced by Dosovitskiy et al. in 2020 [20] that applies the principles of transformers, originally designed for natural language processing, to the domain of computer vision. ViT treats an image as a sequence of patches, where each patch is linearly embedded and processed similarly to tokens in NLP models. The architecture relies heavily on self-attention mechanisms to capture global relationships between different parts of the image. ViT demonstrated that with sufficient training data, transformers could outperform traditional convolutional neural networks (CNNs) in various image classification tasks, marking a significant shift in how visual data is processed.

Swin Transformer (Shifted Window Transformer) is an advanced vision transformer architecture introduced by Liu et al. in 2021 [21]. It builds on the Vision Transformer by incorporating a hierarchical structure and the use of shifted windows. The architecture processes images by dividing them into non-overlapping windows, within which self-attention is calculated. The windows are shifted at each layer to capture cross-window connections and global information more effectively. Swin Transformer achieves state-of-the-art performance across various vision tasks, including object detection and semantic segmentation, while being computationally efficient.

MaxViT is a recent hybrid vision transformer architecture introduced by Tu et al. in 2022 [22]. It combines the strengths of convolutional neural networks (CNNs) and transformers, utilizing multi-axis attention mechanisms. MaxViT incorporates both local and global attention by combining grid and window-based self-attention, enhancing its ability to model both fine-grained details and long-range dependencies in images. This architecture is designed to efficiently scale across various image resolutions and achieve high performance in image classification, object detection, and other vision tasks, making it a versatile and powerful model in the realm of computer vision.

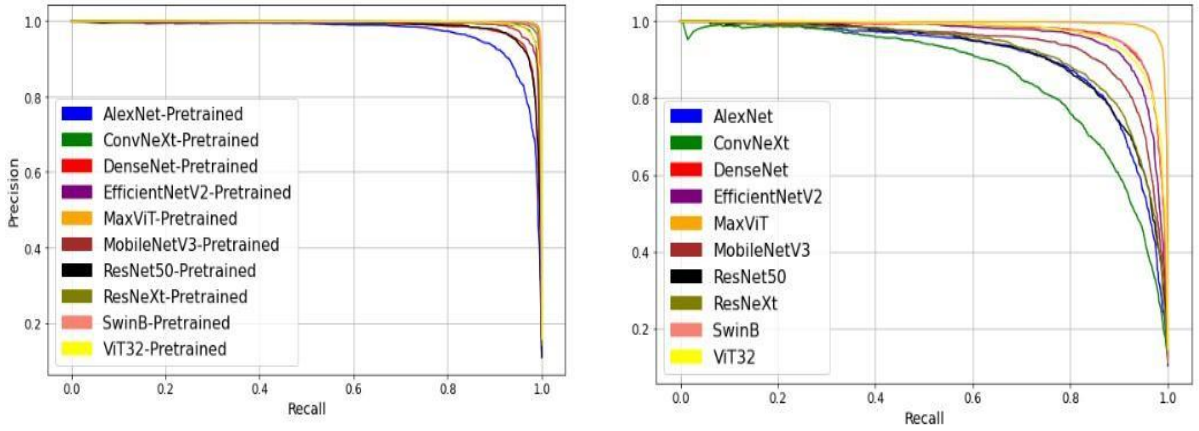

Fig.2: Precision/Recall curves for the ten models.
(a) Trained from scratch **(b)** Using pre-trained weights on ImageNet
The precision/ recall curves for ten models show that ConvNeXt is on the lower end and MaxViT on the higher end when trained from scratch whereas for pre-trained models the curves are closer to each other, with AlexNet as the lowest-performance model and MaxViT on the higher end proving that transformer-based model outperform convolutional neural networks for land cover classification method.

3. Hybrid Models:

a. MMAFNet

The Multi-Modality and Multi-Scale Attention Fusion Network (MMAFNet) introduces a comprehensive approach to land cover classification from Very High Resolution (VHR) remote sensing images by integrating several advanced techniques. The network employs a Multi-Modality Fusion Module (MFM) that adeptly processes Infrared Red Green (IRRG) and Digital Surface Model (DSM) data through its three branches—spectral, depth, and fusion—while utilizing channel attention for effective feature reorganization. For multi-scale feature extraction, the Multi-Scale Spatial Context Enhancement Module (MSCEM) combines Atrous Spatial Pyramid Pooling (ASPP) with a non-local block, leveraging atrous convolutions with rates of 3, 6, and 9, along with 1x1 convolution and global average pooling to capture extensive spatial context and long-range dependencies. Additional techniques include residual skip connections to enhance feature preservation and depth separable convolutions to manage parameter count. The network is trained using the PyTorch framework with a ResNet50 backbone pre-trained on ImageNet, employing an SGD optimizer with momentum (0.9) and weight decay (0.004), an initial learning rate of1 × 10−3, and a cross-entropy loss function over 250 epochs with a batch size of 16.

Comparative evaluations with methods such as DeepLab v3+, MANet, DSMFNet, DP-DCN, and REMSNet reveal that MMAFNet surpasses these approaches in both the Potsdam and Vaihingen datasets, notably enhancing classification accuracy for trees, low vegetation, and complex scenes, while effectively mitigating interference from shadows and occlusions. Ablation studies demonstrate that the MFM, MSCEM, and residual skip connections contribute significant improvements to the mean

F1-score and overall accuracy, with the full MMAFNet achieving a mean F1-score enhancement of 5.8% and an overall accuracy improvement of 5.3% over baseline methods. (Lei et al. (2021b) [23]. MMAFNet proposed by Tao Lei, outperforms other methods in all evaluation metrics, achieving a mean F1-score and OA of 92.34% and 1.04%, respectively. They also find that their method is more robust to shadows and occlusions. Additionally, they find that their method can capture targets well at multiple scales. These findings suggest that MMAFNet is a promising method for land cover classification from remote sensing images.

b. REMSNet

Relation-Enhanced Multiscale Convolutional Network proposed by Liu Chun uses high-resolution aerial images for urban land cover classification using several techniques. The system architecture of this hybrid net is divided into 5 different groups:

- Dense Connectivity Pattern: The REMSNet utilizes a dense connectivity pattern, which allows for better feature propagation and reuse throughout the network. This design helps in capturing more complex features from the input images.
- Parallel Multi-Kernel Convolution: The model employs parallel multi-kernel convolution, enabling it to process features at different scales simultaneously. This is particularly useful for handling multiscale objects commonly found in urban environments.
- Spatial Relation-Enhanced Block: This component of the model is designed to learn global contextual relations between any two positions in the feature maps. By enhancing feature representations, it helps the model understand the spatial relationships between different land cover types.
- Channel Relation-Enhanced Block: Similar to the spatial block, this block focuses on learning relationships between different feature maps, further improving the model's ability to classify land cover accurately.
- Parallel Multi-Kernel Deconvolution Module: This module aggregates information from different scales, allowing the model to effectively reconstruct and classify features at various resolutions.

This model was tested on real-world datasets that are in the area of Shanghai covering an area of approximately 143 km² which achieved an overall accuracy of 88.55% and a mean intersection-over-union of 0.7394 for urban land cover classification such as buildings, roads, vegetation, and water bodies and identifying these different land cover types in densely populated regions.

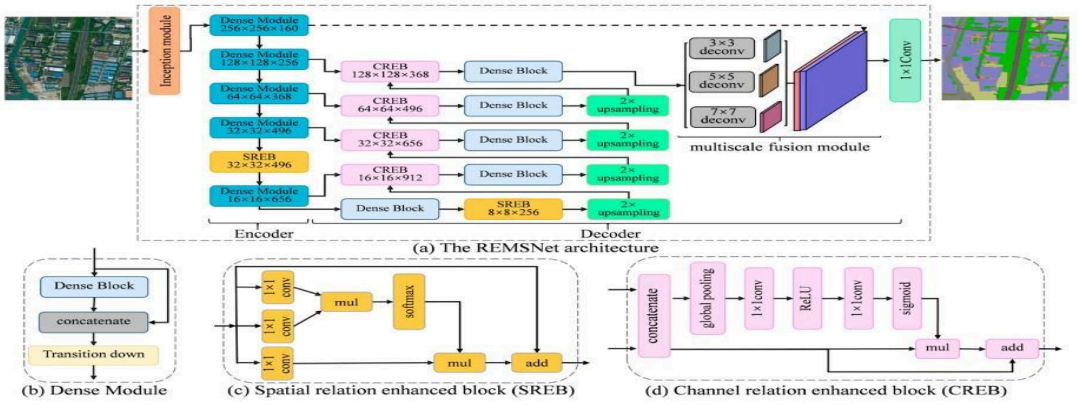

Fig.4: REMSNet Architecture

3. APPLICATIONS IN URBAN PLANNING

A. Land Use And Land Cover Mapping

Resource management in urban planning is a crucial topic, involving the analysis of how different areas—like residential, commercial, and industrial zones—allocate key resources such as land, water, energy, and materials. The environment is one of the most impacted factors when it comes to land use, making it important to assess how new developments might affect ecosystems, biodiversity, and overall environmental quality Setting proper guidelines, or policy-making, provides a clear set of rules to ensure that land use plans are followed, making the process more organized. Identifying areas that are prone to risks—whether from natural disasters or industrial accidents—helps in understanding land use patterns and in taking steps to mitigate these risks. Evaluating how land is used also gives us insights into economic development. Collectively, these elements contribute to the broader concept of land use and land cover mapping, which plays a vital role in urban planning. The figure beside is an urban map export featuring natural and man-made features for understanding the composition of different land usages.

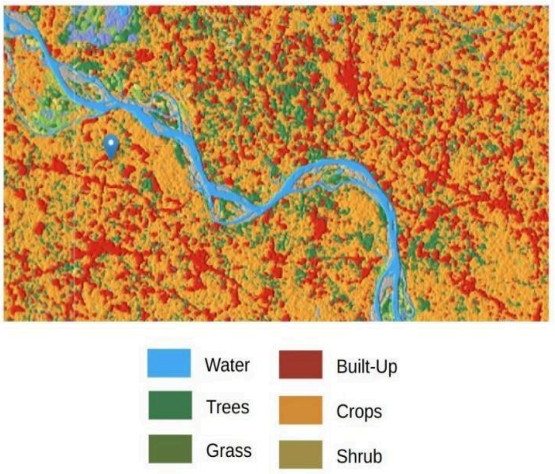

Fig.5: Land Cover Application showing land coverage with water, trees, and buildings for urban planning and policy-making.

B. Environmental Impact Assessment

Most cities face the growing challenge of having to effectively govern, plan, develop infrastructure and support their rapidly growing population while dealing with the impacts of the triple planetary crisis of climate change, biodiversity loss and pollution [UNEP]. As a result of development pressures on green fields, urban green areas become small, scattered and polluted. On the other hand, the degradation and depletion of urban landscapes threaten health related quality of life of the population. Due to the causes of these environmental problems, it is necessary to revise the current urban policies and develop new planning models for sustainable urban development [Dr. Amira Mersal, 2016] [24].

The rapid growth of urban space and its environmental challenges require precise mapping techniques to represent complex earth surface features more accurately. In the study [Rajesh Bahadur Thapa and Yuji Murayama, 2009] [25], four mapping approaches (unsupervised, supervised, fuzzy supervised, and GIS post-processing) were examined using Advanced Land Observing Satellite images to predict urban land use and land cover of Tsukuba city in Japan. Intensive fieldwork was conducted to collect ground truth data. A random stratified sampling method was chosen to generate geographic reference data for each map to assess the accuracy. The accuracies of the maps were measured, producing error matrices and Kappa indices. The GIS post-processing approach proposed in this research improved the mapping results, showing the highest overall accuracy of 89.33% compared to other approaches. The fuzzy supervised approach yielded a better accuracy (87.67%) than the supervised and unsupervised approaches and effectively dealt with the heterogeneous surface features in residential areas.

Earth observation and remote sensing are fundamental for classifying land use, enabling the collection of terrestrial and atmospheric data without being on-site. Sentinel-2 offers a broad spectrum of data, capturing both visible and non-visible light, and allowing the identification of key land features like vegetation, water bodies, and urban areas, which is crucial for the classification task. Scene classification was developed to distinguish between cloudy pixels, clear pixels, and water pixels of Sentinel-2 data and is a result of ESA's Scene classification algorithm. Twelve different classifications are provided, including classes of clouds, vegetation, soils/desert, water, and snow. While it does not constitute a land cover classification map in a strict sense, it helps us understand the steps we take to confirm that the area we are classifying is indeed what we hypothesize it to be. The closer we are to ground truth, the more accurate the classification will be.

Various machine learning classifiers such as Decision Tree, Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) have been employed for urban area classification. Additionally, combining advanced statistical methods with LiDAR metrics derived from multispectral LiDAR data can improve land cover classification accuracy in urban areas. Textural features such as correlation, homogeneity, energy, and contrast can be used to train the classification model in grey-scale images, with techniques like Support Vector Machine (SVM) and Naive Bayes yielding good results. Moreover, supervised classification can also make use of Volunteered Geographic Information (VGI) projects like OpenStreetMap (OSM) to automatically collect training samples for built-up area classification, providing performances similar to manual approaches.

The methodologies discussed above individually convey the cruciality of image classification in Environmental Impact Assessment with its precise, accurate and data-driven approach to the analysis of complex environmental features. Therefore, conducting accurate environmental assessment is crucial not just for urban growth but for the surrounding as well. These 'assessments help in recognizing how developments impact the singular and overall environmental health.

5. CHALLENGES AND FUTURE DIRECTIONS

A. Challenges

1. Data Limitations and Weather Challenges: CV has its own challenges, such as data limitations and constraints of information and weather conditions like rainy, foggy, and cloudy create challenges in collecting aerial data.
2. Outdated GIS and Manual Planning Methods: Traditional Geographic Information Systems (GIS) and manual planning methods are less effective in modern urban planning.
3. Requirement of High Computational Resources: The number of parameters and computation of advanced neural models like BGC-Net (79.73M and 29.46G Mac), SegNet (16.31 M and 23.77GMac), and U-Net (13.4M and 23.77 GMac) exceeds the number of parameters so it requires more complex GPU and computational resources.
4. Low Computational Efficiency and Complexity: Convolutional Neural Networks (CNNs), and Graph Convolutional Neural Networks (GNNs) often struggle, especially when applied to larger datasets. For instance, BGC-Net trains an epoch on the WHU dataset and the CHN dataset is 256s and 294s and the complexity of the model makes it less efficient.
5. Dependency on high-resolution data and external factor: Poor-quality images or those with significant noise adversely affected data samples which directly affected classification accuracy limiting the model's effectiveness.
6. Imbalance Datasets: Relatively small dataset size and limited remote sensing data which affected the accuracy of the study.

B. Mitigation Methods Used In Papers:

1. Occlusion Handling Strategies:

The blocked or covered part of an image, typically by an object in the foreground is referred to as an occlusion. It prevents the full view of the image hampering it from proper analysis. In terms of aerial imagery, occlusions might look like a building, tree, or other structures standing in the way between a viewer and the land or any particular feature behind them.

There are multiple strategies developed to overcome this challenge.

a. Multi-view: Through the integration of multiple viewpoints, i.e., capturing the same area but from different angles helps rebuild the occluded portion of the image.
b. Deep learning: By understanding and analyzing the context of the image, models can be trained to recognize an image despite the presence of occlusions.

c. Multi-Scale and Multi-Modality Fusion**:** Multi-modality fusion implements different types of images or sensors allowing it to see what's hidden.
d. Classification Refinement: On the basis of what is already known about the scenario of the surrounding area, post-classification refinement adjusts the outcome after the initial analysis.
e. Relation-Enhanced Networks: Each part that makes up an image is in some way related to one another. Relation-enhanced networks focus on this factor to help themselves understand what is blocked by the occlusion.

C. Future Directions

In future work, high-resolution datasets could be employed to analyze various urban parameters, such as streets, sidewalk presence, roof geometry, vehicles, and tree canopies. The use of semi-supervised learning techniques should be explored to reduce the data cost associated with deep learning models. Additionally, the proposed method could be applied to monitor dynamic changes in ground objects within a given area, providing valuable insights into land use patterns and supporting government departments in macro-level decision-making. Future research may also focus on integrating REMSNet with technologies like Geographic Information Systems (GIS) and real-time data analytics to enhance analysis. Further refinement of climate models, particularly in assessing how variations in urban form impact climate responses, could provide new insights. Moreover, expanding feature spaces to include geometric features, hyperspectral imagery, and the detection of small or complex objects (e.g., houses under construction) would enhance the depth and applicability of future studies.

6. COMPARATIVE ANALYSIS

High-resolution images, especially in the range of 0.1–0.2 meters per pixel, consistently outperform low-resolution images. For example, Lempesis (2023) used 0.15 m aerial resolution for urban issue detection and achieved an F-score of 70.72%; Lei et al. (2021) reported a land cover classification with a data resolution of 5–9 cm and an overall accuracy of over 90%. However, the effectiveness of classification does not only depend on resolution; the selection of the classification algorithm and feature extraction method plays an important role in making the most of the available data. Classical techniques such as having a perfect and random forest in aerial image classification. Multi-Modal Multi-Scale Attention Fusion Network (MMAFNet) proposed by Lei achieved an average F1 score of 92.34%, which is better than traditional methods. Furthermore, advanced removal techniques including multiple removal and attention techniques show a significant improvement in capturing relevant information from images. Relationship enhanced multiple contact network (REMSNet) proposed by Liu reported this model, which achieved 90.46% and 88.55% overall accuracy on Vaihingen and Shanghai datasets, respectively. Solving problems such as building and tree occultation is still an important focus in urban aerial imagery delivery. Strategies to mitigate these issues include the use of multiple lifting and tracking techniques as demonstrated in the REMSNet model. Furthermore, data enhancement techniques such as rotation, scaling, and translation have proven effective in improving the power and detail of the model. The integration of various data, including spectral data, LiDAR data, and auxiliary data such as points of interest

(POIs), holds promise for improving the distribution of people and solving urban problems. Future research in this area includes using urban morphology data to improve urban climate models, investigating the impact of urban heterogeneity on urban climate responses, and extending it to global data to gain a better understanding of urban forms.

A. Comparison And Analysis Of Results At High-Resolution Vs Low-Resolution Images

The resolution of images is crucial in aerial imagery classification, significantly impacting accuracy and detail. Higher-resolution images typically offer finer spatial information, leading to better classification results, while lower-resolution images often suffer from reduced performance due to the loss of critical details.

Standard Resolution Type Ranges

- Low Resolution: Below 256x256 pixels or greater than 30 meters per pixel
- Medium Resolution: 256x256 to 1024x1024 pixels or 10 to 30 meters per pixel
- High Resolution: 1024x1024 to 10000x10000 pixels or 1 to 10 meters per pixel
- Very High Resolution: Below 10 cm per pixel

1. High-Resolution Images

High-resolution images, ranging from 1 meter to 10 cm per pixel, are especially beneficial for tasks requiring detailed classification. For example:

a. Xiaoling Xie et al. (2022) used GF-2 satellite images with a resolution of 6800x7200 pixels. Their U-net model, enhanced with batch normalization and SeLU activation functions, achieved a precision of 94.3%, recall of 91.5%, and overall accuracy of 94.1%, demonstrating the advantages of high-resolution data in urban and rural planning.
b. Chun Liu et al. (2020) demonstrated the effectiveness of high-resolution images (0.1 to 1 meter per pixel) using the ISPRS Vaihingen and Shanghai datasets. Their Relation-Enhanced Multiscale Convolutional Network (REMSNet) achieved overall accuracies of 90.46% for Vaihingen and 88.55% for Shanghai, emphasizing the role of high-resolution imagery in capturing detailed urban features.
c. Tao Lei et al. (2021) used very high-resolution imagery (as fine as 5 cm per pixel) from the Vienna and Potsdam datasets. Their multi-modality and multi-scale attention fusion network achieved a mean F1-score of 92.34% and overall accuracy of 91.04%, showcasing the exceptional performance of very high-resolution data combined with advanced models.
d. Jinlin Jia et al. (2022) worked with UAV images and DSM data with resolutions of 0.057 m and 0.114 m, respectively. Their Random Forest classifier achieved the highest kappa coefficient of 0.807, demonstrating the value of very high-resolution
imagery for urban infrastructure classification.
e. Yasmine Megahed et al. (2021) fused airborne LiDAR point clouds with aerial images (40 cm resolution) for urban mapping. Their multi-layer perceptron model

achieved an overall accuracy of 97.75%, illustrating the benefits of high-resolution data for heterogeneous land-use mapping.

2. Low-Resolution Images

Low-resolution images (below 256x256 pixels or greater than 30 meters per pixel) often present challenges due to reduced spatial detail:

a. Antonio Rangel et al. (2024) worked with the EuroSAT dataset, which had a resolution of 64x64 pixels. Advanced models like MaxViT achieved a top-1 accuracy of 0.973, but models such as AlexNet (0.837), ResNet (0.835), and ResNeXt (0.843) demonstrated lower performance, reflecting the limitations of low-resolution images.

b. Jingwen Yuan et al. (2022) utilized hyperspectral imagery from the GaoFen-5 satellite with a resolution of 30 meters per pixel. Their Spectral-Spatial Unified Network combined with Conditional Random Fields (SSUN-CRF) achieved an overall accuracy of 93.86% and a Kappa coefficient of 92.08%, but the performance could likely be improved with higher-resolution imagery that would provide more detailed spatial information for better class differentiation.

3. Medium-Resolution Images

a. Wenzhuo Zhang et al. (2022) combined deep fully convolutional networks with graph convolutional neural networks for building extraction from aerial images. Using medium-resolution images (WHU dataset: 512x512 pixels; CHN dataset: 500x500 pixels), their BGC-Net achieved an overall accuracy of 97.6% on the WHU dataset, highlighting the effectiveness of medium-resolution images combined with advanced models.

Summary of High-Resolution vs Low-Resolution:

High-resolution images generally deliver superior classification performance due to their ability to capture finer spatial details, which are crucial for distinguishing various land cover types. In contrast,

low-resolution images, while still useful, often lead to reduced accuracy and precision. These reviewed studies confirm that although advanced models and techniques can mitigate some limitations of low-resolution data, high-resolution imagery remains preferable for tasks requiring high accuracy and detailed spatial analysis.

Table II: Resolution Comparison

Factors	High resolution	Low resolution
Performance	Makes the most out of detailed data allowing models to detect quality features in an image.	Might result in loss of critical information resulting in a negative impact on the performance of a model.
	A more time-consuming process due to its requirement of computational power and resource- intensiveness.	Offers faster processing along with reduced computational requirements.
	Less scalable to implement in large areas because of the volume of data.	Preferable in situations where the data in general is of more importance than the details.
Accuracy	High accuracy caused by its tendency to identify and classify various land cover types and urban features.	Because of its inclination towards reduced detailing, loss of important features takes place making it more vulnerable to in accuracy.
	At times, increased accuracy might result in overfitting. When the model is too aligned with the details but shows negligence towards other data sets.	Generalization across different data sets can be found to be better done due to their emphasis on general data.
	Gaining high accuracy Hand-in-hand with efficient use of computational resources can be challenging.	A balance between speed and accuracy must be attentively managed.

B. Classical Vs Modern Deep Learning Methods:

Performance Comparison

1. Modern Techniques:

Modern deep learning methods consistently perform better than classical techniques across multiple evaluation metrics. In" Combining Deep Fully Convolutional Networks with GIS," BGC-net achieved an overall accuracy of 97.6%, along with higher F1 scores and Kappa statistics, highlighting its superior precision, recall, and label agreement.

Similarly," Multi-Modality and Multi-Scale Attention Fusion Networks for Urban Analysis" reported an overall accuracy of 91.04% and an F1-score of 92.34%, further demonstrating the robustness of these modern architectures.

2. Classical Techniques:
   While generally less effective than modern methods, classical models like SVM, MLP, and Bagging can still deliver reasonable results. For instance, in" Integrated Urban Land Cover Analysis Using Classical Methods," these techniques achieved an overall accuracy of 90.25% on smaller datasets, though they typically showed lower F1 scores and Kappa statistics compared to modern approaches.

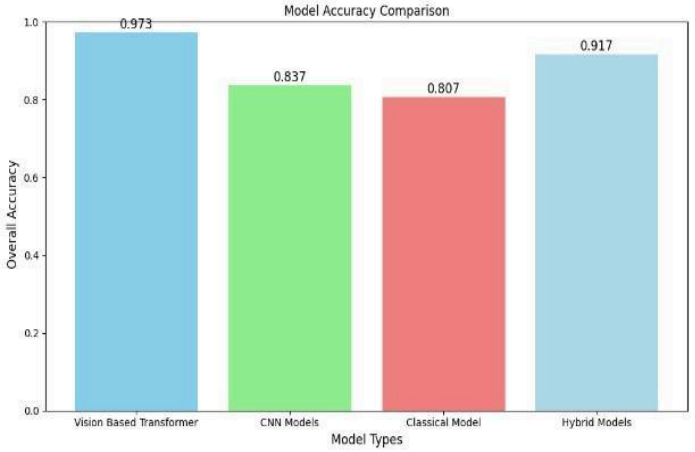

Fig.6: Overall Accuracy comparison between 4 types of models: Hybrid, CNN-based, Vision-based Transformer, Classical models

C. Discussion
   In aerial image classification, modern deep learning techniques offer significant advantages over classical approaches, particularly in terms of accuracy, precision, F1 score, and Kappa statistics. The trend across various metrics, including Overall Accuracy, F1 Score, and Kappa Statistics, indicates that modern deep learning techniques generally outperform classical methods in urban image classification tasks. In the above figure, Vision vision-based transformer model and hybrid model have accuracy greater than 90% whereas CNN and classical model are below 85%. Their ability to manage large datasets and complex patterns makes them more suitable for contemporary urban planning needs. However, classical methods may still be relevant in scenarios with limited computational resources or smaller datasets. Hence, the choice of technique should be context-dependent, considering factors such as computational resources and the nature of the data involved.

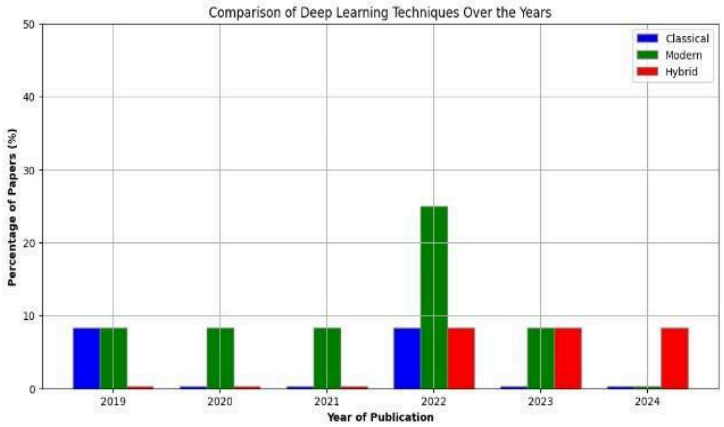

Fig.7: Comparison of different deep learning techniques used in the years from 2019 to 2024 A.D.

7. CONCLUSION

This review has systematically evaluated the advancements in aerial image classification techniques and their applications in urban planning. The integration of high-resolution imagery with modern deep learning approaches, such as Convolutional Neural Networks (CNNs) and Vision Transformers, has significantly enhanced classification accuracy and efficiency. These developments have facilitated more precise assessments of urban environments, including land use, infrastructure, and environmental changes.

The transition from classical to deep learning methods has led to notable improvements in classification performance. High-resolution imagery, particularly with resolutions as fine as 5 cm per pixel, consistently provides superior results compared to lower resolutions. Studies such as those by Lei et al. (2021) and Lempesis (2023) highlight the benefits of high-resolution data in achieving higher classification accuracy. Examination of case studies involving various satellite and drone imagery, including Sentinel-2 and GaoFen-2, demonstrates the effectiveness of advanced models across diverse urban settings. High-resolution and multi-modal data have proven essential for detailed urban planning.

High-resolution images typically offer superior classification performance by capturing finer spatial details, crucial for distinguishing land cover types. In contrast, low-resolution images often result in reduced accuracy due to insufficient detail.

Although advanced techniques can address some limitations of low-resolution data, high-resolution imagery remains preferred for tasks requiring detailed spatial analysis. While classical machine learning techniques, such as Support Vector Machines (SVMs) and Random Forests, retain value in certain contexts, deep learning methods, particularly CNNs, provide superior performance for complex, high-dimensional data. A hybrid model like MMAFNet has greater accuracy than other CNN models. The ability of these to automatically extract hierarchical features makes them the preferred choice for modern urban image classification applications.

Each imaging method satellite, drone, and ground camera offers unique advantages. Satellites are ideal for large-scale monitoring, drones provide high-resolution updates, and ground cameras offer detailed insights into specific features. The selection of the appropriate method depends on the context of the task and the level of detail required.

Despite significant progress, challenges remain, including issues related to image

resolution, data variability, and labelled dataset availability. Future research should focus on enhancing the integration of high-resolution data, exploring semi-supervised learning techniques, and improving the interoperability of classification models with Geographic Information Systems (GIS). Additionally, further investigation into the impacts of urban form on climate responses and the refinement of climate models will be crucial for advancing urban planning practices.

In summary, while considerable advancements have been made in aerial image classification for urban planning, continued research and technological innovation are essential. Future developments should aim to enhance model robustness, adaptability, and data acquisition methods to support more effective and precise urban management strategies.

REFERENCES

[1] R. P. Singh and J. Dhakal, "Problems and prospects of urbanization in Kathmandu valley," International Journal of Atharva, vol. 2, pp. 19–33, mar 2024.

[2] N. Lempesis, "Automatic mapping of physical urban problems using remotely sensed imagery," International Journal of E-Planning Research, vol. 12, pp. 1–21, apr 2023.

[3] European Commission, P. Rode, and P. Hoffman, "Report on urban development and sustainability," 2017. Accessed: 2024-09-06.

[4] T. Fyleris, A. Krisciunas, V. Gruzauskas, D. Calneryte, and R. Barauskas, "Urbanˇ change detection from aerial images using convolutional neural networks and transfer learning," ISPRS International Journal of Geo-Information, vol. 11, p. 246, apr 2022.

[5] W. Lin, D. Zhang, F. Liu, Y. Guo, S. Chen, T. Wu, and Q. Hou, "A lightweight multi-label classification method for urban green space in high-resolution remote sensing imagery," ISPRS International Journal of Geo-Information, vol. 13, p. 252, jul 2024.

[6] M. Rumpler, A. Tscharf, C. Mostegel, S. Daftry, C. Hoppe, R. Prettenthaler, F. Fraundorfer, G. Mayer, and H. Bischof, "Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance," Computer Vision and Image Understanding, vol. 157, pp. 255–273, 2017.

[7] A. Albert, J. Kaur, and M. C. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1357–1366, ACM, 2017.

[8] M. R. Ibrahim, J. Haworth, and T. Cheng, "Understanding cities with machine eyes: A review of deep computer vision in urban analytics," Cities, vol. 96, p. 102481, 2020.

[9] L. Liu and A. Sevtsuk, "Clarity or confusion: A review of computer vision street attributes in urban studies and planning," Cities, vol. 150, p. 105022, 2024.

[10] A. Janowski, M. Renigier-Bi lozor, M. Walacik, and A. Chmielewska, "Remote measurement of building usable floor area – algorithms fusion," Land Use Policy, vol. 100, p. 104938, 2021.

[11] L. Liu and A. Sevtsuk, "Clarity or confusion: A review of computer vision street attributes in urban studies and planning," Cities, vol. 150, p. 105022, 2024.

[12] S. C. K. Tekouabou, E. B. Diop, R. Azmi, R. Jaligot, and J. Chenal, "Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5943–5967, 2022.
[13] R. D. Majd, M. Momeni, and P. Moallem, "Transferable object-based framework based on deep convolutional neural networks for building extraction," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 8, pp. 2627–2635, 2019.
[14] S. Tehsin, S. Kausar, A. Jameel, M. Humayun, and D. K. Almofarreh, "Satellite image categorization using scalable deep learning," Applied Sciences, vol. 13, no. 8, p. 5108, 2023.
[15] Y. Zhang, J. Luttrell, and C. Zhang, "How to detect occluded crosswalks in overview images? comparing three methods in a heavily occluded area," International Journal of Transportation Science and Technology, 2024.
[16] J. Arndt and D. Lunga, "Large-scale classification of urban structural units from remote sensing imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 2634–2648, 2021.
[17] X. Xie, X. Kang, L. Yan, L. Zeng, and L. Ye, "Land use classification method of remote sensing images for urban and rural planning monitoring using deep learning," Scientific Programming, vol. 2022, pp. 1–9, 2022.
[18] T. Fyleris, A. Krisciunas, V. Gruzauskas, D. Calneryte, and R. Barauskas, "Urbanˇ change detection from aerial images using convolutional neural networks and transfer learning," ISPRS International Journal of Geo-Information, vol. 11, no. 4, p. 246, 2022.
[19] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.
[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, others, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.
[22] Z. Tu, H. Talebi, H. Zhang, T. Yang, X. Zhang, L. Zhang, others, and C. Choi, "Maxvit: Multi-axis vision transformer," arXiv preprint arXiv:2204.01697, 2022.
[23] T. Lei, L. Li, Z. Lv, M. Zhu, X. Du, and A. K. Nandi, "Multi-modality and multi-scale attention fusion network for land cover classification from vhr remote sensing images," Remote Sensing, vol. 13, no. 18, p. 3771, 2021.
[24] A. M. Dr, "Sustainable urban futures: Environmental planning for sustainable urban development," Procedia Environmental Sciences, vol. 34, pp. 49–61, 2016.
[25] R. B. Thapa and Y. Murayama, "Drivers of urban growth in the kathmandu valley, Nepal: Examining the efficacy of the analytic hierarchy process," Applied Geography, 2010.