Automated Dilated Convolutional Attention for Land Use Classification

Mr. Abhale B.A[1], Mr. Wadghule Y.M[2], Miss.Chaudhari Nisha[3], Miss.Pagar Shruti[4], Miss.Nikam Aarti[5], Miss.Darade Nikita[6]

[1,2,3,4,5,6]*S.N.D College of Engineering and Research Center, Savitribai Phule Pune University*

Abstract: Accurate and timely Land Use and Land Cover (LULC) classification is a cornerstone of environmental management, urban planning, and sustainable resource allocation. Traditional methods, reliant on manual digitization or classical machine learning, are inefficient, slow, and struggle with the scale and complexity of modern remote sensing data. While Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), has significantly advanced LULC mapping, a critical operational gap remains: data siloing. Most DL models are designed for a single modality (e.g., 3-channel RGB *or* 4-channel multispectral) and produce outputs tied to a specific dataset's legend. This survey paper explores the evolution of LULC classification, focusing on the challenge of multi-modal data ingestion and output harmonization. We review foundational architectures like U-Net and DeepLabV3+ and investigate the use of dilated convolutions and attention mechanisms for high-resolution semantic segmentation. The core focus is on a hybrid architecture that intelligently routes multi-modal inputs (3-channel vs. 4-channel) to specialized models, yet produces a single, unified, and harmonized classification map. This study concludes that such a "router-based" hybrid approach is essential for creating a truly robust, scalable, and user-friendly LULC system that can adapt to diverse, real-world data sources.

Keywords: Land Use Land Cover (LULC), Deep Learning, Semantic Segmentation, High-Resolution Imagery, MultiModal Data, Data Fusion, Data Harmonization, U-Net, DeepLabV3+, Dilated Convolutions, Attention Mechanism, Multispectral, RGB, Remote Sensing, Input Router.

I. INTRODUCTION

Land Use and Land Cover (LULC) mapping, the process of classifying the Earth's surface, is a critical task. From tracking deforestation (Forest) and managing water resources (Water) to planning urban expansion (Building) and monitoring food security (Agriculture), LULC data forms the basis of countless geopolitical and environmental decisions.

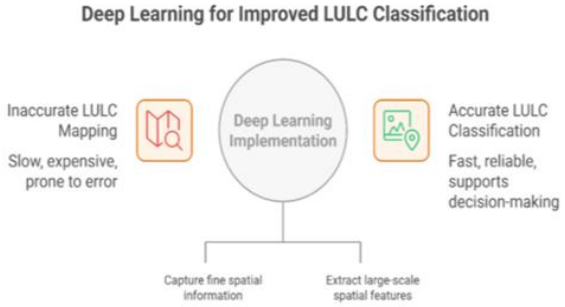

Fig. 1. Deep Learning for Improved LULC Classification.

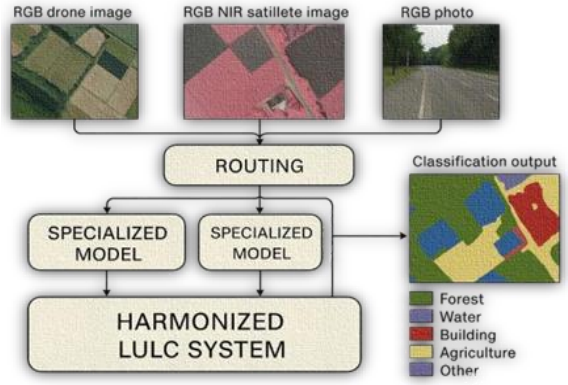

Fig. 2. Introduction to LULC

Historically, this process was manual, slow, and expensive. The advent of remote sensing and classical machine learning (e.g., Random Forest, SVM) offered partial automation but still required significant feature engineering and failed to capture the complex spatial relationships in high-resolution imagery.

Deep Learning, specifically Fully Convolutional Networks (FCNs), has revolutionized the field. However, this has led to a new problem: model- data

rigidity. A model trained on 3-channel RGB drone imagery cannot process a 4-channel (RGB+NIR) satellite image. Conversely, a multispectral model fails if the user provides a simple RGB photo. Furthermore, a model trained on the LoveDA dataset (7 classes) produces a different map than one trained on the Potsdam dataset (6 classes). This lack of flexibility makes "state-of the-art" models operationally useless for an end-user who has different types of data.

This paper surveys the key technologies required to build a truly automated, multi-modal, and harmonized LULC system. We explore the architectures that enable pixel-level accuracy (dilated convolutions, attention) and propose a novel framework that intelligently routes inputs to specialized models, ensuring a single, consistent classification output regardless of the input source.

II. PROBLEM DEFINITION

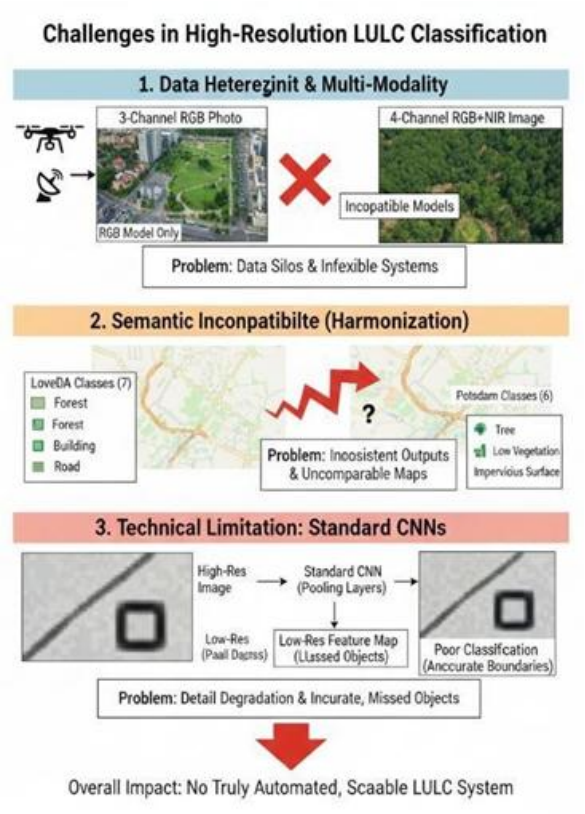

Fig. 3. Challenges in High-Resolution LULC Classification.

High-resolution Land Use and Land Cover (LULC) classification, a cornerstone of effective urban planning, environmental monitoring, and disaster response, faces three significant and interconnected challenges that hinder the development of a truly automated, scalable, and universally applicable system.

1. The Challenge of Data Heterogeneity and Multi-Modality

In the real world, LULC data is not standardized. It is captured by a wide array of sensors on drones, aircraft, and satellites, resulting in a heterogeneous mix of data modalities. An enduser may have 3-channel RGB aerial photographs for one region and 4-channel (RGB + NearInfrared) multispectral imagery for another. Standard Deep Learning models are rigidly architected; a model trained on 3-channel data cannot accept 4-channel input, and vice versa. This "data silo" problem forces organizations to develop and maintain a portfolio of separate, single-use models, a process that is inefficient, costly, and not scalable.

2. The Crisis of Semantic Incompatibility (Harmonization)

A more profound challenge is the lack of semantic interoperability between LULC datasets. Different mapping projects produce data with incompatible legends, terminologies, and class definitions. For example, one dataset (like LoveDA) may define a single Forest class, while another (like ISPRS Potsdam) may split vegetation into Tree and Low Vegetation. This semantic mismatch means that models trained on different data sources produce outputs that are fundamentally incompatible, limiting their comparability and making largescale analysis impossible.

3. The Technical Limitation of Standard CNNs for High-Resolution Data

At the model level, high-resolution imagery presents a paradox. To accurately classify a pixel (e.g., as Building), the model needs a wide contextual view to see its surroundings. However, to draw a precise boundary, it needs to maintain high spatial resolution. Standard Convolutional Neural Networks (CNNs) achieve wide context by using repeated pooling layers, which degrade spatial resolution and cause the loss of finegrained details. This "degradation problem" makes them perform poorly on small, complex

features like cars, thin roads, or building edges, which are critical in high-resolution data.

The Proposed Solution: A Harmonized Multi-Modal Framework

This project directly addresses these three challenges by proposing an intelligent, end-to-end framework that goes beyond a single-model approach.

1. To solve Data Heterogeneity, we will develop an Automated Input Router. This lightweight module will sit in front of the system, detect the input image's band count (3-channel vs. 4-channel), and dynamically deploy the correct, specialized model.
2. To solve Semantic Incompatibility, we will implement a Data Harmonization Pipeline. All models will be trained to output a single, unified classification legend (the LoveDA classes). For the multispectral model, this involves a novel preprocessing step that re-maps the Potsdam dataset's classes and uses the Normalized Difference Water Index (NDWI) to "autolabel" water bodies, creating a semantically consistent training dataset.
3. To solve the CNN Limitations, the core of our system will be an Automated Dilated Convolutional Attention Network (built on DeepLabV3+).
    - Dilated Convolutions will be used to capture multi- scale context without degrading resolution.
    - Attention Mechanisms will enable the model to learn and focus on the most salient features for a given class.

The final product will be a single, robust system that can accept diverse data inputs and automatically produce a single, harmonized, and highly accurate LULC map.

III. OBJECTIVES OF THE PROPOSED SYSTEM

Based on the identified gaps, a truly robust system must achieve four primary objectives:

1. Develop a High-Accuracy RGB Specialist Model: To train a model (Model A) on a 3channel RGB dataset (LoveDA) that is pre-trained on ImageNet to expertly classify features based on shape, color, and texture.
2. Develop a High-Accuracy Multispectral Specialist Model: To train a model (Model B) on a 4-channel RGB+IR dataset (Potsdam) that can leverage the NearInfrared band for superior separation of vegetation and water.
3. Achieve 100% Output Harmonization: To ensure *both* Model A and Model B output the *exact same* 7-class legend (from LoveDA). This involves a novel preprocessing step to "patch" the Potsdam labels, using NDWI to auto-generate Water labels.
4. Implement an Automated Input Router: To build a "brain" that sits in front of the models. This router will automatically detect the input image's channel count (3 vs. 4) and route it to the correct, specialized model for inference, providing a seamless user experience.

IV. LITERATURE SURVEY

The evolution of LULC classification has been rapid, moving from manual analysis to classical machine learning and, most recently, to deep learning. Classical methods like Random Forest (RF) and Support Vector Machines (SVM) were popular but relied on "hand-crafted features" and often failed to capture complex spatial context, producing "noisy" or "salt-and-pepper" classification maps.

The breakthrough came with deep learning, specifically semantic segmentation models. U-Net (Ronneberger et al., 2015) introduced the encoderdecoder architecture with "skip connections," which proved exceptionally effective at preserving fine-grained details and boundaries— critical for high-resolution LULC. DeepLabV3+ (Chen et al., 2018) addressed a different problem: context. By using atrous (dilated) convolutions, it could "see" a wider area without losing resolution, improving its ability to classify large, contiguous objects.

Further refinements, such as the Attention U-Net (Oktay et al., 2018), added attention gates, allowing the model to learn where to focus and suppress irrelevant features.

However, as the table below shows, most research focuses on a single dataset or single modality. The critical research gap, which this project addresses, is the fusion and harmonization of these separate efforts.

Sr. no.	Paper Title	Author(s)	Year	Key Contribution
1	U-Net: Convolutional Networks for Biomedical Image Segmentation	Ronneberger, O., Fischer, P., & Brox, T.	2015	The foundational encoder- decoder architecture with "skip connections" for precise segmentation.
2	Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation	Chen, L.C., Zhu, Y., Papandreou, G., et al.	2018	Introduced the DeepLabV3+ model with Atrous Spatial Pyramid Pooling (ASPP), the gold standard for dilated convolution.
3	Attention U-Net: Learning Where to Look for the Pancreas	Oktay, O., Schlemper, J., et al.	2018	Integrated attention gates into the U-Net skip connections, allowing the model to focus on the most salient features.
4	LoveDA: A Deep Learning Dataset for Land-Use and Land-Cover...	Wang, J., Zheng, Z., Ma, A., et al.	2021	Provided a large-scale, high- resolution 3-channel (RGB) dataset with 7 LULC classes, ideal for training Model A.
5	ISPRS 2D Semantic Labeling Challenge - Potsdam	ISPRS	2018	Provided a benchmark 4- channel (RGB+IR) high- resolution dataset, critical for training Model B (our multispectral specialist).
6	Land Use and Land Cover Classification Meets Deep Learning: A Review	Zhao, S., Tu, K., Ye, S., et al.	2023	A comprehensive review confirming that CNN-based models (U-Net, DeepLab) are the state-of-the-art for LULC.
7	A Survey of Multimodal Data Fusion in Earth Observation-Remote Sensing	Yuan, K., Zhu, Z., Pang, Y., et al.	2025	A recent survey on fusing heterogeneous data (like optical, radar, LiDAR), confirming the importance of multi-modal approaches.
8	Lulc Image Classification with Convolutional Neural Network	Balarabe, A. T., & Jordanov, I.	2021	Directly addresses class heterogeneity by grouping 21 classes into 4 "superclasses"[8888], supporting our harmonization goal.
9	LULC Change Detection Using Combined Machine and Deep Learning Classifiers	Tahraoui, A., & Kheddam, R.	2024	Proves the value of combining classifiers (RF and DNN) 9999and fusing their decisions10, which is conceptually similar to our "Router" system.
10	Image harmonization: A review of statistical and deep learning methods...	Hu, F., Chen, A. A., Shinohara, R. T., et al.	2023	A key review on data harmonization methods to remove "batch effects" (e.g., sensor differences), which is the core problem our router solves.

V. PROPOSED SYSTEM

System Workflow: The entire process is managed by an "Input Router." When a user uploads a high-resolution GeoTIFF, the router (using a library like rasterio) instantly checks the file's metadata to count the bands.

- Path A (3-Channels): If the router detects a 3-channel (RGB) image, it passes the data to Model A.
  - Model A: A DeepLabV3+ model (with SCSE attention) pre-trained on ImageNet and fine-tuned on the 3-channel LoveDA dataset.
  - Output: A 7-class harmonized LULC map.
- Path B (4-Channels): If the router detects a 4-channel (RGB+IR) image, it passes the data to Model B.
  - Model B: A DeepLabV3+ model (with SCSE attention) trained *from scratch* (as ImageNet weights are 3- channel) on our novel, harmonized Potsdam dataset.
  - Output: A 7-class harmonized LULC map.

The user is unaware of this complexity; they simply upload any supported image and receive a consistent, standardized LULC map.

Data Harmonization Workflow (Model B): The key innovation is the creation of the training data forModel B. We cannot simply train on the Potsdam labels, as they are inconsistent with LoveDA. Our system uses a two-step harmonization process:

1. Class Mapping: We create a "translation key" that maps Potsdam classes to LoveDA classes (e.g., Potsdam's Impervious Surface becomes LoveDA's Road; Tree becomes Forest, etc.).
2. Water Patching (NDWI): To solve the missing Water class in Potsdam, we leverage the 4th (NIR) channel. We calculate the Normalized Difference Water Index (NDWI = (Green - NIR) / (Green + NIR)) for the entire image. Any pixel that Potsdam labeled as Clutter but has a high NDWI score is automatically re-labeled as Water. This teaches

Model B to associate the (low-NIR) signature of water with the correct class.

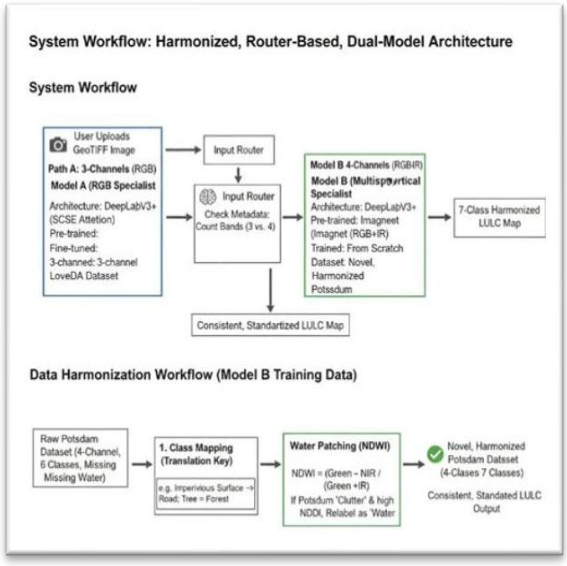

Fig. 4. System Workflow.

VI. ARCHITECTURE

The system is designed with a modern, 3-tier architecture, separating presentation, logic, and inference.

1. User Layer (Presentation): A web-based frontend (built with Gradio) that provides a simple UI for a user to upload their .tif or .png file. This layer is responsible for displaying the final, colorcoded, and annotated LULC map.
2. Application Layer (Logic): The "brain" of the project, written in Python. This layer contains:
   - The Input Router: Code that uses rasterio to check the band count of the uploaded file.
   - Preprocessing Module: Scripts that handle image "tiling" or "patching" (cutting large images into 256x256 tiles), normalization, and data augmentation.
   - Post-processing Module: A script that "stitches" the predicted tiles back together to form a seamless, fullresolution output map.
3. Model Layer (Inference): The core deep learning engine. This layer holds the saved model weights (model_A_rgb.pth and model_B_multispectral.pth) and uses PyTorch and segmentation-models-pytorch (SMP) to load the correct model and run the model.eval() inference on the GPU.

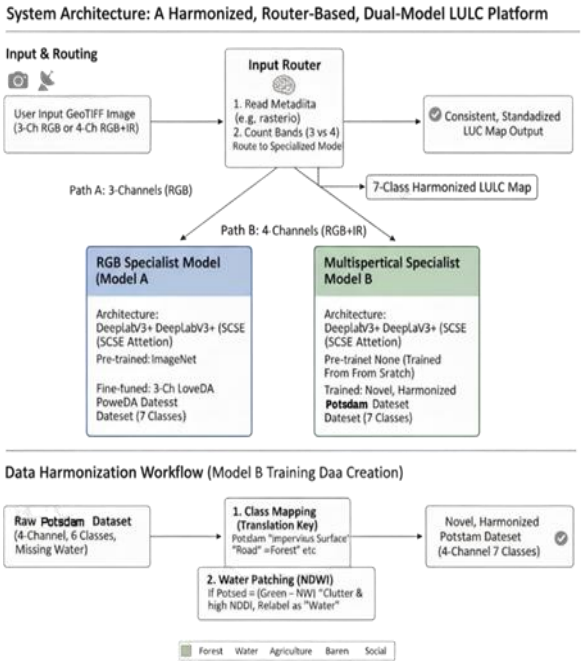

Fig. 5. System Architecture.

VII. WORKING (PROCESS)

The system's development is divided into two distinct, parallel training phases: one for the 3channel RGB specialist and one for the 4-channel multispectral specialist.

1. Model A (RGB) Training

Model A is trained to be the "RGB Specialist." This model, a DeepLabV3+ with an SCSE attention mechanism, uses a ResNet34 encoder pre-trained on the ImageNet dataset (encoder_weights="imagenet"). This pre-training is critical as it provides a strong foundation for recognizing shapes, colors, and textures. The model's input layer is configured for 3-channels (in_channels=3). It is then trained on the 3channel RGB images and 7-class segmentation masks from the LoveDA dataset. After training, the final weights are saved as model_A_rgb.pth.

2. Model B (Multispectral) Training & Harmonization

Model B is the "Multispectral Specialist," built on the identical DeepLabV3+ and SCSE attention architecture. However, its input layer is configured for 4-channels (in_channels=4) to accept RGB plus Near-Infrared (NIR) data.

Because no pre-trained 4-channel weights exist, the

encoder must be trained from scratch (encoder_weights=None). This model's training data is created through our novel Data Harmonization Pipeline. We take the 4-channel images from the ISPRS Potsdam dataset and its corresponding 6-class labels. We then apply two transformations: first, we re-map Potsdam's 6 classes to our 7-class unified LoveDA legend. Second, we auto-generate the missing Water class by calculating the NDWI from the Green and NIR bands. Any pixel that was Background but had a high NDWI score is "patched" and re-labeled as Water. The model is then trained on this new, harmonized 4-channel dataset and its weights are saved as model_B_multispectral.pth.

VIII. CONCLUSION

This paper has surveyed the landscape of deep learning for LULC classification, identifying a critical gap in operational flexibility. We have proposed a robust, dual-model, router-based architecture that moves beyond single-data-source limitations.

By training two specialized models—one for 3channel RGB data and another for 4-channel multispectral data—and using an intelligent Input Router to select the correct model, the system can serve a much wider range of users and data types. The key innovation presented is the data harmonization pipeline for the multispectral model, which uses a "translation key" and a "water patching" technique based on the NDWI to ensure both models output a single, consistent, 7class LULC map.

This framework creates a system that is not only accurate at the pixel level (thanks to dilated convolutions and attention) but is also scalable, flexible, and truly automated, bridging the gap between academic research and a real-world operational tool. Future work could involve expanding the router to include more models, such as for hyperspectral or SAR data, further enhancing its capabilities as a universal LULC classification platform.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U- Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image

[3] Segmentation," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[4] O. Oktay, J. Schlemper, L. L. Folle, et al., "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A Deep Learning Dataset for Land-Use and Land-Cover Classification," *arXiv preprint arXiv:2110.08733*, 2021.

[6] ISPRS, "ISPRS 2D Semantic Labeling Challenge - Potsdam," 2018. [Online]. Available: https://www.isprs.org/education/benchmarks/ Urba nSemLab/2d-sem-label-potsdam.aspx

[7] S. Zhao, K. Tu, S. Ye, et al., "Land Use and Land Cover Classification Meets Deep Learning: A Review," *Remote Sensing*, vol. 15, no. 1, p. 236, 2023.

[8] K. Yuan, Z. Zhu, Y. Pang, et al., "A Survey of Multimodal Data Fusion in Earth Observation-Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-28, 2025.

[9] A. T. Balarabe and I. Jordanov, "LULC Image Classification with Convolutional Neural Network," in Proc. 2021 IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2021, pp. 59855988.

[10] A. Tahraoui and R. Kheddam, "LULC Change Detection Using Combined Machine and Deep

[11] Learning Classifiers," in Proc. 2024 IEEE 7th Int. Conf. Advanced Technologies, Signal and Image Processing (ATSIP), 2024, pp. 1-6.

[12] F. Hu, A. A. Chen, R. T. Shinohara, et al., "Image harmonization: A review of statistical and deep learning methods for removing batch effects and site effects," *NeuroImage*, vol. 280, p. 120355, 2023.