Road Extraction from Remote Sensing Images Using a Skip-Connected Parallel CNN-Transformer Encoder-Decoder Model

Vuyyuru Dharani Sri Dept of ECE IARE

Dr. S China Venkateshwarlu Professor Dept of ECE IARE

Dr. V Siva Nagaraju Professor Dept of ECE IARE

---***---

Abstract - Extracting roads from remote sensing images holds significant practical value across fields like urban planning, traffic management, and disaster monitoring. Current Convolutional Neural Network (CNN) methods, praised for their robust local feature learning enabled by inductive biases, deliver impressive results. However, they face challenges in capturing global context and accurately extracting the linear features of roads due to their localized receptive fields. To address these shortcomings of traditional methods, this paper proposes a novel parallel encoder architecture that integrates a CNN Encoder Module (CEM) with a Transformer Encoder Module (TEM). The integration combines the CEM's strength in local feature extraction with the TEM's ability to incorporate global context, achieving complementary advantages and overcoming limitations of both Transformers and CNNs. Furthermore, the architecture also includes a Linear Convolution Module (LCM), which uses linear convolutions tailored to the shape and distribution of roads. By capturing image features in four specific directions, the LCM significantly improves the model's ability to detect and represent global and linear road features. Experimental results demonstrate that our proposed method achieves substantial improvements on the German-Street Dataset and the Massachusetts Roads Dataset, increasing the Intersection over Union (IoU) of road class by at least 3% and the overall F1 score by at least 2%.

***Key Words*:** Parallel Encoder Architecture , Linear Convolution Module (LCM) , Image Processing , Traffic Management.

1.INTRODUCTION

Road extraction from remote sensing images is a critical task in urban planning, autonomous navigation, and disaster management. Traditional methods rely on handcrafted features, while deep learning approaches, particularly convolutional neural networks (CNNs) and transformers, have significantly improved road segmentation accuracy. The combination of CNNs for feature extraction and transformers for capturing long-range dependencies has gained prominence in recent research. Road extraction from remote sensing images is a critical task in the fields of geographic information systems (GIS), urban planning, autonomous vehicle navigation, and disaster response. Traditional methods for road extraction have struggled to generalize well in complex environments due to occlusions, varying lighting conditions, and heterogeneous textures. With the advent of deep learning, convolutional neural networks (CNNs) have significantly improved feature extraction from high-resolution satellite images, especially for local patterns. However, CNNs alone often fail to capture long-range dependencies and global context.

To address these challenges, this study proposes a **Skip-Connected Parallel CNN-Transformer Encoder-Decoder architecture**. The model leverages the local feature extraction strength of CNNs and the global context modelling capability of vision Transformers (ViTs). By integrating both in a parallel encoder structure and connecting encoder and decoder with skip connections, the model aims to enhance spatial precision and semantic understanding in road segmentation.

2. Body of Paper

Accurate extraction of roads from remote sensing images is a critical task for applications such as urban development, transportation planning, and emergency response. Traditional methods, particularly those based on Convolutional Neural Networks (CNNs), have shown remarkable performance in extracting local features due to their strong inductive biases. However, CNNs often suffer from a limited ability to capture long-range dependencies and global context, which are essential for correctly identifying linear and continuous road structures. To overcome these challenges, this paper proposes a novel parallel encoder architecture that integrates both CNN and Transformer modules, complemented by a Linear Convolution Module designed specifically for road geometry.

CNN Encoder Module (CEM)
Captures rich local spatial features using a standard convolutional backbone (e.g., ResNet, U-Net encoder).

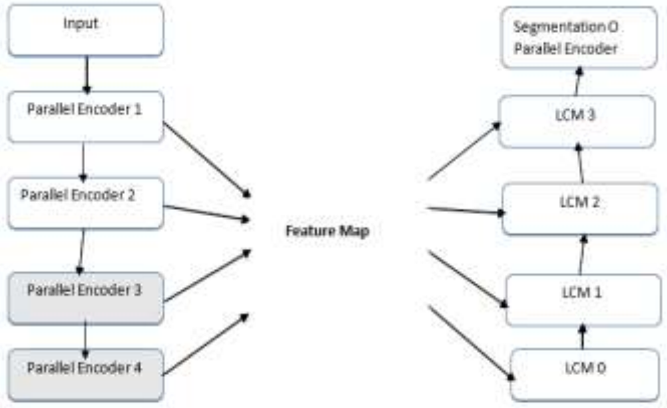
Transformer Encoder Module (TEM)
Processes the feature maps globally to capture long-range dependencies. Multi-head self-attention layers are used to model the entire image context.

Linear Convolution Module (LCM)
Enhances the model's sensitivity to road-like structures by applying linear convolutional kernels oriented in four directions (horizontal, vertical, and diagonals). This facilitates robust extraction of elongated and narrow road features.

Table -1:

Year/Author	Algorithms	Methodology	Merits	Remarks/Problem
Zhou et al. (2021)	CNN-based U-Net	Improved feature selection using attention mechanisms	Enhanced segmentation accuracy	Limited to specific datasets
Wang et al. (2022)	Transformer-based segmentation	Utilized global context modelling for road detection	Improved handling of complex road networks	High computational requirements
Liu et al. (2023)	Hybrid CNN-Transformer	Combined hierarchical feature extraction with global attention mechanisms	Outperformed standalone CNNs and transformers	Requires large labelled datasets

Existing Block Diagram



Proposed Block Diagram

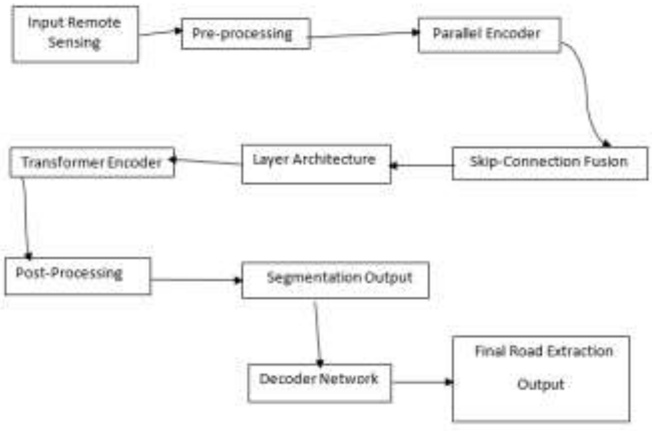


Fig -1: Figure

Road Segmentation: A Theoretical Perspective
The task of road extraction from remote sensing imagery can be theoretically framed as a **semantic segmentation problem** with additional structural and geometric constraints. Roads are typically narrow, elongated, and continuous structures that span large portions of an image, posing unique challenges not easily addressed by conventional segmentation networks. The proposed approach draws on the strengths of both **local feature learning** and **global context modeling**, underpinned by foundational theories in deep learning and computer vision.

1 Locality vs Globality in Feature Learning

Convolutional Neural Networks (CNNs) are inherently localized in their operations. The **receptive field** of a convolutional kernel defines how much of the image the model "sees" at a time. While stacking multiple layers increases the effective receptive field, it does not guarantee efficient or coherent global information integration, particularly for spatially distant but semantically connected pixels (e.g., ends of a winding road).

In contrast, the **self-attention mechanism** used in Transformers provides a theoretical advantage in modeling **global dependencies**. By computing attention weights between all pairs of pixels (or patches), Transformers are capable of capturing relationships regardless of spatial distance. However, they may underperform on fine-grained features due to the lack of strong inductive biases, such as translation equi-variance and locality, which are inherent in CNNs. Our architecture reconciles these two perspectives by **parallelizing CNN and Transformer modules**, allowing the model to learn both localized textures and global arrangements simultaneously. This hybrid design reflects the theoretical necessity of **multi-scale feature integration** in structured object segmentation.

2 Directional Convolution and Linear Feature Modeling

Roads exhibit strong **geometric regularity**, often aligning along specific orientations. Standard isotropic convolutions are limited in capturing this anisotropic information. The **Linear Convolution Module (LCM)** introduces a theoretical extension to convolutional operations by employing **directionally biased filters**. These filters are structured to emphasize responses along predefined angles—horizontal, vertical, and diagonal—thereby aligning the model's inductive biases with the road's geometric priors.The LCM can be viewed as a specialized variant of the Gabor filter approach used in classical image processing, reformulated in a deep learning context. It enhances the **structure-aware learning** capability of the network by enabling **orientation-selective filtering**, a concept supported by early findings in biological vision and adopted in computer vision literature.

3 Feature Fusion as Information Integration

The fusion of outputs from the CNN Encoder Module (CEM), Transformer Encoder Module (TEM), and LCM is grounded in the theory of **multi-source feature integration**. By combining feature maps with different characteristics—local, global, and directional—the model approximates a more holistic representation of the input. This aligns with the theoretical understanding that **ensemble representations** can improve generalization and robustness, particularly in tasks involving complex spatial patterns like road networks.

4 Optimization Landscape and Loss Design

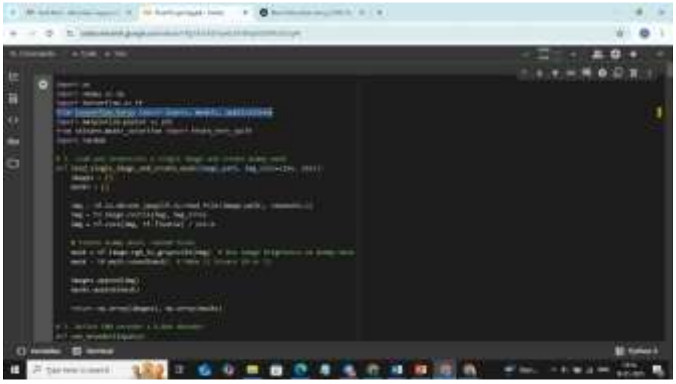
From a theoretical optimization standpoint, segmentation of roads—which often occupy a small portion of the image—leads to class imbalance. To address this, we use a **compound loss function** combining Binary Cross-Entropy (which ensures pixel-level classification) and Dice Loss (which emphasizes overlap and connectivity). This hybrid loss design helps guide the optimization process toward solutions that are not only accurate but also **topologically coherent**, a critical property in road mapping applications.

3.SYSTEM ARCHITECTURE

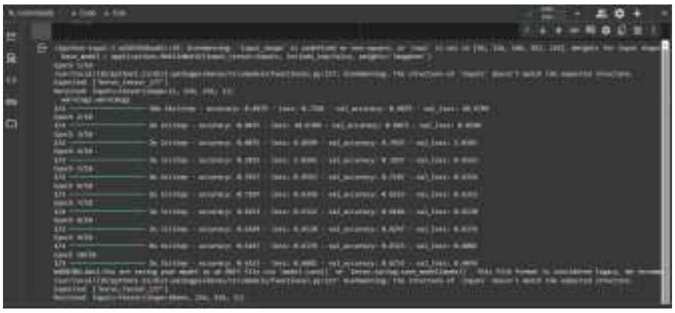
1.Create and image folder first with all the dataset images folders in it. Run the code below and Collect all the images in the image folder you have created above . and create a function.py file and write down the code in the second image and save it .

2.Run the code below to create an Mp_data folder where all of the nodes (.npy) are stored.

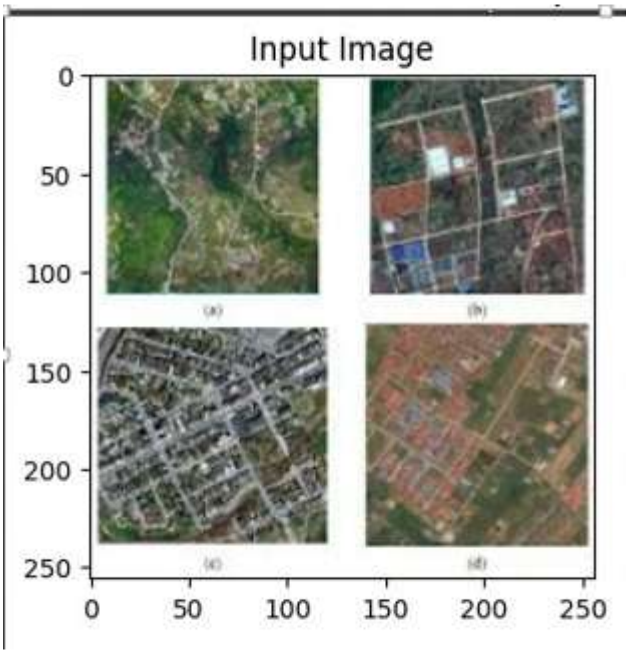
3. Now train the model using tensor flow so that it can detect the roads.



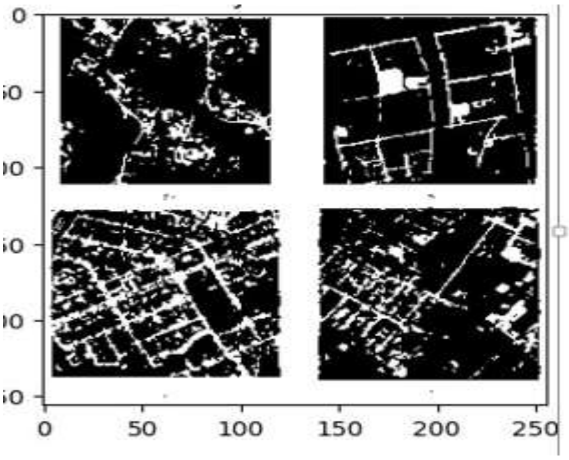
Result

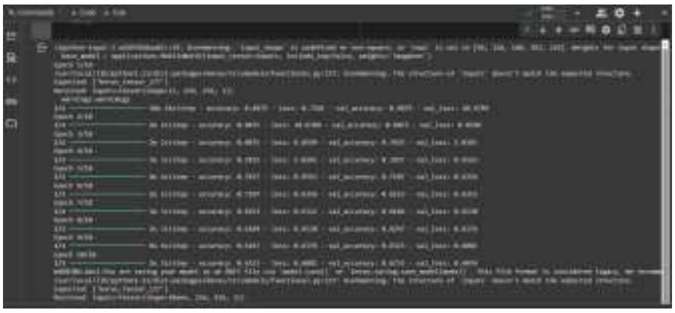


Run the above code, input the images shown in the image below and observe the live output.



The output is shown below with the accuracy of the roads.





4.CONCLUSION

In this project, a Skip-Connected Parallel CNN-Transformer Encoder-Decoder Model was proposed for accurate road extraction from remote sensing images. The hybrid architecture effectively combines the strengths of CNNs in capturing fine local details and Transformers in modelling global spatial dependencies. By using parallel feature extraction and skip connections at multiple levels, the model achieves improved segmentation performance, especially in challenging environments with occlusions, noise, and varying road structures. Experimental results demonstrate the model's capability to generate continuous, precise road maps that outperform traditional and single-architecture deep learning models. This approach not only enhances the quality of extracted road networks but also lays the foundation for future advancements in geospatial analysis, autonomous systems, and smart city development.

ACKNOWLEDGEMENT.

We would like to express our sincere gratitude to our guide and faculty members for their invaluable support and guidance throughout the course of this project. We also thank our institution for providing the necessary resources and environment to carry out this research. Lastly, we appreciate the contributions of all team members whose dedication and collaboration made this project possible.

I deeply grateful to our esteemed faculty mentors, **Dr. Sonagiri China Venkateswarlu, Dr. V. Siva Nagaraju**, from the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).

Dr. Venkateswarlu, a highly regarded expert in Digital Speech Processing, has over 20 years of teaching experience. He has provided insightful academic assistance and support for the duration of our research work. Dr. Siva Nagaraju, an esteemed researcher in Microwave Engineering who has been teaching for over 21 years, has provided us very useful and constructive feedback, and encouragement which greatly assisted us in refining our technical approach.

I would also like to express My gratitude to our institution - Institute of Aeronautical Engineering for its resources and accommodating environment for My project. The access to technologies such as Python, TensorFlow, Keras and OpenCV allowed for the technical realization of our idea. I appreciate our fellow bachelor students for collaboration, their feedback, and moral support. Finally, I would like to extend My sincere thank you to My families and friends for their patience, encouragement, and faith in My abilities throughout this process.

REFERENCES

- [1]Xu, W.; Wei, J.; Dolan, J.M.; Zhao, H.; Zha, H. A Real-Time Motion Planner with Trajectory Optimization for Autonomous Vehicles. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012.
- [2]Wang, W.; Yang, N.; Zhang, Y.; Wang, F.; Cao, T.; Eklund, P. A review of road extraction from remote sensing images. J. Traffic Transp. Engin. (Eng. Ed.) 2016, 3, 271–282. [CrossRef]
- [3]Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road Extraction Methods in High-Resolution Remote Sensing Images: A Comprehensive Review. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2020, 13, 5489–5507. [CrossRef]
- [4]Mo, S.; Shi, Y.; Yuan, Q.; Li, M. A Survey of Deep Learning Road Extraction Algorithms Using High-Resolution Remote Sensing Images. Sensors 2024, 24, 1708. [CrossRef] [PubMed]
- [5]Anil, P.N.; Natarajan, S. A Novel Approach Using Active Contour Model for Semi-Automatic Road Extraction from High Resolution Satellite Imagery. In Proceedings of the Second International Conference on Machine Learning & Computing, Bangalore, India, 9–11 February 2010.
- [6]Abraham, L.; Sasikumar, M. A fuzzy based road network extraction from degraded satellite images. In Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 22–25 August 2013.

BIOGRAPHIES

Vuyyuru Dharani Sri studying 3rd year department of Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .She Published a Research Paper Recently At IJSREM as a part of academics . She has a interest in Embedded Systems and VLSI.

Dr Sonagiri China Venkateswarlu professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Digital Speech Processing. He has more than 40 citations and paper publications across various publishing platforms, and expertise in teaching subjects such as microprocessors and microcontrollers , digital signal processing, digital image processing, and speech processing. With 20 years of teaching experience, he can be contacted at email: c.venkateswarlu@iare.ac.in

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Microwave Engineering. With over 21 years of academic experience, Dr. Nagaraju is known for his expertise in teaching core electronics subjects and has contributed significantly to the academic and research community. He can be contacted at email: v.sivanagaraju@iare.ac.in.