1. Introduction

Road detection from satellite imagery has become a critical tool in various applications such as urban planning, transportation management, disaster response, and environmental monitoring. In the Indian context, road detection plays a vital role due to the country's rapid urbanization, diverse geographical terrains, and complex infrastructure. The effective extraction of road networks from satellite images helps in improving transportation systems, providing updated maps, and supporting disaster management efforts. High-resolution satellite imagery, captured from advanced sensors such as IKONOS, QuickBird, and GeoEye, offers detailed geographical information essential for various sectors, including land-use planning, agriculture, and traffic management [1][2]. However, despite these advancements, extracting road networks in India remains a significant challenge due to the complex road infrastructure, varied terrains, and frequent occlusions caused by vegetation, buildings, and vehicles [1][5]. Roads in India often exhibit diverse patterns, from well-maintained highways in urban areas to poorly defined rural roads, which further complicates detection tasks.

To tackle these challenges, we propose INRnet, the modified U-Net architecture with attention blocks. This approach is particularly effective in road extraction because of its ability to focus on relevant road features while minimizing the impact of irrelevant elements in the image, such as trees, buildings, and vehicles. The attention mechanism allows the model to adapt to the complexity of road networks in India, where road types and conditions vary greatly between urban and rural areas [3]. The modified U-Net, when enhanced with attention blocks, significantly improves detection accuracy by concentrating on fine-grained road features and mitigating occlusions. It also excels in scenarios where labelled data is scarce, making it highly suitable for regions in India where acquiring large-scale labelled datasets can be challenging [3][5]. The model is able to generalize better across different terrains and infrastructure types, offering a more robust and scalable solution compared to traditional methods.

Despite the potential of various road extraction methods, many existing algorithms face significant drawbacks when applied to Indian road networks. Traditional approaches, have shown limited effectiveness in dealing with the unique challenges posed by India's road infrastructure. For example, supervised methods like random forests and support vector machines, although effective in some contexts, require large labelled datasets and are often unable to accurately detect roads in areas with poor visibility or significant occlusion. Additionally, these models tend to struggle with road-like patterns in cluttered environments, where roads are not easily distinguishable from other objects such as buildings and vegetation [5]. Unsupervised techniques, such as clustering, also face limitations due to their reliance on predefined criteria and their inability to adapt to diverse road features. Furthermore, many traditional methods are time-consuming and prone to human errors, particularly when manual intervention is required for dataset labelling or post-processing [5].

Our approach INRnet the modified U-Net with attention blocks overcomes many of these challenges. By using deep learning, the model is capable of performing efficient, pixel-by-pixel classification with higher accuracy and fewer errors compared to conventional methods. The attention mechanism enhances the model's

focus on critical road features, such as road boundaries and intersections, while reducing the influence of non-road objects in the image. This approach not only improves detection accuracy in urban environments but also adapts well to rural areas with less-defined roads, where traditional methods tend to fail. Additionally, the model's ability to function with minimal labelled data makes it an ideal solution for regions where data annotation is scarce or expensive. By overcoming the limitations of existing algorithms, our approach provides a more reliable, efficient, and scalable solution for road extraction from satellite imagery in India [3][5]. This can significantly contribute to better urban planning, disaster management, and transportation network optimization in India's rapidly evolving landscape.

2. Related Work:

In recent years, research on road detection from satellite imagery has grown rapidly, thanks to advances in remote sensing and machine learning. Earlier methods, like clustering and morphological operations, performed well in simpler settings but often struggled with the complexity of urban and rural landscapes. More recently, deep learning models have shown improved results by using layered feature extraction and attention mechanisms to tackle challenges like occlusions and spectral similarities in high-resolution images. In this section, we group the existing solutions for road detection into modern techniques.

2.1 Modern Techniques

Remove 2.1 title

The U-Net architecture, widely used in image segmentation tasks, has been extensively explored and modified in research to improve its performance for specific applications. Variants such as Dense U-Net incorporate dense connections between layers to enhance feature propagation and gradient flow [1]. The standard U-Net employs an encoder-decoder structure with skip connections, thereby allowing effective feature extraction and precise localization by capturing both local and global contextual information within images [10]. The encoding phase processes three-channel RGB input images through convolutional and max-pooling layers, reducing spatial dimensions progressively while increasing the number of feature channels to extract high-level features [10]. Furthermore, lightweight modifications, such as U-Net Mini, utilize fewer layers and smaller filters, are suitable for low-resource settings or smaller dataset applications for the sake of computational efficiency [11].

Other works on Fully Convolutional Networks (FCNs) report both the advantages and challenges of using such networks. Although convolutional layers are very efficient in extracting features from input images, down sampling operations in them decrease the spatial resolution of feature maps, which is challenging for pixel-wise prediction at the original image resolution required for semantic segmentation [2], [6]. To overcome this, FCNs use up sampling techniques like deconvolution or transposed convolution and bilinear interpolation [2], [6]. Deconvolution uses learnable filters to reconstruct higher-resolution feature maps with higher accuracy, albeit at a higher computational cost, whereas bilinear interpolation offers a simpler and faster method by estimating up sampled pixel values through a weighted average of neighbouring pixels [2]. However, FCNs still face challenges, including the high computational cost of processing large-scale, high-resolution images, particularly in remote sensing applications [7].

Building on these approaches, subsequent models of deep learning have been developed to overcome bottlenecks in segmentation tasks. For instance, high resolution images can be processed efficiently through patch-based CNNs in small patches of an image individually, which can easily facilitate parallel processing and processing without overloading the computations as large images [4]. The predictions from all of the patches are integrated afterwards to produce the final segmentation map. Variants of U-Net include the deep residual U-Net, which includes features such as residual connections for the simplification of the training of deeper networks and an improvement in accuracy by learning more complex features [8]. Similarly, VNet uses convolutional layers rather than pooling layers to preserve resolution and utilizes a dual loss function referred to as CEDL combining cross-entropy loss with Dice loss to effectively handle class imbalance which is particularly advantageous in the case of road extraction tasks [5]. ResNet and SegNet, amongst others, have been customized for these purposes as well. ResNet uses residual connections to speed up training, while SegNet uses an encoder-decoder architecture to produce highly accurate and high-resolution segmentation maps by effectively up sampling [5], [9].

Refer the table here and state that none of the works are available for Indian Roads upto our knowledge

Method	Year	Remarks	Indian Roads	Attention
Proposed	2024	Based on a U-Net backbone, it applies encoder-decoder structure for encoding features at high and low levels toward very accurate pixel-wise image segmentation.	✓	✓
INFOROAD framework [9]	2024	A modified U-TAE network architecture is used on images from the Seich-Sou Forest in Thessaloniki, Greece. This modification improves segmentation performance and feature extraction.	✗	✓
VGG-UNET [11]	2024	The VGG architecture is used because it can extract rich and hierarchical features from images. It is particularly useful for handling complex image segmentation tasks in the IEEE Dataport dataset.	✗	✓
UNET MINI [12]	2024	UNet Mini is designed to perform efficient segmentation on small datasets or within resource-limited environments. The light design is ensured to give accurate segmentation with minimal computation.	✗	✓
UNET [13]	2024	UNet is used in the Coast Train Dataset to perform edge detection. The implementation in this case targets coastline segmentation, which it does accurately and in high detail.	✗	✗
UNET-Attention [1]	2023	BRISQUE preprocessing is used on the Massachusetts road dataset to improve image quality and enhance segmentation	✗	✓

		performance. This helps in preparing the data for better model input and analysis.		
FCN [6]	2021	The methodology consists of Very High Resolution (VHR) images acquisition and developing a reference map based on OpenStreetMap data. It will use an FCN model consisting of an encoder-decoder network for segmentation tasks.	✗	✗
MRF CNN [3]	2020	A hybrid model is used for road network extraction, combining multiple techniques to enhance accuracy. This approach effectively maps the road networks within the dataset.	✗	✓
VNET CEDL [5]	2020	The Ottawa Road Dataset contains Cross-Entropy Dice Loss, a loss function aimed at increasing segmentation accuracy. It guarantees well-balanced predictions and sharp delineation of edges.	✗	✓
FCN [7]	2018	For SAR images, fully Convolutional Neural Networks (FCNNs) are applied in order to extract the road network. The utilization of SAR images benefits from data unique properties of SAR, so better extraction of the roads.	✗	✗

give table caption and cite the table in text of previous paragraph

PROBLEM STATEMENT Remove the title

Extracting road networks from satellite imagery in India comes with unique challenges that make it especially complex. Indian roads are diverse, ranging from well-maintained highways in cities to unmarked, uneven paths in rural areas, creating inconsistencies that models struggle to interpret. Shadows from trees and buildings, frequent vehicle presence, and spectral similarities between road surfaces and surrounding land make it hard for algorithms to distinguish roads accurately. Seasonal changes, like flooding during monsoon seasons, alter road appearance, adding another layer of complexity to accurate detection. In dense urban areas, intricate networks with tight intersections and alleyways make segmentation harder. Furthermore, there is often limited access to well-annotated datasets for different regions, which limits training and testing accuracy. These factors together make road extraction in India challenging, requiring models that can handle a high degree of variability and adapt to changing environments. The Attention Gate (AG) mechanism is really helpful for improving road detection, especially when dealing with complex datasets like the Indian Road Dataset (INRdataset). In India, road networks are often obstructed by things like trees, buildings, or vehicles, which makes road detection harder. The attention gate helps the model by focusing on the important parts of the image, like the roads, and ignoring unnecessary areas. This

Shift the content inside proposed method

way, the model can detect and segment roads more accurately, even in messy environments that are common in both urban and rural areas of India.

The following are the significant contributions in this research:
1. INRDataset Curation: We curated INRDataset which is highly essential to validate our goal
2. INRNet Design:.....
3. Evaluation:.....
4. COmparison:.....

Creation of dataset:

To create the dataset, we first selected high-resolution satellite images from Google Earth Pro, focusing on the roads of Kelambakkam near VIT. This area was chosen for its diverse urban features and detailed spatial resolution, making it ideal for training and evaluating road and building detection models.

The selected satellite images were carefully cropped to target specific regions within Kelambakkam, narrowing down the dataset to relevant sections. We used CVAT for annotation, manually marking building and road boundaries to create accurate ground-truth data essential for model training and validation. This manual annotation phase was critical in ensuring the dataset's accuracy and relevance.

To further refine the dataset, we applied preprocessing techniques to enhance image visibility and clarity, improving the model's ability to detect subtle features. Finally, the dataset was divided into training and testing sets, with most images allocated for training, while a smaller portion was reserved for testing to evaluate model performance. This structured approach allows the model to generalize better across different urban environments, making it a solid foundation for high-accuracy road and building detection in similar contexts.

DEVELOPING INRnet:

The INRNET architecture includes three main parts: an encoder, a decoder, and attention gate (AG) blocks. These components together allow the model to focus more effectively on road detection within satellite images, making the segmentation more accurate.

1. **Encoder Block**
   The encoder is responsible for extracting features by passing the input image through multiple convolutional layers.
2. **Decoder Block**
   The decoder works to reconstruct the image back to its original resolution using up sampling layers.
3. **Attention Gate (AG) Block**
   The attention gates sit on the skip connections between the encoder and decoder, filtering out unimportant features and highlighting road-specific information.

Overall, INRNET's combination of encoder-decoder with attention gates gives it the ability to capture essential details, ignore distractions, and adapt to challenging imagery, making it particularly effective for complex road extraction tasks.

Implementing INRnet on INRdataset:

The project makes use of the Indian Road Dataset (INRdataset), which comprises 1536 high-resolution satellite images split into 256x256 tiles, resulting in 1,506 tiles. Specifically, these tiles capture particular portions of roads, thus enhancing efficiency and accuracy in the models. The road detection model, INRnet, is based on U-Net architecture with an encoder-decoder structure and attention gates, which focus on road elements while ignoring irrelevant ones. With pixel-wise classification approach and the loss function of Dice coefficient, INRnet effectively handles imbalanced datasets. Preliminary experimental results indicate that it is a useful tool in urban planning and road transportation management across roads in India for various regions.

Remove these content and have it in separate document

3. INRDataset Curation
paste 3.1 content under this heading

4. Proposed Methodology: INRNet Architecture

3.Proposed Methodology :

3.1)IR Dataset creation :

In this study, an Indian roads (IR) dataset was meticulously developed to aid in road detection within high-resolution satellite imagery. Captured through Google Earth Pro, the dataset comprises two images taken around VIT Chennai, each with a remarkable resolution of 8192x4320 pixels, equivalent to 8k UHD at 96 dpi. This high resolution was selected to ensure that the captured images offer exceptional detail, which is particularly advantageous for applications requiring precise road detection and segmentation in infrared remote sensing tasks.

To accurately delineate roads and enhance the dataset's utility, the images were processed using the Computer Vision Annotation Tool (CVAT), a robust platform designed for detailed image annotation. Within CVAT, roads were manually masked across the high-resolution images, isolating these features as the primary areas of interest. This careful masking process involved highlighting only the road sections, thereby filtering out extraneous details such as vegetation, buildings, and water bodies. The manual approach ensured that the masking remained highly accurate and specific to roadways, thereby emphasizing these critical features for the subsequent machine learning tasks.

Following the annotation phase, each masked image was systematically segmented into smaller tiles of 256x256 pixels. This step not only optimizes computational efficiency but also retains essential spatial information for each road section, allowing machine learning models to effectively recognize and analyze road patterns within manageable image sizes. This segmentation process yielded a total of 1,506 individual image tiles, each containing a distinct portion of the masked roads from the original captures. By working with these smaller, context-preserving image sections, the dataset is structured to facilitate efficient training of models aimed at road detection within infrared satellite imagery.

This dataset creation approach lays a strong foundation for road detection in Geosatellite imagery, providing clear and precise data that are essential for developing and refining machine learning models tailored to remote sensing applications. The careful annotation, segmentation, and focus on road-specific features make this dataset a valuable resource for advancing road detection capabilities.

S.No	Area	Train/validation	Number of images
1	Around VIT Chennai	Train	1305
2	Around VIT Chennai	Validation	231
		Total	1536

3.2)INRNET Architecture

3.2.1 U-Net Architecture in INRNET

4.1 Fundamental Block in INRNet

INRNET is designed with a U-Net backbone, a fully convolutional network architecture known for its encoder-decoder structure. This structure is particularly suited for image segmentation as it captures both high-level and low-level features through the sequential compression and expansion of the feature maps, thus enabling precise pixel-wise predictions.

Encoder Path (Down-Sampling)

In the encoder path, the network extracts feature representations at multiple scales, with each layer progressively capturing more abstracted information from the input image. Each encoding layer C_i in the encoder path operates as follows:

$$C_i = \text{ReLU}(W_i * C_{i-1} + b_i)$$

where:

- W_i represents the convolutional filter applied at the i -th layer,
- C_{i-1} is the input feature map from the preceding layer,
- b_i is the bias term added to each convolution, and
- $\text{ReLU}(x)=\max(0,x)$ serves as the activation function.

Each convolutional layer in the encoder path is followed by a max-pooling operation, which reduces the spatial dimensions while preserving salient features, thus achieving a compact feature representation at each stage. As depth increases, the number of feature channels doubles, allowing the network to store a higher volume of spatial and contextual details.

Decoder Path (Up-Sampling)

The decoder path mirrors the encoder path, progressively restoring the spatial dimensions of the feature maps to reconstruct the original image resolution. Each layer in the decoder performs up-sampling through transpose convolutions, which increase the spatial resolution of feature maps, effectively reversing the down-sampling effect seen in the encoder. This process can be represented as:

$$T_i = W_i^T * T_{i-1} + b_i$$

where:

- T_i denotes the up-sampled feature map,
- W_i^T is the transpose convolutional filter,
- T_{i-1} is the output from the previous layer.

In addition to up-sampling, each stage in the decoder incorporates a skip connection from the corresponding encoder layer. These connections help retain high-resolution spatial details lost during down-sampling, effectively merging fine-grained features with the abstracted representations. The concatenation operation in each decoder layer is represented as:

$$U_i = \text{Concatenate}(T_i, C_j)$$

where U_i represents the fused feature map at decoder layer i , with T_i being the up-sampled feature map and C_j being the feature map from the corresponding encoder layer.

4.2 Attention Block in INRNet

3.2.2 Attention Gate Mechanism in INRNET

The attention gates (AGs) incorporated into INRNET operate to refine and enhance feature maps by focusing on salient regions of the image while suppressing non-essential regions. These AGs are particularly useful for pixel-level segmentation tasks, such as identifying road features within infrared images, by highlighting only relevant spatial regions.

Attention gates in INRNET take two inputs:

1. Feature Map x : Extracted from the encoder path, providing spatial information about the regions of interest.

2. Gating Signal g : Extracted from a lower-resolution decoder layer, offering high-level contextual information.

Each AG modulates the feature map based on its relevance to the gating signal, through the following transformations:

1. Linear Transformation of Inputs:

To align dimensions between x and g , both inputs are linearly transformed:

$$\theta_x = W_\theta * x + b_\theta$$

$$\phi_g = W_\phi * g + b_\phi$$

where:

- W_ϕ and W_θ are weight matrices applied to g and x, respectively,

- b_θ and b_ϕ are biases,

- θ_x and ϕ_g are the transformed inputs.

2. Combination via Addition and Activation:

The transformed features are added element-wise, producing an intermediate response map that reflects regions of interest, based on both spatial features from x and contextual cues from g:

$$f = \mathrm{ReLU}(\theta_x + \phi_g)$$

3. Sigmoid Transformation for Masking:

The combined response map f undergoes a sigmoid transformation to produce a scaling mask, effectively selecting features with high relevance:

$$\psi = \sigma(W_\psi * f + b_\psi)$$

where W_ψ and b_ψ are additional learned parameters, and σ denotes the sigmoid function, producing output in the range [0, 1].

The resulting mask ψ is applied to the original feature map x through element-wise multiplication, yielding an attention-modulated feature map x′ that retains only the most relevant regions for segmentation:

$$x' = x \cdot \psi$$

where (·) denotes element-wise multiplication.

Through this attention gate mechanism, INRNET achieves refined focus on critical image regions, enabling it to discriminate between relevant and irrelevant pixels, which is particularly essential for detecting roads in complex, infrared satellite images. This refined focus enhances INRNET's ability to perform precise segmentation by dynamically adjusting to the image's contextual needs at each pixel level.

Paste the INRNet architecture and attention block architecture

4)Experimentation

4.1)Experimental Setup

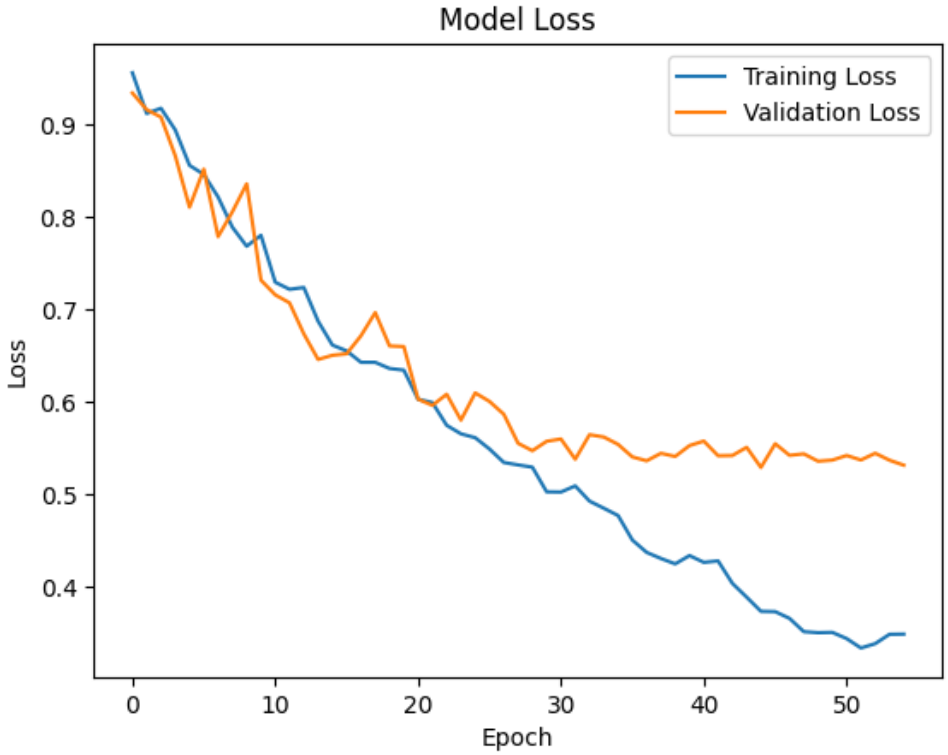
This experiment was conducted on Google Colab, utilizing an NVIDIA T4 GPU to manage model training efficiently. By leveraging Colab's GPU acceleration, we were able to expedite the training process, achieving results faster than on a CPU setup. The core libraries for this study included TensorFlow (v2.x) and Keras for model building and training, alongside OpenCV for image processing tasks. Additional essential Python packages, such as NumPy, Pandas, and Scikit-learn, were used to support data preprocessing, analysis, and various operations throughout model development. The image dataset, specifically curated for road segmentation tasks, was stored on Google Drive, allowing seamless integration with Colab for direct data access and management without the need for extensive data transfers.

Preprocessing steps were applied to prepare the images, including resizing to 256x256 pixels to ensure a uniform input size, normalization to standardize pixel values, and patch extraction to enhance the model's feature-capturing ability. These preprocessing steps aimed to optimize input consistency and boost the model's capacity to detect fine-grained features within road images. The proposed model, named INRNET, is an advanced variation of U-Net. INRNET integrates attention mechanisms and Conv2DTranspose layers within the upsampling path, which are specifically designed to preserve spatial details crucial for accurate road boundary detection. In the downsampling path, Conv2D layers with ReLU activation were employed to capture essential features, while a Dropout rate of 0.25 was applied to mitigate overfitting risks and enhance generalization. To further refine boundary delineation, the upsampling path incorporated attention

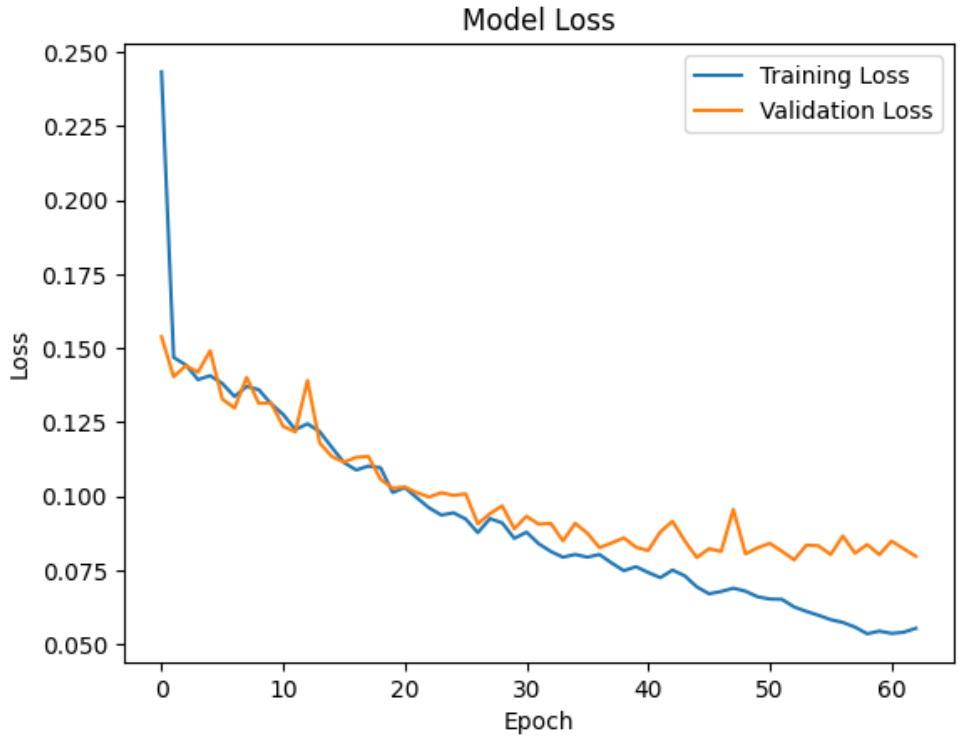
gates and difference-based convolution blocks, which enhance the model's focus on key spatial details.

The training setup for INRNET involved a batch size of 32 and a total of 100 epochs. Early stopping on validation loss was implemented with a patience threshold of 10 epochs to ensure efficient convergence while avoiding unnecessary computation in cases of stalled improvement. The Adam optimizer was utilized for parameter updates, paired with binary cross-entropy as the loss function to optimize the model's performance on binary segmentation tasks. Intersection over Union (IoU) and accuracy metrics were employed as primary evaluation metrics to assess segmentation quality on the validation set, providing a robust measure of the model's precision and recall in segmenting road areas. These configurations collectively contributed to achieving a balance between model performance, computational efficiency, and segmentation accuracy, demonstrating INRNET's potential for practical applications in road image analysis.

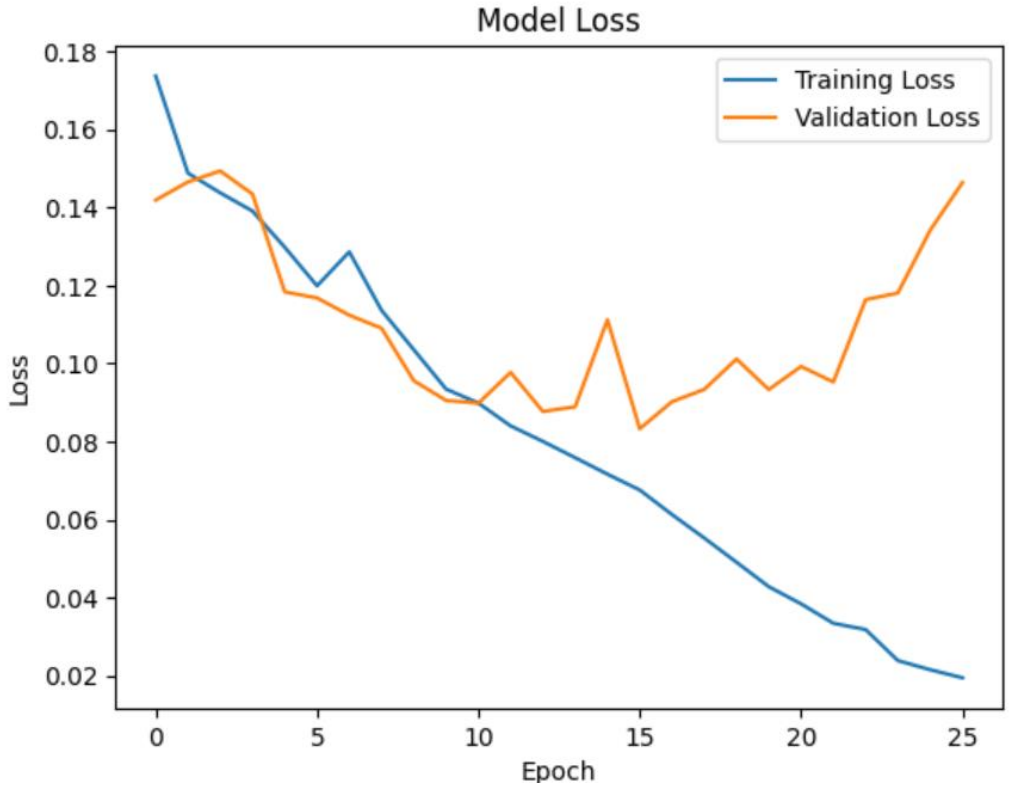
4.2)Results



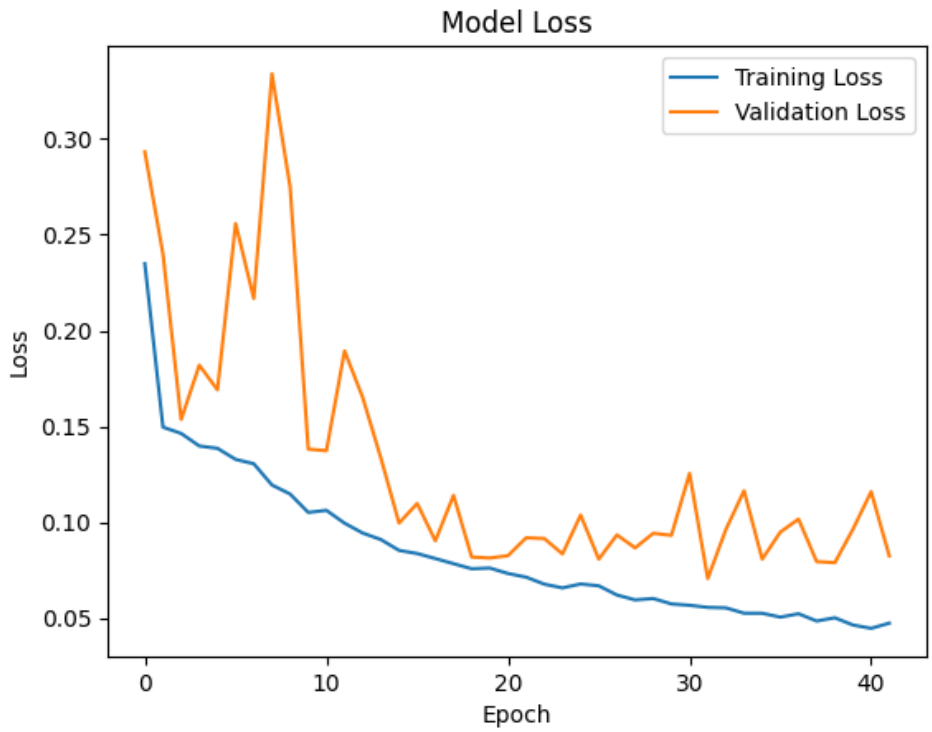
Learning curve for Unet



Learning curve for FCN model



Learning curves for Unet++ model



Learning curve for INRNet

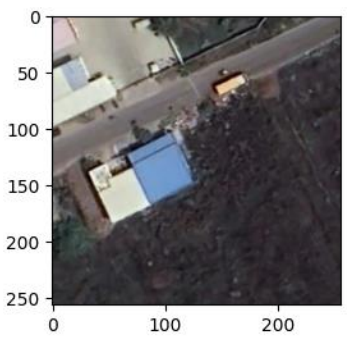
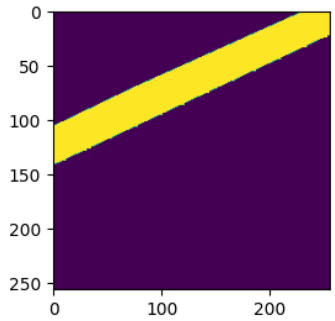
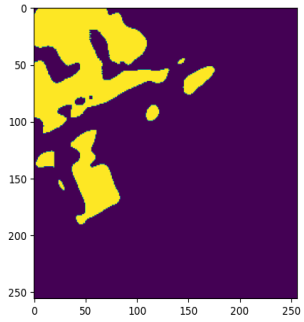
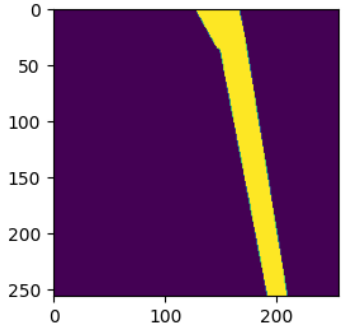
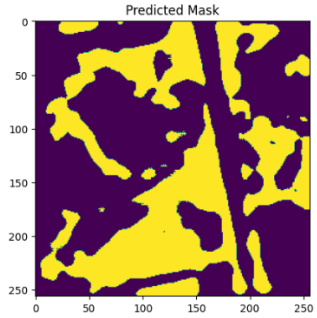
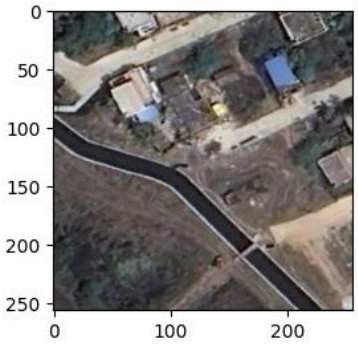
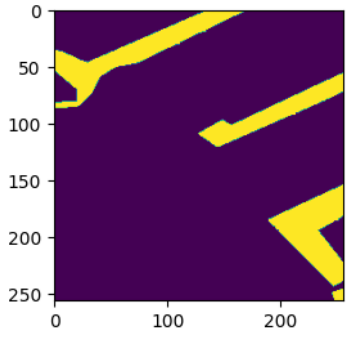
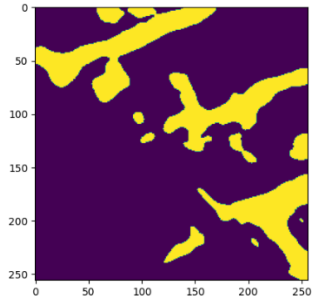
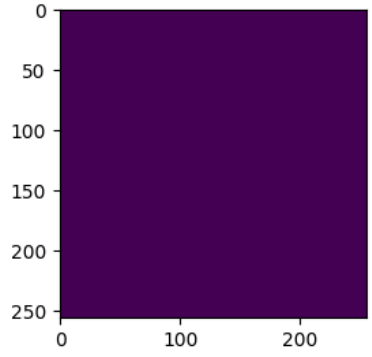
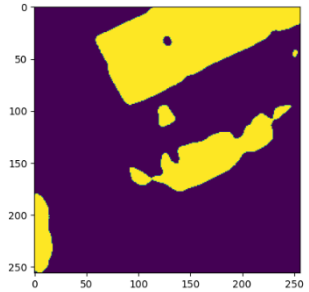
Model	Mean Average Precision	Recall	IoU	Precision	Best Epoch
Unet	0.5530	0.6214	0.5288	0.6390	45
Unet++	0.4352	0.4656	0.2654	0.5435	16
FCN	0.5507	0.4782	0.3076	0.6959	53
INRNet	0.5746	0.5576	0.3397	0.6973	37

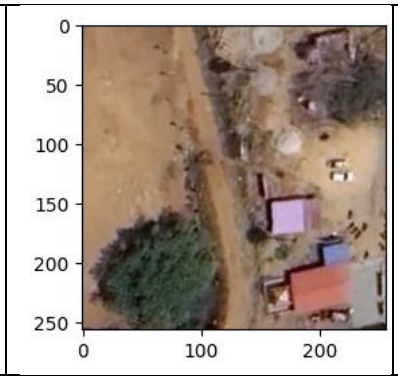
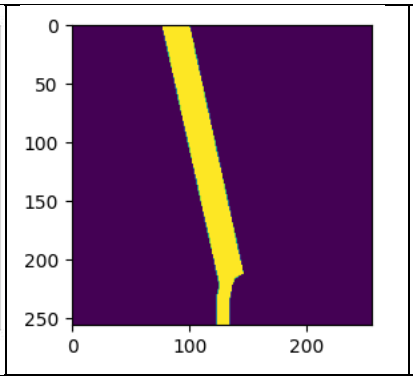
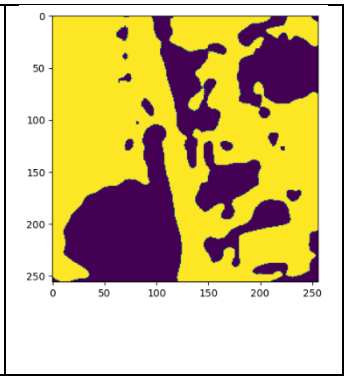
comparsion table

The table presents a comparison of four models—U-Net, U-Net++, FCN, and INRNet—for the task of road detection using geosatellite images. Among these, INRNet outperforms others in terms of mean average precision (0.5746), indicating its ability to provide more accurate predictions for road segments. While U-Net has a slightly lower mean average precision (0.5530), it excels in recall (0.6214), suggesting it is more capable of detecting road features comprehensively. INRNet, however, achieves a better balance between precision (0.6973) and recall (0.5576), alongside the highest IoU (0.3397), which reflects superior alignment between its predictions and ground truth. In contrast, U-Net++ underperforms across most metrics, with the lowest mean average precision (0.4352) and IoU (0.2654), making it less effective for this task. FCN, while achieving a comparable mean average precision to U-Net (0.5507), performs moderately in recall (0.4782) and IoU (0.3076). INRNet also achieves its best results at epoch 37, showcasing faster convergence compared to U-Net (45 epochs) and FCN (53 epochs), making it the most efficient and effective model for road detection from geosatellite imagery.

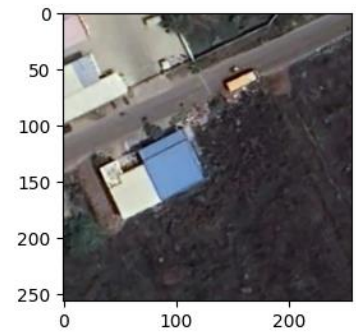
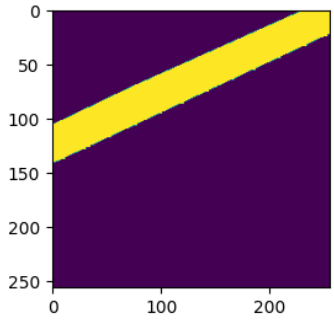
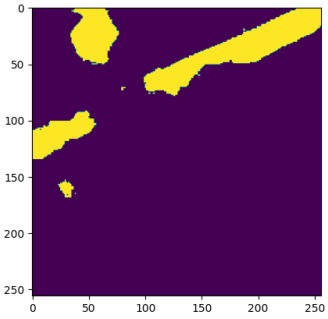
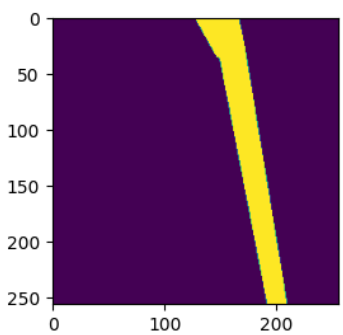
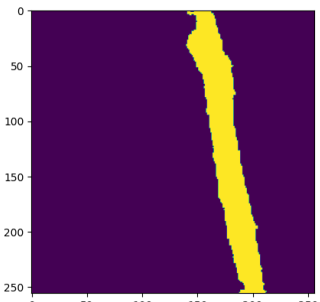
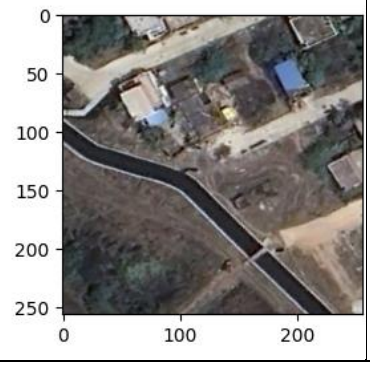
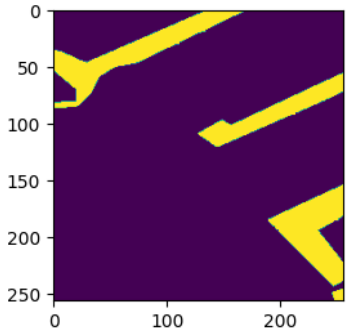
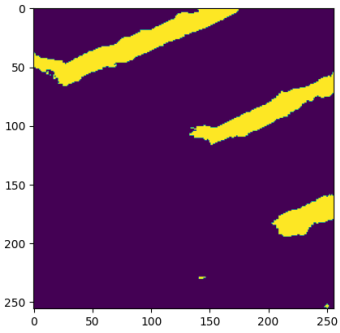
The following section will showcase 5 images (2 urban and 3 rural) with their respective image IDs, original images, masks, and predicted masks.

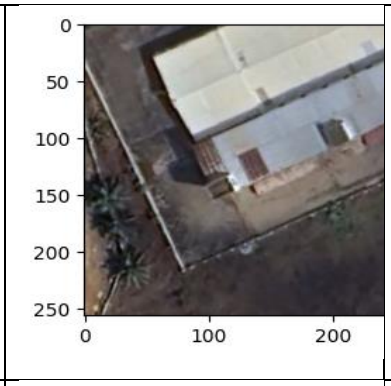
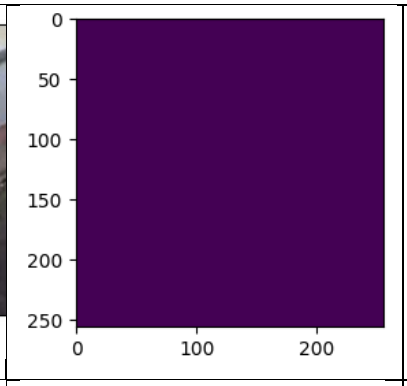
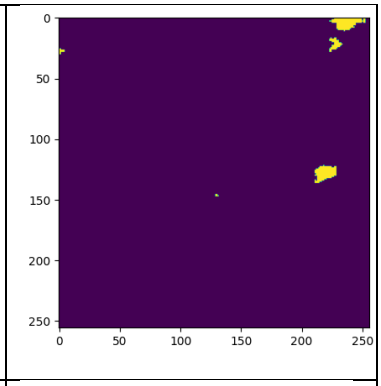
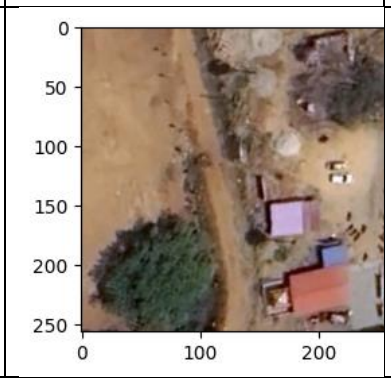
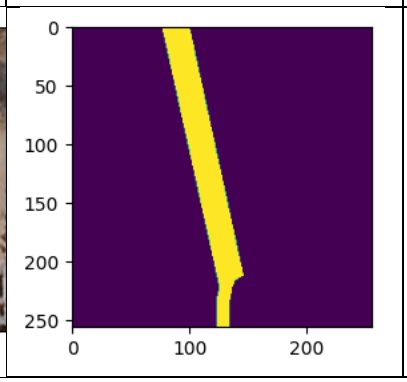
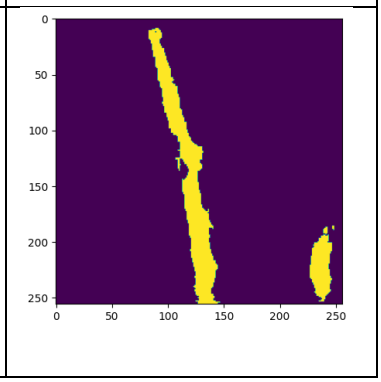
The below is the section that shows the results of Unet

ID	Original Image	Original Mask	Predicted Mask
30			
32			
55			
69			

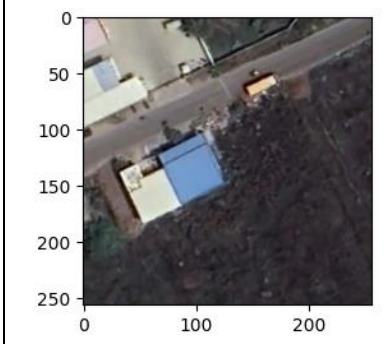
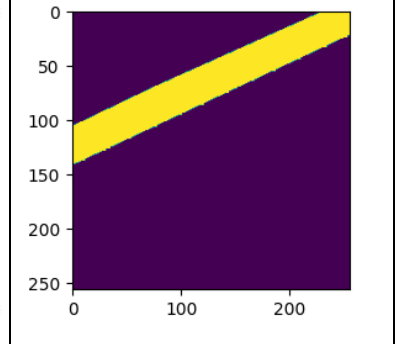
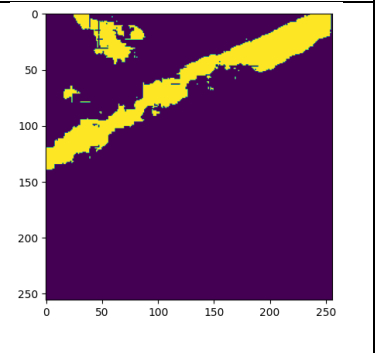
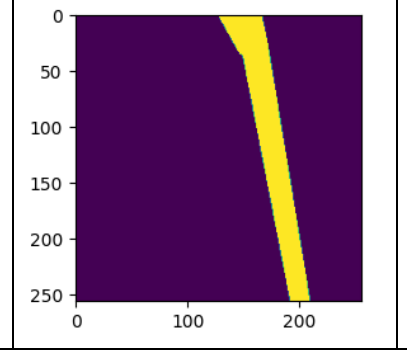
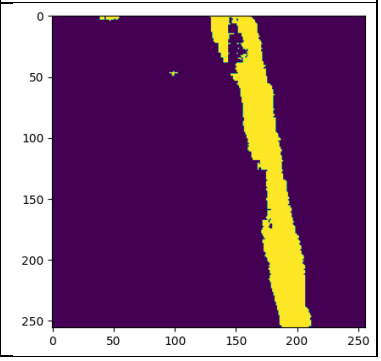
90			

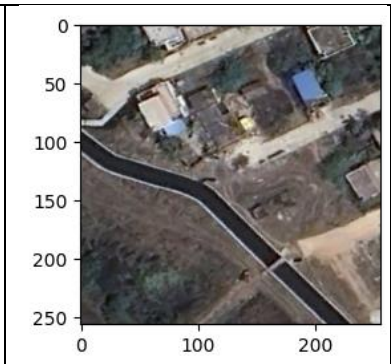
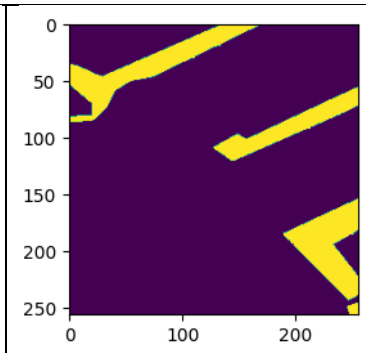
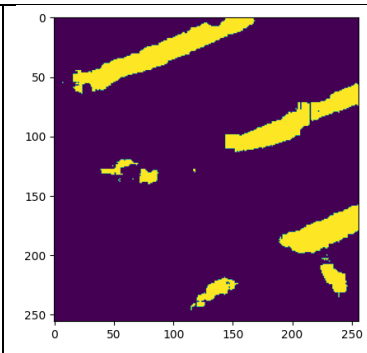
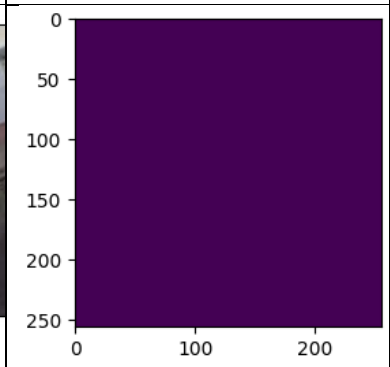
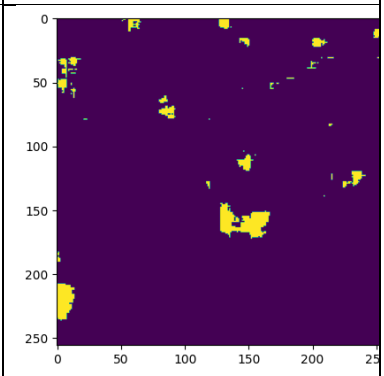
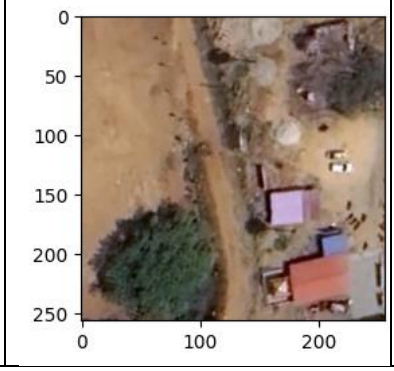
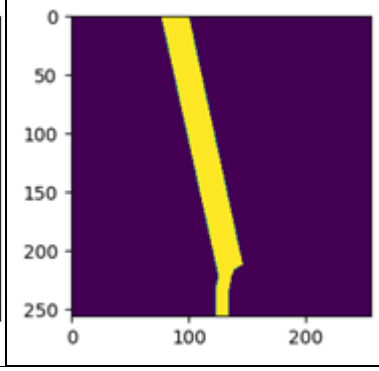
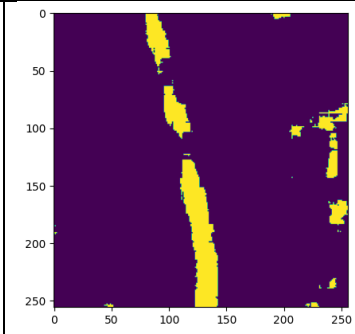
The below is the section that shows the results of Unet++

ID	Original Image	Original Mask	Predicted Mask
30			
32			
55			

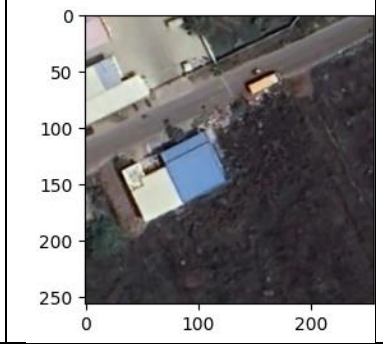
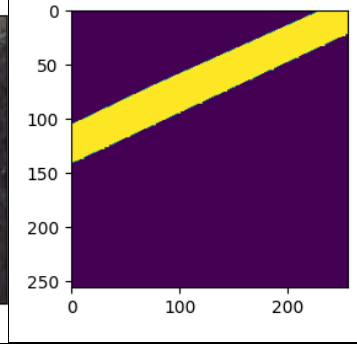
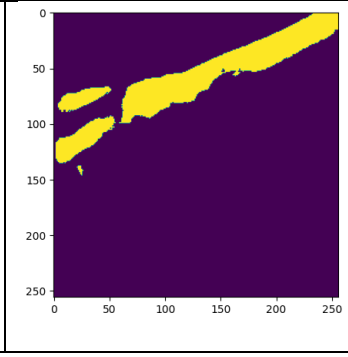
69			
90			

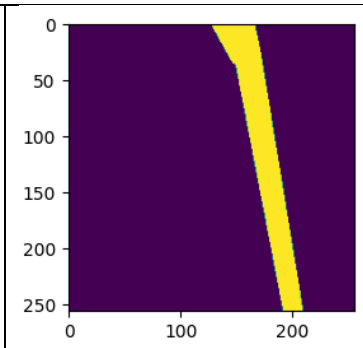
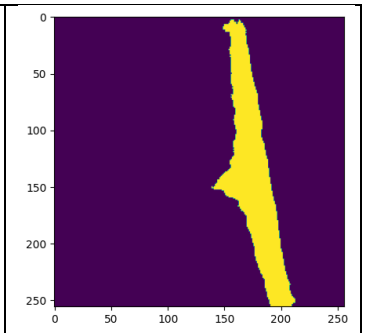
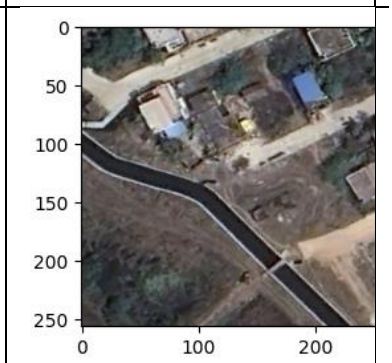
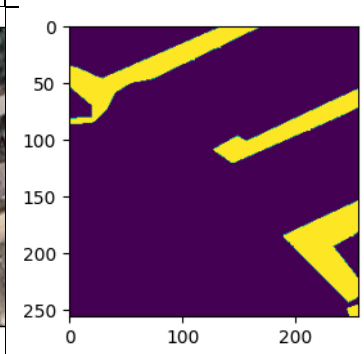
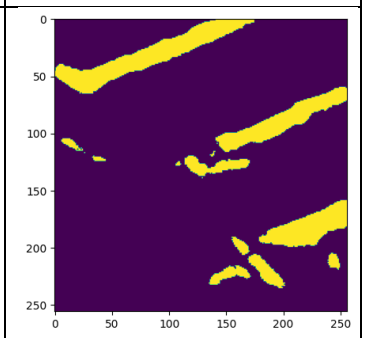
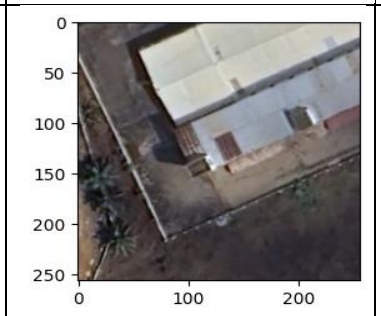
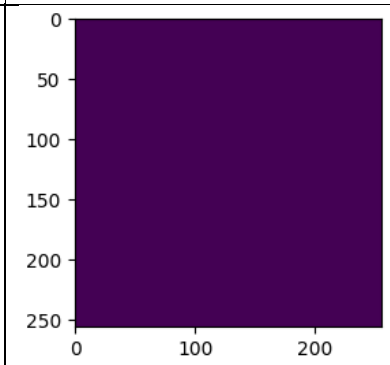
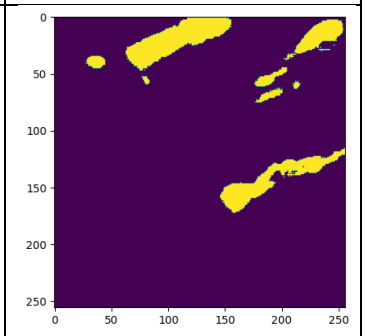
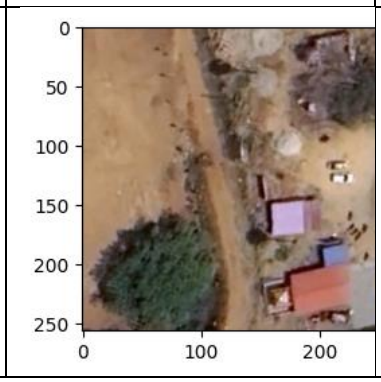
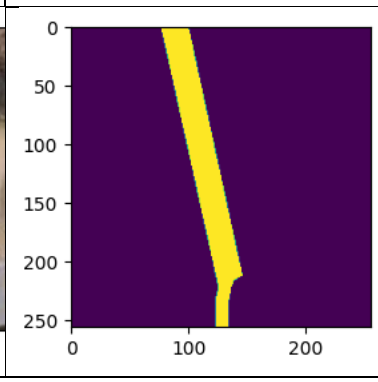
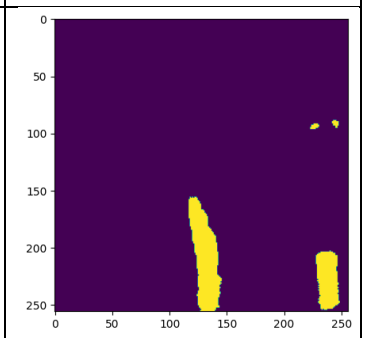
The below is the section that shows the results of FCN

ID	Original Image	Original Mask	Predicted Mask
30			
32			

55			
69			
90			

The below is the section that shows the results of INRNET

ID	Original Image	Original Mask	Predicted Mask
30			

32			
55			
69			
90			

4.3)Abalation studies

Loss used : binary cross entropy

Model	Mean Average Precision	Recall	IoU	Precision	Best Epoch
Unet	0.5087	0.6161	0.4342	0.6090	87
INRNet	0.5746	0.5576	0.3396	0.6973	37

we compared the performance of the baseline U-Net model with our proposed INRNet model using binary cross-entropy loss to assess their effectiveness in road image segmentation. The INRNet model shows significant improvement across multiple evaluation metrics, emphasizing its enhanced segmentation capabilities. Notably, the mean average precision (mAP) increased from 0.5087 for U-Net to 0.5746 for INRNet, suggesting a higher reliability in detecting relevant features and fewer missed objects. Although INRNet's recall (0.5576) is slightly lower than U-Net's (0.6161), it compensates with a higher precision, which improved from 0.6090 to 0.6973. This increase indicates that INRNet has reduced false positives, leading to more accurate segmentation. However, the Intersection over Union (IoU) value decreased from 0.4342 with U-Net to 0.3396 with INRNet, which might suggest that further refinement is needed to improve spatial accuracy in overlapping regions. Importantly, INRNet achieved its best results much earlier, at epoch 37, compared to epoch 87 for U-Net, showing faster convergence and better efficiency in training. Overall, these results underscore INRNet's potential as a superior model for road image segmentation, with promising avenues for future enhancement.

4.4)SoTa review

Road extraction from satellite imagery has been extensively studied due to its critical role in urban planning, transportation management, and disaster response. A study by Mohd Jawed Khan et al. (2023) proposed a fine-tuned U-Net model with data preprocessing techniques such as BRISQUE and data augmentation. Their approach, applied to the Massachusetts Roads Dataset, achieved significant results with an accuracy of 95.48% and an IoU of 60.97%. This U-Net variant utilized a standard encoder-decoder architecture with four hidden layers and ReLU activation, focusing on maintaining road geometry and minimizing segmentation errors in complex environments.

In contrast, INRNet, a modified U-Net architecture designed specifically for road detection in Indian satellite imagery, addresses the unique challenges posed by the region's diverse road networks. Indian roads are characterized by frequent occlusions from vegetation, buildings, and vehicles, as well as highly variable infrastructure patterns ranging from urban highways to rural paths. INRNet incorporates attention gates (AGs) within its encoder-decoder framework to enhance the model's focus on road-specific features, reducing the influence of irrelevant objects and improving segmentation performance in noisy and cluttered scenarios. The preprocessing pipeline for INRNet involves manual annotation of high-resolution satellite images using CVAT, followed by segmentation into 256x256 pixel tiles, enabling efficient training and improved adaptability across diverse terrains.

When comparing the performance metrics, INRNet exhibits superior precision and reliability in road segmentation tasks. It achieves a mean average precision of 0.5746, a precision of 0.6873, a recall of 0.5576, and an IoU of 0.3397. While its IoU is lower than the U-Net model's 60.97%, the higher precision and mean average precision indicate INRNet's robustness in identifying road

features while minimizing false positives. This tailored focus makes INRNet particularly effective in addressing the complexities of Indian road networks.

The data preparation strategies further highlight the differences between the two models. The U-Net model utilized BRISQUE to evaluate image quality and select high-quality samples for training, focusing on 200 images from the Massachusetts Roads Dataset and augmenting them with random transformations. In contrast, INRNet relied on manually annotated high-resolution images segmented into smaller tiles, ensuring detailed feature capture and effective training. Both approaches demonstrate the versatility of U-Net-based architectures for road extraction; however, INRNet's integration of attention mechanisms and its specialized preprocessing pipeline enable it to perform better in scenarios involving highly variable road conditions and challenging environments, such as those found in India.

Give the figure caption for all the figures and cite wherever required in the text..

Write abstract and conclusion