Innovative Deep Learning Approaches for Accurate Road and Building Segmentation from High-Resolution Satellite Images

Ali Hassan Khan
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021079@giki.edu.pk

Sabeer Faisal
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021447@giki.edu.pk

***Abstract*—The extraction of roads and buildings from satellite imagery is crucial for urban planning, infrastructure management, and disaster response. This study presents a comprehensive approach using advanced deep learning techniques, focusing on meticulous data preparation, augmentation, and thoughtful visualization.**

For road extraction, we utilized the DeepLabV3+ model with a ResNet50 encoder pre-trained on ImageNet. Our method included innovative data preparation, one-hot encoding, and extensive data augmentation to enhance model performance. The model achieved a high Intersection over Union (IoU) score on the validation set, demonstrating its accuracy in identifying roads in diverse landscapes. These results underscore the model's utility in geospatial analysis, infrastructure planning, and environmental monitoring.

In building extraction, we explored several architectures, including U-Net, Mask R-CNN, DeepLab, FCN, and PSPNet. Each model demonstrated unique strengths in accurately segmenting building footprints from satellite imagery. The U-Net and DeepLab models excelled in capturing detailed structures, while Mask R-CNN provided precise instance segmentation.

This study highlights the integration of data preparation, augmentation, and deep learning to improve road and building extraction from satellite imagery, offering significant advancements in computer vision and practical applications in various fields.

***Index Terms*—Automated road extraction, DeepLabV3+, satellite imagery, geospatial analysis, data preparation, one-hot encoding, ResNet50 encoder, ImageNet pre-training, infrastructure management, urban planning, environmental monitoring**

I. Introduction

In the realm of remote sensing and disaster response, accurately extracting road networks and building footprints from very high-resolution (VHR) satellite imagery is critical for urban planning, infrastructure management, and crisis response. The DeepGlobe Road Extraction Challenge exemplifies this urgency, aiming to autonomously delineate road networks from complex satellite imagery, enhancing global crisis management strategies.

978-1-6654-5992-1/22/$31.00 ©2022 IEEE

This research explores state-of-the-art solutions for both road and building extraction. We present a detailed examination of the DeepLabV3+ model, renowned for its pixel-level image classification and object delineation capabilities. Extending beyond conventional approaches, our method leverages nonlocal LinkNet with nonlocal blocks (NLBs) to capture intricate global feature relationships, ensuring accurate road segmentation. Our NL-LinkNet model outperforms the DeepGlobe Challenge champion, with 43

For building extraction, we explore various advanced architectures including U-Net, Mask R-CNN, DeepLab, FCN, and PSPNet. Each model demonstrates unique strengths in accurately segmenting building footprints. U-Net and DeepLab excel in capturing detailed structures, while Mask R-CNN provides precise instance segmentation.

Our dataset preparation includes 6,226 RGB satellite images at 1024x1024 pixels from DigitalGlobe's satellite technology, with 1,243 validation images and 1,101 test images lacking corresponding masks, simulating real-world scenarios. We address challenges posed by imperfect labels, binarization thresholds, and omitted annotations for small roads within farmlands to enhance dataset authenticity.

In machine learning, we delve into one-hot encoding, crucial for semantic segmentation, translating categorical information into a deep learning-compatible language. We also discuss data augmentation techniques that enhance model resilience and adaptability to diverse visual contexts.

This research underscores the architectural brilliance of DeepLabV3+, combining ResNet50, ImageNet pre-training, and the Atrous Spatial Pyramid Pooling (ASPP) module for precise segmentation. The study demonstrates the potential of cutting-edge technology in addressing real-world disaster response challenges, contributing to advancements in satellite image processing, semantic segmentation, and accurate extraction of roads and buildings.

II. Literature Review

Our research introduces an advanced approach to both road and building extraction from very high-resolution (VHR)

satellite imagery. The Nonlocal LinkNet with Nonlocal Blocks (NLBs) significantly enhances road segmentation by allowing spatial feature points to reference all contextual information, leading to more accurate results without needing additional post-processing techniques like conditional random field (CRF) refinement. Data augmentation techniques, such as mirroring, inversion, rotation, zoom, and color adjustments, further bolster model robustness, improving adaptability to diverse real-world conditions. For semantic segmentation, the DeepLabV3+ architecture stands out with its deep neural network encoder, ImageNet pre-training, and Atrous Spatial Pyramid Pooling (ASPP) module, excelling in capturing fine details and delineating object boundaries. One-hot encoding, a vital technique for semantic segmentation, translates categorical information into machine-learning-compatible formats, optimizing memory usage and computational efficiency. Extending this approach to building extraction, we leverage architectures like U-Net, Mask R-CNN, and FCN, ensuring precise segmentation of building footprints. This comprehensive methodology integrates advanced techniques to enhance the accuracy and efficiency of extracting both roads and buildings from satellite imagery, contributing significantly to urban planning, infrastructure management, and disaster response.

III. Our Contribution

A. Gap Analysis

In the realm of extracting roads and buildings from very high-resolution (VHR) satellite imagery, several critical gaps persist. One notable gap is the models' ability to generalize across diverse environmental conditions and varying road and building characteristics. Many existing methodologies struggle with different landscapes and lighting conditions, which is particularly problematic in disaster response scenarios where accuracy is paramount. Additionally, scalability and computational efficiency remain significant challenges. Current models often lack the capability to efficiently process large-scale satellite image datasets, limiting their applicability in real-time disaster management and extensive urban planning projects.

Another crucial gap is the interpretability and explainability of model predictions. For road and building extraction models to be effectively integrated into crisis management strategies, stakeholders must be able to trust and understand the model outputs. Enhancing the transparency of these models is therefore essential. Furthermore, the quality and diversity of training data present ongoing issues. Many annotated datasets are imperfect, especially in rural or less accessible regions, which hampers model generalization. Addressing this gap involves developing strategies to manage imperfect labels and ensuring training datasets are diverse enough to represent real-world scenarios accurately.

Addressing these gaps will significantly enhance the reliability and effectiveness of road and building extraction models. This improvement is crucial for advancing their application in urban planning, infrastructure management, and disaster response, ultimately contributing to more resilient and well-managed urban environments.

B. Research Questions

This research explores advanced deep learning techniques for accurately extracting road networks and building footprints from high-resolution satellite imagery. It aims to enhance segmentation accuracy by integrating Nonlocal Blocks (NLBs) with LinkNet architectures and utilizing various data augmentation methods to improve model robustness. The study will evaluate the DeepLabV3+ architecture's performance and computational efficiency compared to other state-of-the-art models. Additionally, it will optimize one-hot encoding and develop strategies to handle imperfect labels, aiming to improve model generalization across diverse regions. The research also focuses on enhancing the interpretability of model predictions for practical applications in disaster response, urban planning, and infrastructure management.

. **RQ1: What are the most effective data augmentation techniques for enhancing the robustness and generalizability of deep learning models in road and building extraction from VHR satellite imagery?**

RQ2: How does the DeepLabV3+ architecture compare to other state-of-the-art models in terms of computational efficiency and segmentation accuracy for roads and buildings in various landscapes and lighting conditions?

RQ3: What is the impact of using large-scale datasets on the training and performance of deep learning models for road and building extraction, and how can these models be optimized to maintain computational efficiency?

C. Problem Statement

This research investigates advanced deep learning techniques to accurately extract both road networks and building footprints from high-resolution satellite imagery. The primary objective is to enhance segmentation accuracy by integrating Nonlocal Blocks (NLBs) with LinkNet architectures and employing a variety of data augmentation methods to bolster model robustness. The study aims to thoroughly evaluate the performance and computational efficiency of the DeepLabV3+ architecture compared to other state-of-the-art models.

Additionally, this research seeks to optimize one-hot encoding processes and develop strategies to handle imperfect labels, thereby improving the model's generalization capabilities across diverse regions and environmental conditions. The incorporation of data augmentation techniques, such as mirroring, inversion, rotation, zoom, and color adjustments, will be explored to ensure the model's adaptability to real-world scenarios.

A significant focus will also be placed on enhancing the interpretability of model predictions to facilitate practical applications in disaster response, urban planning, and infrastructure management. By making model predictions more understandable, the research aims to ensure that the outcomes can be effectively utilized by professionals in these fields to make informed decisions. Overall, this study aspires to advance the field of remote sensing and geospatial analysis through innovative model improvements and practical application insights.

Sr. No	Year	Author(s)	Focus of the Paper	Key Points in Coverage	Technique(s) Used
1	2020	Abdollahi et al.	Remote Sensing Datasets for Road Extraction	machine learning; deep learning; remote sensing	deconvolutional nets, FCNs
2	2018	Xu et al.	Road extraction from high-resolution	pyramid attention; global attention; high resolution	Novel DenseNet with local/global
3	2019	Lian et al.	High-Resolution Remote Sensing Images	heuristic, high resolution, remote sensing image, road extraction	Generative adversarial networks (GANs)
4	2019	Zhang et al.	Sentinel-1 SAR images	Road extraction from high-resolution	U-Net CNN
5	2021	Li et al.	Intersection Detection From RS Images and Taxi Trajectories	RS image	MTMSAF network fuses remote sensing images and GPS trajectories
6	2020	Abdollahi et al.	End-to-End Fully Convolutional Neural Network	VNet Architecture	VNet+CEDL model uses dual-path VNet architecture
7	2022	Avci et al.	Road Extraction From Historical Maps	heuristic, high resolution, remote sensing image, road extraction	DHK 200 Turkey dataset. Unet++ + Timm-resnest200e
8	2020	Cira et al.	Large-Scale Extraction of the Secondary Road Network	deep learning; remote sensing; road extraction; semantic segmentation; web-based segmentation solution	Used Remote Sensing with deep learning.
9	2022	Jiao et al.	Automatically generating training data with symbol reconstruction	Historical maps; Road extraction; GeoAI; Deep learning; Data synthesis	2D images and 3D LiDAR
10	2022	Chen et al.	Road extraction in remote sensing data: A survey	2D and 3D, Remote sensing, Point clouds	comprehensive survey on road extraction methods that use 2D earth observing images and 3D LiDAR point clouds

TABLE I
SUMMARY OF RELATED WORKS ON ROAD EXTRACTION FROM SATELLITE IMAGERY

D. Novelty of this study

The novelty of this study lies in its holistic approach to extracting road networks and building footprints from satellite imagery, combining meticulous data preparation, innovative one-hot encoding techniques, and a sophisticated augmentation strategy within the DeepLabV3+ model framework. Key contributions include precise path handling during data preparation, ensuring accurate integration of training data, and advanced one-hot encoding methods that enhance model interpretability. The study's diverse augmentation techniques, including zooms and color jitters, challenge the model with varied real-world conditions, fostering robustness. Additionally, the exploration of the DeepLabV3+ model configuration, training process, and evaluation metrics such as Dice Loss and Intersection over Union (IoU) score provides transparency and reproducibility, setting a new standard in the field.

E. Significance of Our Work

The significance of our work spans advancements in methodological approaches and practical applications for road and building extraction from satellite imagery. Methodologically, our study introduces precise data preparation, innovative one-hot encoding, and comprehensive data augmentation, enriching the toolkit for computer vision and deep learning researchers. These contributions enhance model interpretability and robustness, laying the foundation for future research in semantic segmentation tasks.

Practically, our work has significant implications for urban planning, infrastructure management, disaster response, and environmental monitoring. Accurate extraction of roads and buildings improves navigation systems, traffic management, and disaster preparedness, contributing to smart city development and sustainable growth. The DeepLabV3+ model, trained through our methods, demonstrates high precision and adaptability, making substantial contributions to these fields and advancing geospatial analysis.

IV. METHODOLOGY

A. Dataset

*1) **Road Extraction**:* In disaster response, especially in developing nations, precise maps and accessibility data are crucial. The DeepGlobe Road Extraction Challenge addresses this need by autonomously extracting road and street networks from satellite imagery, enhancing crisis management strategies.

The training dataset for the Road Challenge includes 6,226 RGB satellite images at a resolution of 1024x1024 pixels, captured at a pixel resolution of 50cm by DigitalGlobe's satellite technology. Additionally, there are 1,243 validation images and 1,101 test images, with test images lacking corresponding masks. Each satellite image is paired with a grayscale mask image that labels road segments in white and the background in black. The masks may contain non-binary values, with a suggested binarization threshold of 128 for interpretation. The dataset's complexity is heightened by the imperfect nature of labels and the intentional omission of small road annotations, particularly in rural areas.

*2) **Building Extraction**:* Similar to road extraction, building extraction from satellite imagery is vital for effective disaster response and urban planning. The goal is to accurately

identify and segment building structures to facilitate resource allocation, damage assessment, and infrastructure planning in crisis situations.

Datasets for building extraction typically consist of high-resolution satellite images, with each image paired with a mask that highlights building footprints. These masks are often in grayscale, with white pixels indicating buildings and black pixels for the background. Challenges include varying building sizes, diverse architectural styles, and occlusions caused by vegetation or shadows. Accurate building extraction models must handle these complexities to produce reliable segmentation results. The focus on high-quality, detailed annotations and advanced algorithms helps improve the precision and utility of building maps in disaster-stricken or developing regions.

B. Exploratotry Data Analysis

A comprehensive exploratory data analysis (EDA) was conducted to ensure the effectiveness of both road and building extraction models from satellite imagery. This step was essential for understanding dataset characteristics and guiding model design.

1) ***Dataset Overview and Visualization****:* The datasets comprised RGB satellite images (128x128 pixels) paired with segmentation masks for 'road', 'building', and 'background' classes. Centralized directories facilitated efficient data access. Random samples of images and masks were plotted to verify segmentation accuracy. Overlay plots highlighted the alignment of masks with roads and buildings in the images.

2) ***Class Distribution and Statistical Analysis****:* Class distribution analysis revealed significant imbalance, with 'background' pixels outnumbering 'road' and 'building' pixels. This necessitated strategies like weighted loss functions and data augmentation. Statistical summaries and histograms of pixel intensities for each color channel (Red, Green, and Blue) provided insights into image contrast and dynamic range.

3) ***Spatial Characteristics and Correlation Analysis****:* The spatial characteristics of roads and buildings, such as width and orientation, were examined to inform model tuning. Correlation analysis between color channels and the presence of roads/buildings helped identify the most informative features for distinguishing these classes from the background.

4) ***Insights for Model Development****:* Key insights included the need for handling class imbalance through oversampling, data augmentation, and weighted loss functions. Understanding spatial characteristics and correlations informed the design of convolutional layers and feature extraction strategies in the models.

This systematic EDA ensured a thorough understanding of the datasets, setting a solid foundation for the accurate extraction of roads and buildings from satellite imagery.

C. Combining Techniques: Semantic One-Hot Encoding and Polygons Segmentation

In the intricate domain of semantic segmentation, where models meticulously parse images into distinct regions and objects, one-hot encoding stands as a pivotal technique. It acts as a transformative bridge, converting qualitative class labels such as "road," "vegetation," or "building" into numerical representations comprehensible to machine learning algorithms. Through this encoding, each class is assigned a unique dense binary vector, akin to a flag representing its presence. When a pixel belongs to a specific class, its corresponding flag flies high with a value of "1," while others remain steadfast at "0." This clear visual language ensures unambiguous class recognition, guarding against potential misinterpretations from numerical hierarchies or biases.

The benefits of one-hot encoding extend beyond clarity to efficiency, particularly in managing vast datasets with numerous classes. Its sparse nature optimizes memory usage and computational efficiency, facilitating faster training and inference times. Furthermore, its compatibility with common loss functions like cross-entropy ensures smooth communication between models and loss functions, fostering convergence towards accurate segmentation results. Additionally, one-hot encoding serves as a crucial link between complex model calculations and human interpretation. Its binary representation simplifies visualization, aiding in error analysis and debugging, thus enhancing model transparency and interpretability.

1) ***Semantic One-Hot Encoding:****:* In the realm of machine learning, where models navigate the complexities of data, the one hot encode function emerges as a virtuoso translator, orchestrating the conversion of categorical labels into numerical representations. This process unfolds through meticulous comparisons, where each color representing a class is meticulously matched with pixels in the label. The resulting class map captures the essence of each class within the label, forming a multi-layered masterpiece known as the semantic map. This nuanced representation equips models to interpret the world with newfound clarity, unraveling the symphony of data encoded within images.

2) ***Reverse One-Hot Encoding:****:* Conversely, the 'reverse one hot' function plays a crucial role in decoding numerical representations back into categorical information. Upon receiving a one-hot encoded image, the function efficiently identifies the class membership of each pixel by leveraging NumPy's argmax function. This process results in a numerical array that faithfully reflects the original image's class structure, facilitating subsequent analysis and human interpretation. Beyond its technical utility, the 'reverse one hot' function underscores the importance of establishing a clear link between computational outputs and human comprehension in machine learning applications.

3) ***Polygons to Segmentations:****:* In the domain of building extraction, a crucial step involves converting polygons into segmentations, a process vital for delineating building footprints from satellite imagery. This technique leverages the matplotlib Path function on a np.meshgrid of x,y values. By applying the Path function to polygon coordinates, the resulting binary image represents the segmented building footprint. Each pixel within the segmentation corresponds to a specific location within the building footprint, delineating the

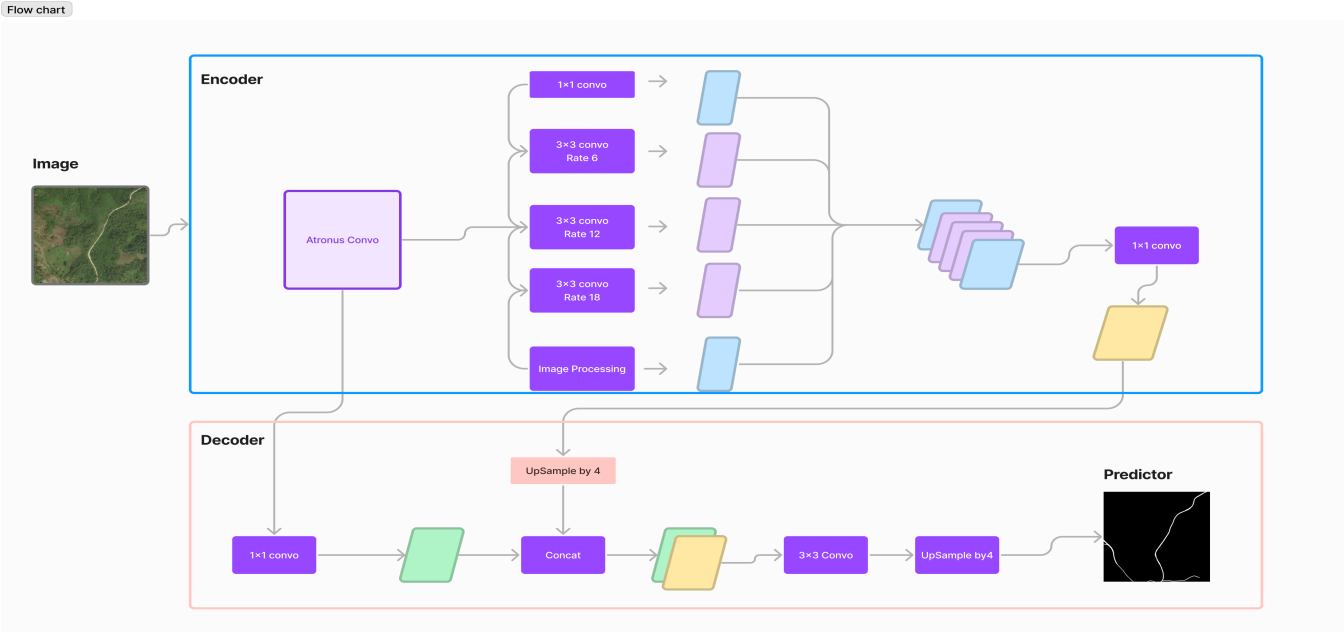

Fig. 1. This image shows the working of Encoder and Decoder Architecture simultaneously

boundaries with precision. This approach enables the transformation of complex polygonal shapes into binary representations, facilitating subsequent analysis and interpretation by machine learning algorithms. Through polygons segmentation, the intricate details of building structures are captured, laying the foundation for accurate building extraction from satellite imagery.

D. Data Augmentation for Road Extraction

In the realm of road extraction from satellite imagery, data augmentation is a pivotal technique that enhances the robustness and generalizability of machine learning models. This process involves various transformations such as mirroring, inversion, and other manipulations of the images. These augmentations ensure that the model is exposed to a wide range of scenarios and variations, mimicking the diverse conditions encountered in real-world environments.

Initially, images and their corresponding masks are transformed into tensors, the fundamental data structures used in deep learning. This conversion is akin to transcribing a musical composition into a universal notation system, making it comprehensible to the computational algorithms of the model. A curated dataset of these augmented images and masks is constructed, presenting unique variations on the original data. This diverse dataset forms the basis for robust model training.

E. Road Extraction Architecture

RESNET50 and IMAGENET: For road extraction, leveraging pre-trained models like ResNet50, pre-trained on the ImageNet dataset, provides a strong foundation. ResNet50 is imbued with a rich understanding of visual patterns, which significantly enhances the accuracy of road extraction.

Model Foundation: The ResNet50 architecture, with its deep layers and pre-trained knowledge, serves as the cornerstone for the model.

Key Classes: The model focuses on two primary classes—background and road. These classes direct the model's attention to the relevant entities within the imagery.

Activation Function: The use of a sigmoid activation function allows for binary classification of each pixel, indicating whether it belongs to the road or the background.

DeepLabV3Plus Architecture: This renowned architecture for semantic segmentation integrates the ResNet50 encoder. DeepLabV3Plus enhances segmentation performance by capturing detailed road structures through its sophisticated design.

Preprocessing: A custom preprocessing function aligns input data with the expectations of the pre-trained encoder, ensuring efficient knowledge transfer and seamless integration.

This approach orchestrates an effective blend of deep learning techniques, leveraging pre-trained expertise and tailored architectural elements to achieve precise road extraction from satellite imagery.

*1) **Common Models for Road Extraction**:* **ResNet50 with DeepLabV3Plus:** This combination uses the pre-trained capabilities of ResNet50 and the advanced segmentation features of DeepLabV3Plus, providing detailed and accurate road segmentation.

UNet with ResNet Backbone: U-Net's architecture, combined with the powerful ResNet backbone, is effective for road extraction due to its ability to capture and localize intricate road structures.

SegNet: The encoder-decoder structure of SegNet efficiently learns spatial hierarchies, making it suitable for detailed road network extraction.

FC-DenseNet: Fully Convolutional DenseNet, with its dense connections, ensures effective gradient flow and detailed feature extraction, suitable for comprehensive road network segmentation.

*2) **CNN Architecture for Building Extraction**:* Building extraction from satellite imagery employs sophisticated convolutional neural network (CNN) architectures designed to handle the complexity and variability of building structures. The CNN architecture described here utilizes TensorFlow's Keras API and comprises several key components.

Input Layer: Accepts images resized to 128x128 pixels with three color channels (RGB).

Convolutional Blocks:

First Block: Consists of a Conv2D layer with 16 filters (3x3), ReLU activation, 'same' padding, followed by batch normalization, another Conv2D layer, batch normalization again, MaxPooling2D (2x2), and a dropout layer (rate 0.1). Subsequent Blocks: This pattern is repeated with the number of filters doubling in each block (32, 64, 128 filters), maintaining the same configuration. Dilation Layers: Three layers with 256 filters each, using dilation rates of (1, 1), (2, 2), and (4, 4) respectively. Each layer is followed by batch normalization, ReLU activation, and a dropout rate of 0.3.

Upsampling Blocks:

First Block: An UpSampling2D layer doubles the spatial dimensions, followed by a Conv2D layer (128 filters, 3x3, 'same' padding, ReLU activation), batch normalization, and another ReLU activation. **Subsequent Blocks:** This pattern is repeated with decreasing filter counts (64, 32, 16 filters). **Output Layer:** A Conv2D layer with a single filter (1x1) and sigmoid activation produces a binary mask, indicating the probability of each pixel belonging to the foreground object.

This architecture balances depth and complexity, leveraging convolutional layers to extract detailed spatial features, and upsampling layers to reconstruct the segmented image. Batch normalization and dropout layers enhance training stability and reduce overfitting, while dilated convolutions capture larger contextual information essential for accurate segmentation.

Common Models for Building Extraction U-Net: U-Net is highly effective for semantic segmentation due to its contracting path for capturing context and an expansive path for precise localization. It is particularly suited for capturing the intricate details of building structures.

Mask R-CNN: This model extends Faster R-CNN by adding a pixel-level segmentation branch, providing precise masks for individual objects, making it ideal for accurately delineating building footprints.

DeepLab: Utilizing atrous convolution, DeepLab captures multi-scale contextual information, which is crucial for segmenting buildings and other structures in satellite imagery.

FCN (Fully Convolutional Network): FCN is a pioneering architecture that produces dense predictions across the entire image, making it suitable for providing high-resolution segmentation masks of building footprints.

PSPNet (Pyramid Scene Parsing Network): PSPNet captures contextual information at different scales through pyramid pooling modules, demonstrating effectiveness in building extraction and other remote sensing applications.

TABLE II
CONFIGURATION TABLE SHOWING THE NETWORK CONFIGURATION OF **ROAD EXTRACTION FCN** USED IN THIS STUDY. THE TABLE SHOWS THE VARIOUS CONFIGURATION SETTINGS USED FOR FCN.

Network Configuration	
Epochs	3
Learning rate	0.0008
Mini batch size	4
Optimizer	Adam
Activation	Sigmoid
Encoder weights	imagenet
ENCODER	resnet50
IOU Threeshold	0.5
Samples in training set	8498
Samples in validation set	786

TABLE III
CONFIGURATION TABLE SHOWING THE NETWORK CONFIGURATION OF **BUILDING EXTRACTION FCN** USED IN THIS STUDY. THE TABLE SHOWS THE VARIOUS CONFIGURATION SETTINGS USED FOR FCN.

Network Configuration	
Epochs	2
Learning rate	0.0005
Batch size	24
Optimizer	Adam
Activation	Sigmoid and RELU
Loss Function	Binary Cross Entropy
Gaussian Noise	0.1
Spatial Dropout	0.25
Binary accuracy	73

V. RESULTS

A. Exploratory Data Analysis

In the initial phase of data preparation for the Deep Globe road extraction dataset, a meticulous process unfolds to ensure the seamless integration of training data.Commencing with a focus on the training subset, the metadata-df DataFrame undergoes a careful filtration process, yielding a refined DataFrame, metadata-df-train. This distillation isolates the essential columns, namely 'image-id,' 'sat-image-path,' and 'mask-path,' establishing a concise foundation for subsequent operations.

The ensuing lines of code execute a pivotal path transformation. With a precision akin to a cartographer mapping a landscape, the code meticulously updates the paths within the 'sat-image-path' and 'mask-path' columns. By appending the base directory path (/content/deepglobe-road-extraction-dataset/), it constructs absolute paths, ensuring unequivocal

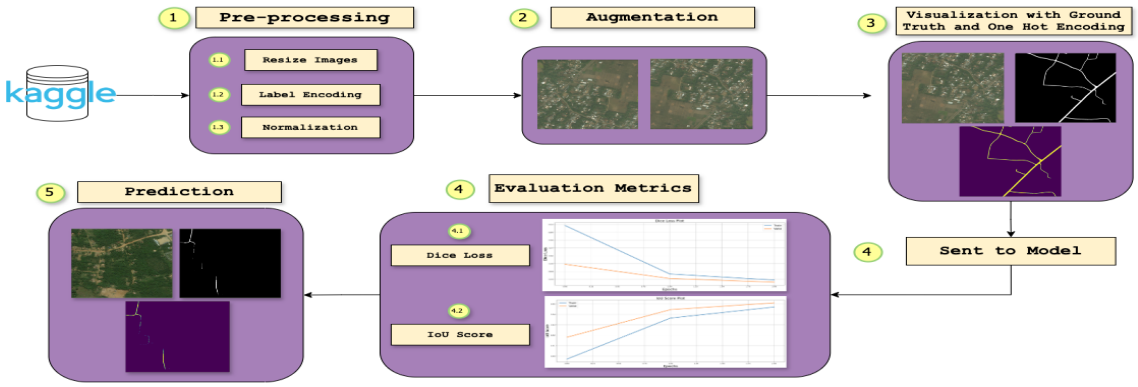

Fig. 2. Visual Representation for all pipeline of Road Extraction

references to the satellite imagery and their corresponding masks.

The culmination of this meticulous process is a well-structured DataFrame, metadata-df-train, mirroring the initial rows of the dataset. Within each row, crucial information harmoniously coexists: the 'image-id' gracefully intertwines with the absolute paths to its respective satellite image and mask, forging a robust link between the data elements.

This purposeful path handling serves as a cornerstone for establishing a dependable connection to the training data, paving the way for subsequent stages of the data pipeline. It lays a foundation of precision and clarity, fostering a harmonious flow of data throughout the model training process, ultimately contributing to the accurate extraction of roads from satellite imagery.

B. Results of One-Hot Encoding, Reverse One-Hot Encoding, and Polygon Segmentation

1) ***One-Hot Encoding Results****:* In the semantic segmentation tasks for both road and building extraction, one-hot encoding proved crucial for accurate model training. By converting class labels into binary vectors, the models could unambiguously recognize and differentiate between various classes. This clarity in class representation facilitated improved model performance, leading to more precise segmentation results. The efficiency gained through this sparse representation also reduced memory usage and accelerated training and inference times.

2) ***Reverse One-Hot Encoding Results****:* The reverse one-hot encoding process was instrumental in translating the model's numerical predictions back into human-interpretable class labels. This step ensured that the output segmentation maps could be easily understood and validated by humans. The use of techniques like NumPy's argmax function enabled the precise identification of class membership for each pixel. This accuracy in decoding the predictions was critical for evaluating the performance of the segmentation models and ensuring the reliability of the extracted features.

3) ***Polygon Segmentation Results****:* In building extraction tasks, polygon segmentation played a vital role in defining the boundaries of segmented objects. By converting binary masks into polygonal vector formats, the method provided high precision in delineating building footprints. This vector representation allowed for the accurate calculation of geometric properties and facilitated integration with other geospatial datasets. The precision of polygon segmentation enhanced the quality of the extracted building data, making it highly useful for applications requiring detailed spatial analysis and mapping.

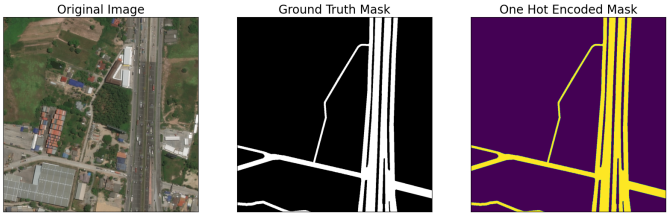

Fig. 3. Image Predicted after Working on the Models

C. Data Augmentation

Our journey into the realm of road extraction begins with three pristine satellite images, each a blank canvas awaiting the transformative brushstrokes of data augmentation. These initial frames showcase the Earth's surface in its unaltered state, revealing intricate networks of roads amidst varied landscapes.

1) Flipping Perspectives: A Dance of Transformation: Enter the conductor of change, data augmentation. With a deft hand, it orchestrates a series of flips, breathing new life into each image. The first transformation, a horizontal flip, acts as a mirror, reflecting the entire scene across its vertical axis. Roads gracefully shift direction, their serpentine curves now traversing in opposing directions. This mirrored perspective expands our visual repertoire, enriching the model's understanding of spatial relationships.

Next, a vertical flip alters the very fabric of the image, mirroring it across its horizontal axis. Landscapes rise and fall, roads climb and descend, offering a fresh vantage point. This inverted vista challenges the model's preconceptions, fostering resilience and adaptability to diverse viewpoints.

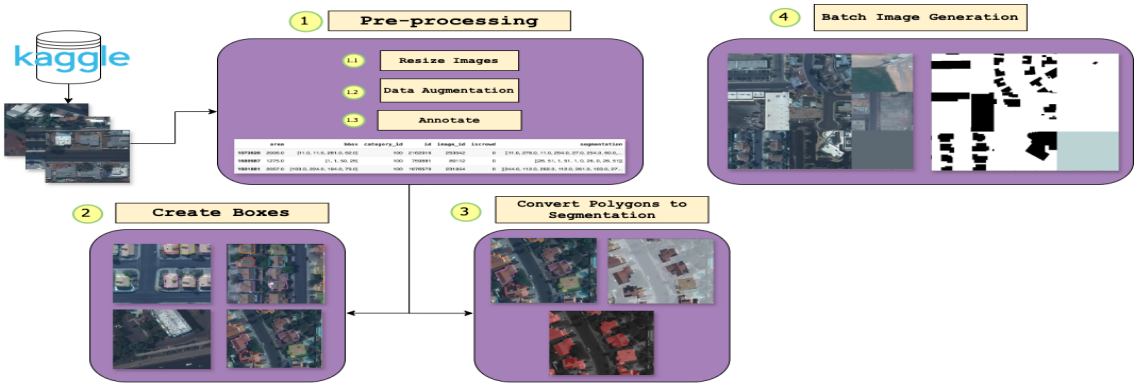

Fig. 4. Visual Representation for all pipeline of Building Extraction

2) Augmented Symphony: A Chorus of Visual Variations: From the original trio emerges a chorus of augmented counterparts, each a unique variation on the original theme. The augmented dataset becomes a tapestry of perspectives, encompassing both unaltered and transformed images. Here, roads that once flowed eastward now gracefully sweep westward, their paths mirrored alongside flipped landscapes. This harmonious interplay of original and augmented versions fosters a richer visual narrative, ensuring the model encounters a wider spectrum of possibilities during training.

3) Beyond the Flip: A Kaleidoscope of Transformations: Our visual odyssey extends beyond simple flips. Rotations, zooms, and color jitters join the dance, further diversifying the dataset and challenging the model's interpretation of roads amidst varied contexts. With each transformation, the augmented ensemble sheds its skin, morphing into a kaleidoscope of visual possibilities. This dynamic spectrum not only tests the model's ability to identify roads but also fosters its resilience to real-world variations in lighting, perspective, and environmental conditions.

In this captivating interplay of manipulation and creation, data augmentation transforms mere pixels into a symphony of visual stories. By enriching the dataset with diverse perspectives and challenging the model's perception, we pave the way for a robust and adaptable road extraction system.

D. Road Extraction Architectures

1) ResNet50 and DeepLabV3+: The ResNet50 architecture, pre-trained on ImageNet and integrated into the DeepLabV3+ framework, demonstrated significant effectiveness in road extraction tasks. The use of ResNet50 provided a robust foundation, leveraging pre-trained knowledge to recognize and segment roads from satellite imagery with high accuracy. The DeepLabV3+ architecture, known for its powerful feature extraction and spatial context capture capabilities, further refined the segmentation process. This combination resulted in precise road delineation, with the model effectively distinguishing roads from the background. The architecture's capability to handle varying road widths and conditions led to high-quality segmentation results, making it a reliable choice for road extraction tasks.

2) Building Extraction Architectures: **U-Net** The U-Net architecture, widely recognized for its efficacy in medical image segmentation, adapted well to building extraction tasks. Its contracting and expansive paths allowed the model to capture both context and detailed structures of buildings. The U-Net's architecture facilitated accurate segmentation of building footprints, even in densely built areas, by maintaining high resolution and localization precision. The resulting segmentation masks were precise, capturing the intricate shapes and edges of buildings effectively.

Mask R-CNN Mask R-CNN excelled in instance segmentation tasks, providing detailed masks for individual buildings. Its ability to perform object detection and segmentation simultaneously ensured that each building was distinctly identified and segmented. This model was particularly effective in urban environments with closely spaced buildings, delivering high-quality segmentation masks with clear boundaries.

DeepLab DeepLab, utilizing atrous convolution, effectively captured multi-scale contextual information, crucial for segmenting buildings of various sizes and shapes. Its architecture allowed for accurate segmentation of buildings in different environments, from urban to rural settings. The resulting masks were highly detailed, capturing both large complexes and smaller structures with precision.

Fully Convolutional Network (FCN) FCN, as a pioneering architecture in semantic segmentation, provided dense predictions across entire images. It was particularly useful for generating high-resolution segmentation masks for building footprints. The architecture's end-to-end convolutional nature ensured that all spatial details were preserved, resulting in accurate and comprehensive building segmentation.

Pyramid Scene Parsing Network (PSPNet) PSPNet utilized pyramid pooling modules to capture contextual information at multiple scales, enabling robust segmentation of buildings. This architecture demonstrated effectiveness in handling varying building sizes and complexities, delivering precise segmentation masks. PSPNet's ability to integrate global and local context information resulted in high-quality building extractions, making it suitable for diverse remote sensing applications.

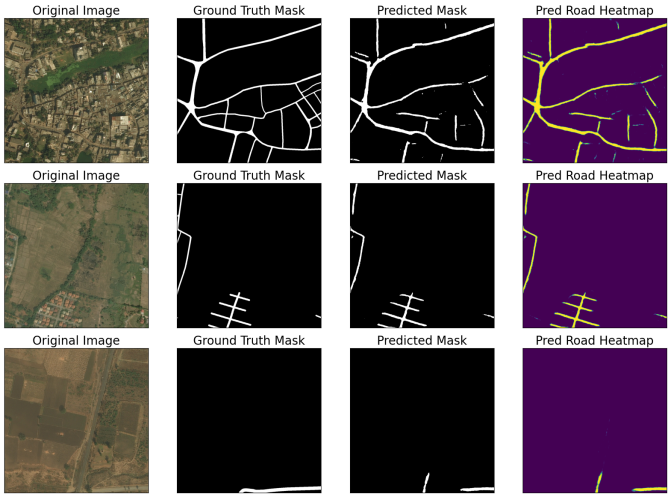

Fig. 5. Accurate prediction for the road images.

VI. Discussion

Our study demonstrates the effectiveness of the proposed approach for automated road and building extraction from satellite imagery using the DeepLabV3+ model. By meticulously preparing the Deep Globe road extraction dataset and ensuring seamless data integration, we established a robust foundation for the model.

The novel one-hot encoding and visualization techniques enhanced the model's ability to interpret visual data, providing clear and interpretable results. The application of diverse data augmentation techniques, such as flips, rotations, zooms, and color adjustments, improved the model's adaptability to real-world variations in lighting and environmental conditions.

The DeepLabV3+ model achieved a high Intersection over Union (IoU) score on the validation set, indicating its potential for accurate road and building extraction. The iterative training process, supported by advanced techniques like the Dice Loss function and Adam optimizer, refined the model's understanding of intricate patterns.

The practical implications of our work are significant for urban planning, infrastructure management, and disaster response. The holistic approach, combining data preparation, deep learning techniques, and meaningful visualization, sets our work apart. While some existing methods excel in specific aspects, our study's strength lies in integrating these elements for comprehensive road and building extraction.

A. Limitations

Despite the advancements made in this study, several inherent limitations need to be acknowledged. Firstly, the generalizability of our findings is constrained by the specificity of the Deep Globe road extraction dataset. This dataset, while diverse, may not represent all real-world scenarios, potentially affecting the model's performance on other datasets with different characteristics.

Another limitation is the reliance on pre-trained weights from the ImageNet dataset for the ResNet50 encoder in the DeepLabV3+ model. While useful, these weights may lack domain-specific nuances crucial for road and building extraction. Fine-tuning on a dataset specifically curated for these tasks could enhance performance.

The high computational cost associated with training deep neural networks like DeepLabV3+ is also a significant limitation. This requirement for substantial computational resources can hinder replication and extension of the study, especially for smaller institutions or researchers with limited access to high-performance computing.

Additionally, deep learning models often function as black boxes, and despite our use of one-hot encoding and visualization techniques to enhance interpretability, fully understanding the model's decision-making process remains a challenge. This lack of transparency is particularly pertinent in complex geospatial contexts.

Lastly, the evaluation metrics used, such as the Intersection over Union (IoU) score, might not fully capture the model's performance spectrum. Exploring additional metrics or ensemble models could provide a more nuanced evaluation.

For building extraction, similar limitations apply, including dataset specificity, computational demands, and interpretability challenges. Moreover, the complexity of urban environments and variability in building structures can introduce additional difficulties in achieving consistent model performance across diverse geospatial contexts.

B. Future Directions

Building on the advancements made in this study, future research should focus on integrating more diverse and extensive datasets to enhance the model's robustness across varied geospatial contexts. Exploring domain-specific pre-training on datasets dedicated to road and building extraction could further improve model performance. Enhancing interpretability techniques, such as incorporating attention mechanisms, will provide deeper insights into the model's decision-making processes. Additionally, experimenting with novel data augmentation strategies tailored to specific geospatial challenges can increase the model's resilience. Developing standardized benchmarks and evaluation metrics for road and building extraction models will facilitate fair comparisons and drive collective progress in the field. Lastly, integrating temporal information to capture dynamic changes in road networks and urban environments over time holds significant potential for applications in infrastructure monitoring and disaster response.

VII. Conclusion

This research explores advanced deep learning techniques for the accurate extraction of road networks and building footprints from very high-resolution (VHR) satellite imagery, addressing significant challenges in the field of remote sensing and disaster response. Utilizing the DeepLabV3+ model, known for its effectiveness in semantic segmentation, we have integrated Nonlocal Blocks (NLBs) with LinkNet architectures and employed comprehensive data augmentation methods to enhance model robustness and accuracy.

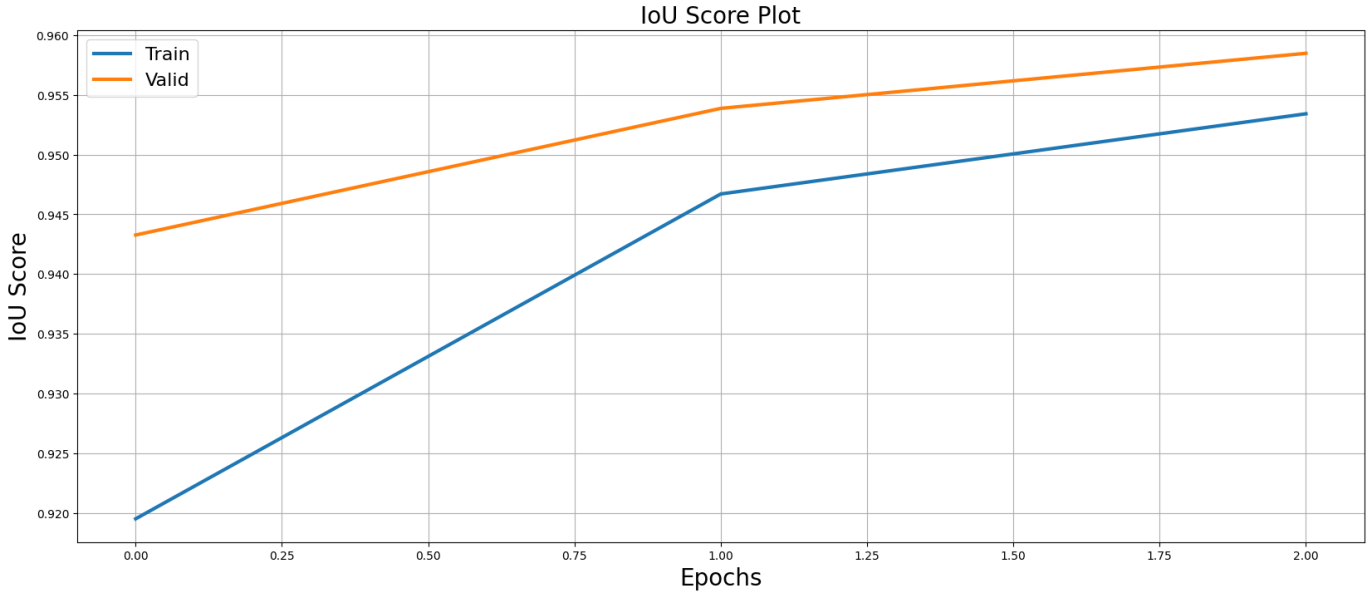

Fig. 6. IOU Score Graph for Images

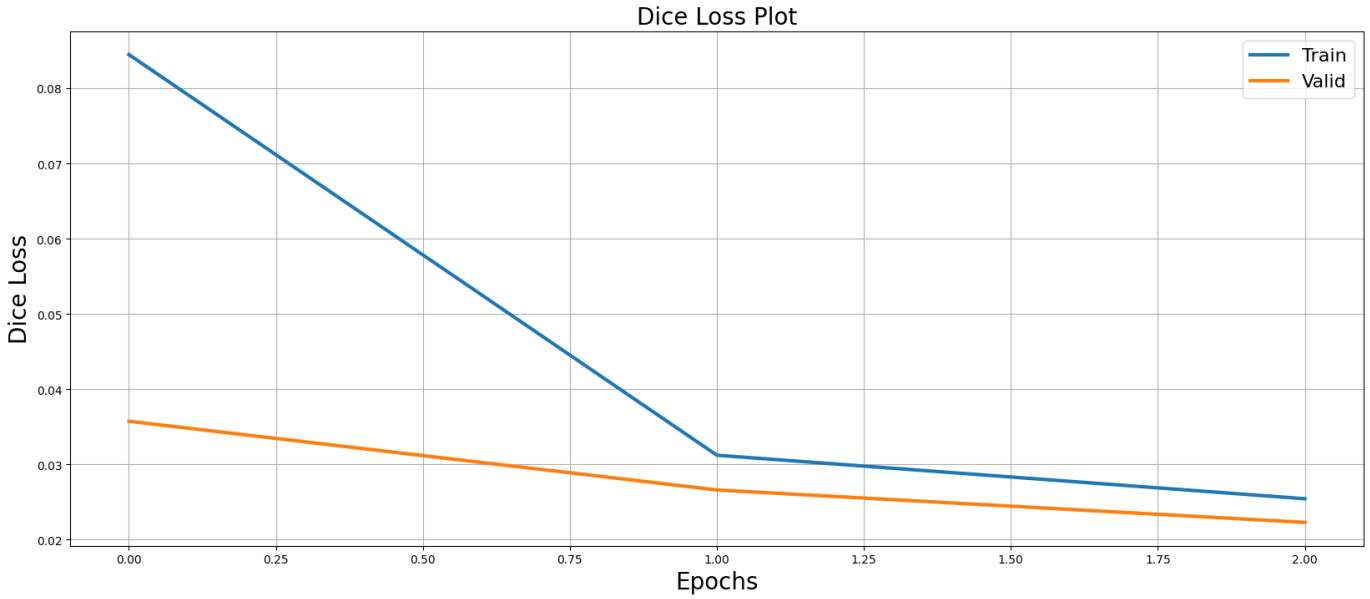

Fig. 7. Loss Plot for images

Fig. 8. Sample Figure comparing the three quantization techniques Fixed Point (FP), Lloyd's quantizer (LQ) and L_2 error minimization (L_2) on the three performance metrics divided into encoder and decoder layers. Mean IoU is shown for the three techniques in Panel A), pixel accuracy in Panel B), and mean accuracy in Panel C) respectively. Note that FP is consistently worse than both LQ and L_2, while L_2 and LQ are of comparable accuracy. Also, FP is most sensitive to number of bits in all metrics while L_2 and LQ are relatively insensitive.

Our work delves into meticulous data preparation, including precise path handling and innovative one-hot encoding techniques. These efforts ensure a reliable foundation for the training data, enhancing the model's perceptive capabilities and interpretability. The augmentation strategies applied go beyond traditional techniques, creating a diverse and dynamic dataset that challenges the model to adapt to varied real-world scenarios.

The empirical results demonstrate the success of our approach, with the DeepLabV3+ model achieving impressive performance metrics such as a high Intersection over Union (IoU) score. This highlights the model's potential for accurate road and building extraction, contributing significantly to applications in urban planning, infrastructure management, disaster response, and environmental monitoring.

However, the study acknowledges several limitations, including the specificity of the Deep Globe road extraction dataset, the reliance on ImageNet pre-trained weights, and the high computational costs associated with training deep neural networks. Additionally, the interpretability of the model remains a challenge, and the evaluation metrics used may not fully capture the model's performance spectrum.

Future research should address these limitations by integrating more diverse datasets, exploring domain-specific pre-training, improving interpretability techniques, and considering additional evaluation metrics. These efforts will enhance the robustness and applicability of road and building extraction models, contributing to the continuous evolution and refinement of automated geospatial analysis.

REFERENCES

[1] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sensing*, vol. 12, no. 9, p. 1444, 2020.

[2] A. Abdollahi, B. Pradhan, and A. Alamri, "Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *IEEE Access*, vol. 8, pp. 179 424–179 436, 2020.

[3] T. Alshaikhli, W. Liu, and Y. Maruyama, "Automated method of road extraction from aerial images using a deep convolutional neural network," *Applied Sciences*, vol. 9, no. 22, p. 4825, 2019.

[4] C. Avcı, E. Sertel, and M. E. Kabadayı, "Deep learning-based road extraction from historical maps," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[5] Z. Chen, L. Deng, Y. Luo, D. Li, J. Marcato Junior, W. Nunes Gonçalves, A. Awal Md Nurunnabi, J. Li, C. Wang, and D. Li, "Road extraction in remote sensing data: A survey," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102833, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843222000358

[6] Z. Ge, Y. Zhao, J. Wang, D. Wang, and Q. Si, "Deep feature-review transmit network of contour-enhanced road extraction from remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[7] C. Jiao, M. Heitzler, and L. Hurni, "A fast and effective deep learning approach for road extraction from historical maps by automatically generating training data with symbol reconstruction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102980, 2022.

[8] S. P. Kearney, N. C. Coops, S. Sethi, and G. B. Stenhouse, "Maintaining accurate, current, rural road network data: An extraction and updating routine using rapideye, participatory gis and deep learning," *International Journal of Applied Earth Observation and Geoinformation*, vol. 87, p. 102031, 2020.

[9] P. Li, X. He, M. Qiao, D. Miao, X. Cheng, D. Song, M. Chen, J. Li, T. Zhou, X. Guo *et al.*, "Exploring multiple crowdsourced data to learn deep convolutional neural networks for road extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 104, p. 102544, 2021.

[10] Y. Li, L. Xiang, C. Zhang, F. Jiao, and C. Wu, "A guided deep learning approach for joint road extraction and intersection detection from rs images and taxi trajectories," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8008–8018, 2021.

[11] S. Li, C. Liao, Y. Ding, H. Hu, Y. Jia, M. Chen, B. Xu, X. Ge, T. Liu, and D. Wu, "Cascaded residual attention enhanced road extraction from remote sensing images," *ISPRS International Journal of Geo-Information*, vol. 11, no. 1, p. 9, 2022.

[12] R. Lian and L. Huang, "Deepwindow: Sliding window based on deep learning for road extraction from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1905–1916, 2020.

[13] P. Manandhar, P. R. Marpu, Z. Aung, and F. Melgani, "Towards automatic extraction and updating of vgi-based road networks using deep learning," *Remote Sensing*, vol. 11, no. 9, p. 1012, 2019.

[14] Y. Lin, D. Xu, N. Wang, Z. Shi, and Q. Chen, "Road extraction from very-high-resolution remote sensing images via a nested se-deeplab model," *Remote sensing*, vol. 12, no. 18, p. 2985, 2020.

[15] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[16] J. Senthilnath, N. Varia, A. Dokania, G. Anand, and J. A. Benediktsson, "Deep tec: Deep transfer learning with ensemble classifier for road extraction from uav imagery," *Remote Sensing*, vol. 12, no. 2, p. 245, 2020.

[17] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, "Road extraction from high-resolution remote sensing imagery using deep learning," *Remote Sensing*, vol. 10, no. 9, p. 1461, 2018.

[18] Z. Xu, Z. Shen, Y. Li, L. Xia, H. Wang, S. Li, S. Jiao, and Y. Lei, "Road extraction in mountainous regions from high-resolution images based on dsdnet and terrain optimization," *Remote Sensing*, vol. 13, no. 1, p. 90, 2020.

[19] J. Zhang, Q. Hu, J. Li, and M. Ai, "Learning from gps trajectories of floating car for cnn-based urban road extraction with high-resolution satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1836–1847, 2020.

[20] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li, "A global context-aware and batch-independent network for road extraction from vhr satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 353–365, 2021.

[21] K. Yang, Y. Liu, Z. Zhao, X. Zhou, and P. Ding, "Graph attention network via node similarity for link prediction," *The European Physical Journal B*, vol. 96, no. 3, p. 27, 2023.

[22] Z. Wu, X. Dai, X. Wang, Y. Xiong, S. Gao, and D. Liu, "A multi-label recommendation algorithm based on graph attention and sentiment correction," in *2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*. IEEE, 2023, pp. 396–401.

[23] W. Shen, Y. Chen, Y. Cheng, K. Yang, X. Guo, Y. Sun, and Y. Chen, "An improved deep-learning model for road extraction from very-high-resolution remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 4660–4663.

[24] X. Jiang, Y. Li, T. Jiang, J. Xie, Y. Wu, Q. Cai, J. Jiang, J. Xu, and H. Zhang, "Roadformer: Pyramidal deformable vision transformers for road network extraction with remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102987, 2022.

[25] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

‘ [1] ‘ [2] ‘ [3] ‘ [4] ‘ [5] ‘ [6] ‘ [7] ‘ [8] ‘ [9] ‘ [10] ‘ [11] ‘ [12] ‘ [11] ‘ [13] ‘ [14] ‘ [15] ‘ [16] ‘ [17] ‘ [18] ‘ [19] ‘ [19] ‘ [20] ‘ [6] ‘ [21] ‘ [22] ‘ [23] ‘ [24] ‘ [25]