University of Tehran

Earth Observation and Geomatics Engineering

Online ISSN: 2588-4360

Homepage: https://eoge.ut.ac.ir/

Road Segmentation in High-Resolution Overhead Imagery Using Deep Encoder–Decoder Architecture

Mohammadreza Tavakoli[1]✉ , **Abbas Abedini**[2]

1. Corresponding author, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran.
   E-mail: mohammadtavakoli@ut.ac.ir
2. School of Surveying and Geospatial Engineering, College of Engineering University of Tehran, Tehran, Iran.
   E-mail: aabedeni@ut.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 2026-02-19
Received in revised form 2026-05-25
Accepted 2026-05-26
Available online 2026-06-02

Keywords:
extraction;
image;
Semantic;
U-Net;
Pyramid Network.

ABSTRACT

This study develops and comparatively evaluates deep-learning-based road segmentation approaches for the automated extraction of road networks from high-resolution overhead imagery, motivated by the need for scalable road-map updating and reliable mask generation for downstream GIS workflows. Experiments were conducted on the Massachusetts Roads Dataset, which comprises 1,171 aerial image tiles of size 1500 × 1500 pixels at approximately 1 m spatial resolution. Three deep encoder–decoder segmentation architectures—U-Net, a U-shaped convolutional encoder–decoder network; Feature Pyramid Network (FPN); and Multi-Scale Attention Network (MA-Net)—were assessed under a controlled experimental setup in which all models used EfficientNet-B7 as the shared convolutional backbone. Performance was evaluated using pixel-level metrics, including Accuracy, Intersection over Union (IoU), F1-score, Precision, and Recall. Quantitative results indicate that U-Net achieves the strongest overall performance (Accuracy = 0.9۷, IoU = 0.89, F1 = 0.94, Precision = 0.9۹, Recall = 0.92), followed by FPN and MA-Net.Visual comparisons show that all methods successfully recover the dominant road layout, while performance differences are more pronounced in thin streets, intersections, and visually complex regions. Error-map analysis further reveals that disagreements are concentrated around connectivity-critical structures such as junctions and narrow links, with MA-Net exhibiting the most widespread error patterns and U-Net demonstrating more spatially localized discrepancies.

Cite this article: Tavakoli, M., Abedini, A. (2025). Road Segmentation in High-Resolution Overhead Imagery Using Deep Encoder–Decoder Architecture, *Earth Observation and Geomatics Engineering*, Volume 9, Issue 2, Pages 118-131. http//doi.org/ 10.22059/eoge.2026.411573.1219

Publisher: University of Tehran.

DOI: http//doi.org/ 10.22059/eoge.2026.411573.1219

1. Introduction

Road networks constitute a primary layer of spatial infrastructure underpinning contemporary societies. Their geometry and connectivity directly influence mobility, logistics, emergency response, and public-service efficiency, while also acting as a proxy for socio-economic activity and urban growth patterns. Accurate, up-to-date road maps are therefore essential for transportation planning, navigation and routing, infrastructure maintenance, land administration, and rapid situational awareness during crises. However, road networks are dynamic: new roads are constructed, existing corridors are widened or re-aligned, and temporary obstructions may appear following natural hazards or conflict. Maintaining current road inventories through conventional field surveys and manual digitization is costly, time-consuming, and often infeasible at regional to national scales—particularly in areas with limited institutional capacity or in environments where access is restricted. These constraints establish a clear necessity for automated, scalable, and reliable road extraction methods.

High-resolution satellite and aerial remote sensing (Bastani et al., 2018, Zhou et al., 2018) provides a uniquely suitable observational basis for addressing this need. Contemporary very-high-resolution (VHR) satellite imagery captures detailed spatial structure over large extents with consistent acquisition geometry, frequent revisit intervals, and broad geographic availability. In parallel, modern aerial imagery (including airborne and UAV-based platforms) offers sub-meter detail, flexible task-driven acquisition, and the ability to collect imagery under specific operational constraints (e.g., targeted corridors, rapid post-disaster mapping). Compared with ground-based collection, overhead data can be obtained rapidly and safely without physical access to the area of interest. Compared with many aerial campaigns, satellite acquisitions can offer more systematic coverage, easier multi-temporal monitoring, and reduced logistical burden for repeated mapping; conversely, aerial data can provide higher spatial fidelity and improved delineation of narrow road structures in localized areas. For road segmentation specifically, high spatial resolution increases the likelihood that narrow or secondary roads occupy a sufficient number of pixels to be detectable, while multi-spectral information (often available in satellite products) can strengthen discrimination between road surfaces and visually similar classes such as rooftops, bare soil, or concrete platforms. Consequently, satellite and aerial imagery are well-positioned to support timely road network updates and to enable consistent monitoring of road evolution over time(Canny and intelligence, 2009, Lafferty et al., 2001, Krähenbühl and Koltun, 2011).

The process of road identification from remote sensing imagery has evolved substantially over the past decades. Early approaches relied heavily on human photointerpretation, where analysts delineated road centerlines or polygons using visual cues such as linearity, contrast, and contextual relationships. Although manual interpretation can produce high-quality cartographic products, it is intrinsically limited by labor intensity, subjectivity, and poor scalability. To reduce dependence on manual work, classical automated and semi-automated techniques were introduced, including edge detection, thresholding, morphological operations, Hough-transform-based line detection, active contours, template matching, and region growing. Later, probabilistic and optimization-based formulations—such as Markov random fields, conditional random fields, and graph-based methods—attempted to incorporate spatial smoothness and contextual constraints. While these methods established important conceptual foundations, they typically depend on hand-crafted features and assumptions about road appearance (e.g., consistent width, homogeneous texture, strong contrast with surroundings). In practice, these assumptions are frequently violated in overhead scenes due to illumination variation, shadows, seasonal change, occlusion by trees or vehicles, heterogeneous paving materials, and complex backgrounds in dense urban areas.

Subsequent advances introduced machine learning strategies that used engineered descriptors (e.g., texture and gradient statistics) together with conventional classifiers. These approaches improved adaptability relative to purely rule-based pipelines, yet their performance remained constrained by feature design, limited representation capacity, and sensitivity to domain shifts between geographic regions, sensors, and acquisition conditions. The emergence of deep learning, particularly convolutional neural networks (CNNs) designed for dense prediction, marked a major shift in road segmentation research. Fully convolutional architectures enabled end-to-end learning of hierarchical features directly from data, supporting pixel-wise classification and the delineation of complex spatial patterns. Encoder–decoder networks, multi-scale context aggregation, and attention mechanisms further improved performance on thin, elongated structures such as roads by combining local edge detail with broader semantic context. More recently, transformer-based components and hybrid models have been explored to capture long-range dependencies and enhance global coherence—capabilities that are especially relevant for maintaining road continuity across occluded or low-contrast segments.

Despite this progress, automated road segmentation remains a challenging task, and these challenges motivate continued research. Roads often occupy a small fraction of pixels relative to background classes, creating strong class imbalance that can degrade learning if not addressed. Their geometry is thin and topologically constrained: a visually plausible segmentation that contains gaps can be operationally inadequate because routing and network analysis require connectivity. Furthermore, road appearance varies widely across rural and urban environments, between paved and unpaved surfaces, and under different imaging conditions. Shadows from buildings and vegetation, occlusions by tree canopies, and confusion with other linear

features (e.g., rivers, railways, building edges) can produce both omission and commission errors. High-resolution overhead imagery, while rich in detail, also increases within-class variability and introduces fine-grained clutter that can mislead models without robust contextual reasoning. These factors underscore the necessity of segmentation approaches that are not only accurate in terms of pixel-wise overlap, but also reliable in preserving continuity and supporting downstream vector-based road network extraction.

Within this landscape, the advantages of combining high-resolution satellite and aerial imagery with deep learning methods over classical approaches are substantial. Overhead imagery offers scalable, repeatable, and wide-area observation with a spatial granularity that can resolve many road types; deep learning provides data-driven feature learning capable of modeling complex, non-linear relationships between spectral–spatial patterns and semantic classes. Unlike classical methods that rely on fixed thresholds, handcrafted filters, or narrowly tuned heuristics, deep models can learn invariances to illumination changes, background heterogeneity, and partial occlusion when trained on representative data. Moreover, modern architectures integrate multi-scale information—an essential property for roads, which require both fine boundary precision (to capture narrow segments) and broader contextual cues (to disambiguate roads from similar surfaces). Deep learning pipelines can also be designed to support operational requirements, such as efficient inference over large mosaics, adaptability via transfer learning to new regions or sensors, and systematic evaluation using standardized metrics.

Accordingly, the present study is situated at the intersection of remote sensing and modern semantic segmentation, focusing on road extraction from high-resolution overhead imagery (satellite and aerial) using deep learning. The overarching objective is to develop a method that produces precise and robust road masks suitable for practical applications, while addressing persistent issues such as discontinuities, false positives in complex urban textures, and generalization across heterogeneous landscapes. In this context, the study emphasizes (i) the necessity of automated road mapping for scalable and timely geospatial information production; (ii) the functional value of road segmentation outputs for cartography, transportation analytics, emergency management, and infrastructure monitoring; and (iii) the methodological rationale for selecting overhead imagery and deep learning as the core technological components. Beyond producing a segmented raster representation, road extraction is commonly a precursor to vectorization and graph construction; therefore, attention to geometric fidelity and continuity is essential for enabling subsequent steps such as centerline derivation, network connectivity analysis, and integration into geographic information systems.

The broader significance of road segmentation extends beyond mapping as an end in itself. Reliable road delineation supports multi-temporal change detection (e.g., expansion of peri-urban road networks), accessibility modeling (e.g., estimating travel time to healthcare facilities), disaster impact assessment (e.g., identifying disrupted links after floods or earthquakes), and humanitarian logistics (e.g., routing relief supplies where base maps are outdated). These applications amplify the societal and scientific value of advancing road segmentation methods, particularly in regions where open road data are incomplete or rapidly changing. By leveraging high-resolution satellite and aerial imagery and deep learning-based segmentation, the present work contributes to the ongoing effort to produce accurate, scalable, and operationally meaningful representations of road networks in diverse real-world conditions.

2. Related Works

Road extraction from overhead imagery has been an active research topic for several decades because road networks are fundamental geospatial layers for navigation, transportation planning, emergency response, infrastructure monitoring, and map updating. Early studies mainly treated road extraction as a low-level image-processing problem in which roads were detected using radiometric contrast, local linearity, edge information, morphological filtering, region growing, active contours, and graph-based linking strategies. These approaches were valuable because they introduced important geometric and contextual assumptions, such as road continuity, elongated shape, and relatively consistent width. However, their performance was often limited by the need for hand-crafted features and manually tuned parameters. In high-resolution aerial and satellite imagery, road appearance can vary substantially due to shadows, vegetation occlusion, vehicles, seasonal changes, road-surface materials, and confusion with visually similar objects such as rooftops, parking lots, bare soil, and building edges. Therefore, classical methods often struggled to generalize across different urban and rural scenes.

With the increasing availability of high-resolution aerial and satellite imagery, supervised learning methods became more common in road extraction. Instead of relying only on manually designed rules, these approaches used labeled samples and engineered descriptors to learn the visual characteristics of roads. A major step in this direction was the work of Mnih and Hinton (Mnih and Hinton, 2010), who used a neural-network-based approach for road detection in high-resolution aerial images and emphasized the difficulty of producing reliable road maps at large scale. Their work also contributed to the development of public aerial image labeling datasets, including the Massachusetts Roads Dataset, which later became a widely used benchmark for road segmentation. The dataset contains high-resolution aerial image tiles and road labels derived from map data, making it suitable for systematic comparison of learning-based road extraction methods.

The emergence of deep convolutional neural networks significantly changed road extraction research. Fully

convolutional and encoder–decoder architectures enabled end-to-end semantic segmentation, allowing models to learn hierarchical spatial features directly from data rather than depending on hand-crafted descriptors. U-Net (Ronneberger et al., 2015) is one of the most influential encoder–decoder models because it combines a contracting path for contextual feature extraction with an expanding path for precise localization through skip connections. This design is especially relevant for road segmentation, where roads are thin, elongated, and sensitive to small localization errors. Skip connections help preserve high-resolution spatial information that may otherwise be lost during downsampling, thereby improving the recovery of narrow road segments and road boundaries.

Multi-scale feature representation has also become important in road segmentation because roads appear at different widths and contextual scales within the same image. Major highways, secondary roads, and narrow residential streets may require different levels of spatial and semantic representation. Feature Pyramid Network (FPN) (Lin et al., 2017) addresses this issue by constructing a top-down feature pyramid with lateral connections, combining low-resolution semantic features with high-resolution spatial features. This architecture has been widely adopted in detection and segmentation tasks because it provides semantically rich feature maps at multiple scales. In the context of road extraction, this is useful for detecting both wide and narrow road structures while maintaining spatial detail.

Several road-specific deep learning architectures have been proposed to improve segmentation accuracy and continuity. For example, D-LinkNet (Zhou et al., 2018) introduced an encoder–decoder architecture with a pretrained encoder and dilated convolution modules for high-resolution satellite imagery road extraction. The use of dilated convolution increases the receptive field without excessive loss of spatial resolution, making it useful for capturing long road structures and broader context. The DeepGlobe Satellite Image Understanding Challenge further accelerated research by providing benchmark datasets and evaluation protocols for road extraction, building detection, and land-cover classification from satellite imagery. These benchmark efforts encouraged more systematic evaluation of deep learning methods for remote-sensing segmentation tasks.

Although pixel-wise road segmentation has achieved strong progress, recent studies have shown that road extraction is not only a mask-generation problem but also a topology-preservation problem. A road mask with high pixel-level overlap may still be unsuitable for routing or GIS analysis if it contains gaps at junctions, missing links, or fragmented road segments. RoadTracer addressed this limitation by generating road-network graphs directly from aerial images through an iterative CNN-guided search procedure, showing that graph-based approaches can better preserve connectivity than conventional segmentation followed by post-processing. This line of research highlights the need to evaluate road extraction not only by overlap-based metrics such as Intersection over Union (IoU) and F1-score, but also by connectivity-aware and graph-based criteria.

Attention mechanisms have also been introduced into segmentation architectures to improve contextual reasoning and feature selection. MA-Net (Fan et al., 2020a), or Multi-Scale Attention Network, uses attention modules such as the Position-wise Attention Block and Multi-scale Fusion Attention Block to capture spatial dependencies and channel-wise relationships across feature maps. Although MA-Net was originally proposed for medical image segmentation, its attention-based design is relevant to road extraction because roads may be partially occluded, locally ambiguous, or confused with visually similar background structures. By modeling long-range dependencies and multi-scale feature importance, attention-enhanced architectures may help improve continuity and reduce false detections in complex overhead scenes.

Backbone selection is another important issue in deep road segmentation. EfficientNet (Tan and Le, 2019) introduced compound scaling, which balances network depth, width, and input resolution to improve accuracy and computational efficiency. In comparative studies, using the same backbone across different segmentation models is important because it controls the representational capacity of the encoder and allows the analysis to focus on differences in decoder structure and feature-fusion strategy. Therefore, in the present study, U-Net, FPN, and MA-Net are evaluated using EfficientNet-B7 as a shared backbone. This controlled design enables a fair comparison among three representative decoder paradigms: skip-connected reconstruction in U-Net, pyramid-based multi-scale fusion in FPN, and attention-guided multi-scale refinement in MA-Net.

Based on the reviewed literature, existing studies demonstrate that deep encoder–decoder architectures are effective for road extraction from high-resolution overhead imagery. However, performance can vary depending on how each model preserves fine spatial details, integrates multi-scale context, and handles visually complex backgrounds. Therefore, the present study contributes by comparing U-Net, FPN, and MA-Net under a shared EfficientNet-B7 backbone on the Massachusetts Roads Dataset. This design provides a controlled assessment of how decoder architecture and feature-fusion strategy influence road segmentation accuracy, error distribution, and the preservation of thin road structures.

3. Materials and Methods

3.1. Dataset

Experiments in this study are conducted on the Massachusetts Roads Dataset, a widely used benchmark for road segmentation in overhead imagery that was introduced

by Mnih in the context of large-scale aerial image labeling. The dataset is publicly distributed with predefined partitions

(training/validation/test) and accompanying vector data used for label generation. The dataset contains 1,171 overhead (aerial) image tiles covering diverse landscapes across the state of Massachusetts, including urban, suburban, and rural areas. Each tile has a spatial extent of 1500 × 1500 pixels, corresponding to 2.25 km² per tile. In the original dataset construction, imagery was rescaled to a ground sampling of 1 pixel per square meter (i.e., approximately 1 m spatial resolution). Collectively, the dataset covers more than 2600 km², with the test set alone spanning over 110 km², which supports robust evaluation across heterogeneous environments.

The dataset provides a standard split that is commonly used for reproducible comparison: 1108 tiles for training, 14 for validation, and 49 for testing. Unless explicitly stated otherwise, studies using this benchmark typically adhere to these partitions to enable fair comparison with prior work.

Ground-truth were generated by rasterizing road centerlines obtained from the OpenStreetMap (OSM) project. Importantly, the labels are not derived from full road-surface polygons; instead, they correspond to a fixed-width raster representation of the centerline network. In the original construction, a line thickness of 7 pixels was used and no smoothing was applied. This definition is consequential for model interpretation: the learning target represents a buffered centerline mask rather than precise road boundaries, and therefore performance reflects the ability to recover a functional road footprint around centerlines as encoded in the benchmark.

The authors of the present study did not manually modify the validation or test ground-truth maps. Instead, the official Massachusetts Roads Dataset target maps were used as distributed by Mnih (2013). In the original dataset construction, road labels were generated by rasterizing OpenStreetMap road centerlines, and the validation and test target maps were hand-corrected by the dataset provider to improve evaluation accuracy. Therefore, the reported results are based on the standard corrected validation/test partitions of the benchmark and remain comparable with studies using the official dataset split. In addition to raster targets, the dataset distribution includes the vector shapefile used to generate target maps, which is useful for downstream analysis or alternative label processing.

The Massachusetts Roads Dataset is considered challenging not only due to the diversity of scenes, but also because road extraction must contend with factors common to overhead imagery, such as (i) occlusions (e.g., tree canopy), (ii) background confusion with visually similar linear features, and (iii) class imbalance, since road pixels typically occupy a small fraction of each tile. Moreover, because labels are derived from map centerlines, residual misregistration and incompleteness may persist—particularly in training data—motivating robust modeling and careful evaluation practices.

3.2. Methodology

In this study, U-Net, Feature Pyramid Network (FPN), and Multi-Scale Attention Network (MA-Net) are treated as deep encoder–decoder semantic segmentation architectures. They share the same encoder/backbone, EfficientNet-B7, while differing mainly in the decoder and feature-fusion strategy. This design isolates the effect of decoder architecture from the representational capacity of the feature extractor. U-Net was selected as a strong skip-connected baseline because its encoder–decoder structure preserves fine spatial detail, which is essential for thin road structures. FPN was selected to represent pyramid-based multi-scale feature fusion, which is relevant because roads appear at different widths and contextual scales in overhead imagery. MA-Net was selected as an attention-enhanced encoder–decoder model to examine whether spatial and channel attention can improve continuity and suppress road-like background clutter. EfficientNet-B7 was used as a common backbone because its compound-scaled design provides a high-capacity feature extractor and enables a controlled comparison among decoder designs (Tan and Le, 2019). This controlled setting supports a direct comparison between (i) skip connected reconstruction (U-Net), (ii) pyramid based multi-scale fusion (FPN), and (iii) attention-guided refinement (MA-Net).

Table 1. Fundamental characteristics and selection rationale of the evaluated segmentation models.

Model	Fundamental nature	Main mechanism	Rationale for selection in this study
U-Net	Encoder–decoder semantic segmentation network	Skip connections between encoder and decoder features	Selected as a strong baseline for dense prediction because skip connections preserve fine spatial details, which are important for narrow and elongated road structures.
FPN	Pyramid-based feature-fusion segmentation network	Top-down pathway with lateral connections and multi-scale feature aggregation	Selected to evaluate the effect of explicit multi-scale feature fusion, since roads appear at different widths and spatial scales in overhead imagery.
MA-Net	Attention-enhanced encoder–decoder segmentation network	Position-wise Attention Block and Multi-scale Fusion Attention Block	Selected to examine whether spatial and channel attention can improve contextual reasoning, continuity, and suppression of road-like background clutter.
EfficientNet-B7	Shared convolutional encoder/backbone	Compound-scaled feature extractor balancing depth, width, and resolution	Used as the common backbone for all models to control encoder capacity and isolate the influence of decoder design and feature-fusion strategy.

3.2.1. U-Net

U-Net follows a U-shaped encoder–decoder architecture consisting of (i) a contracting path that extracts progressively higher-level representations and (ii) an expanding path that reconstructs a full-resolution prediction map (Ronneberger et al., 2015). In the present configuration, the contracting path produces a hierarchy of feature maps across multiple spatial scales, ranging from high-resolution, detail-rich representations in early stages to lower-resolution, semantics-rich representations in deeper stages.

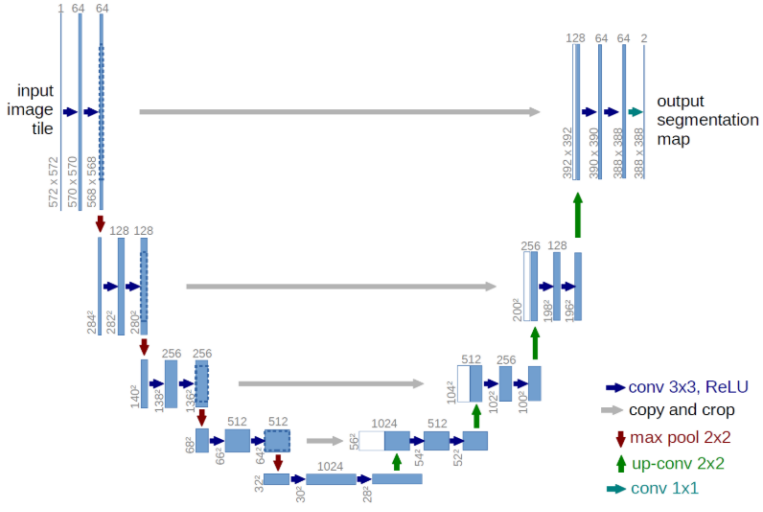

Figure 1. U-Net architecture (Ronneberger et al., 2015)

The defining mechanism of U-Net is its skip connections, which link encoder feature maps to corresponding decoder stages at matched resolutions. This structure is particularly important for road segmentation because roads often occupy only a small number of pixels in width in aerial imagery. Skip connections inject fine-grained spatial cues (edges, local contrast, and geometric detail) into the decoding process, reducing over-smoothing and enabling the recovery of narrow road segments that may otherwise be lost after successive downsampling. The decoder progressively upsamples features and fuses contextual information from deeper layers with localization cues from shallow layers, supporting accurate alignment and improved continuity of thin structures. Overall, U-Net provides a structurally straightforward and stable baseline that emphasizes precise localization through skip-connected reconstruction.

3.2.2. FPN

FPN (Lin et al., 2017, Chang et al., 2022) is built around the principle of constructing a feature pyramid with strong semantics at multiple resolutions. In the present configuration, the contracting path produces a hierarchy of feature maps across several spatial scales, ranging from high-resolution representations that preserve fine spatial detail to low-resolution representations that encode stronger semantic context. FPN then introduces a top-down pathway that progressively upsamples deeper feature maps and merges them with corresponding features from earlier stages through lateral connections. This structure allows high-level semantic information to be propagated to higher-resolution feature maps where fine localization is required.

This pyramid formulation is particularly relevant for road segmentation because roads may appear at different effective scales within the same scene: major roads occupy many pixels, whereas secondary roads can be only a few pixels wide. The top-down pathway helps preserve detectability of narrow roads by reintroducing semantic context at higher resolutions, while lateral fusion retains spatial detail necessary for accurate alignment and continuity. After pyramid construction, the multi-scale features are aggregated and refined to generate the final full-resolution segmentation output.

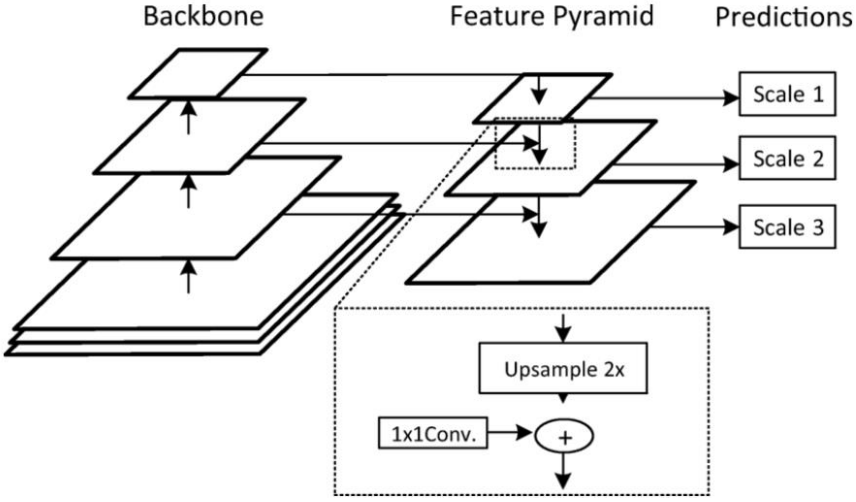

Figure 2. FPN architecture(Chang et al., 2022)

FPN provides a structured and computationally efficient mechanism for multi-scale feature fusion. In this work, it serves as a strong baseline emphasizing scale-consistent representation learning and robust detection across varying road widths and scene contexts.

3.2.3. MA-Net

MA-Net (Fan et al., 2020b) retains the overall encoder–decoder organization but modifies the feature integration process by introducing two attention-based modules designed to capture complementary dependencies. In the original MA-Net formulation, the model is described as integrating local evidence with broader context through self-attention, specifically via a Position-wise Attention Block (PAB) and a Multi-scale Fusion Attention Block (MFAB).

Position-wise Attention Block (PAB) is intended to capture long-range spatial relationships by allowing the model to relate distant pixels in a global view. This is particularly relevant for roads because their appearance can be locally ambiguous or partially missing due to tree canopy, shadows, or sensor/illumination artifacts. By enabling

information from confidently detected road segments to influence uncertain regions, spatial attention can reduce local breaks and promote more globally consistent road traces.

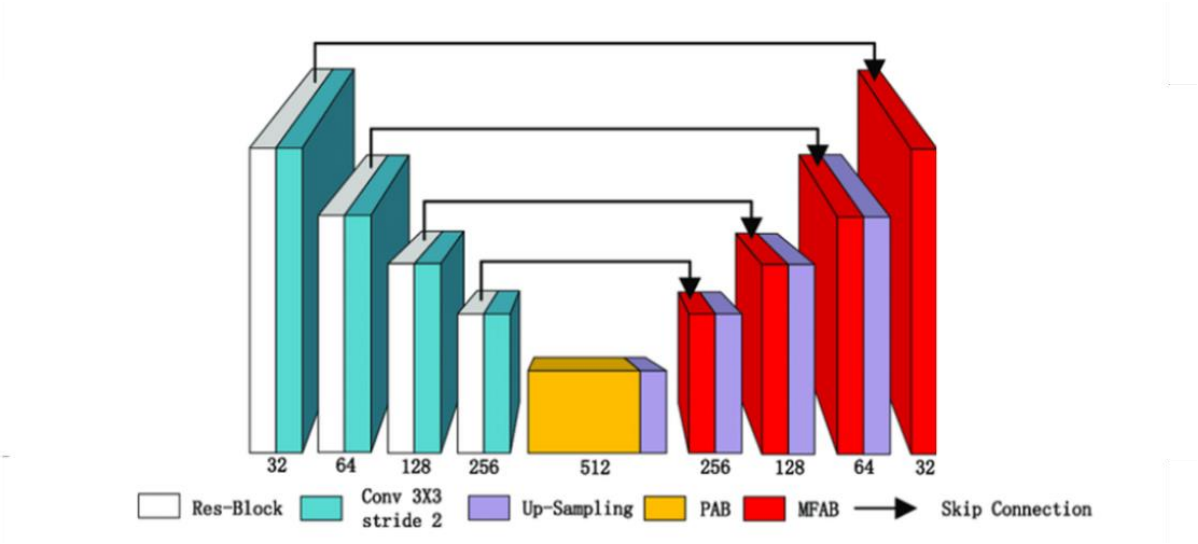

Figure 3. MA-Net architecture (Fan et al., 2020b)

MFAB is designed to capture channel dependencies while fusing information across feature levels, emphasizing informative channels during multi-scale integration. For road segmentation, this mechanism targets a common issue in overhead scenes: background elements such as parking lots, rooftops, bare soil, and linear shadows can exhibit road-like textures. Attention-guided fusion helps the network prioritize feature channels that are more discriminative for roads and suppress channels that amplify distractors. MA-Net can be interpreted as an encoder–decoder that not only reconstructs spatial detail, but also actively re-weights spatial evidence and multi-scale semantics to improve continuity and reduce false positives in visually complex regions.

3.2.4. Comparison of U-Net and MA-Net

Because all models share the same backbone capacity (EfficientNet-B7 in the contracting path), the primary methodological differences arise from how multi-scale information is fused and how contextual information is introduced during decoding.

U-Net relies on deterministic skip connections to transfer encoder detail directly to the decoder at matched resolutions, strengthening localization through explicit feature reuse. Its strengths are most directly linked to precise boundary recovery and the preservation of thin structures, though performance can degrade when local evidence is weak, leading to fragmentation in challenging regions.

FPN relies on deterministic pyramid construction: the top-down pathway and lateral connections merge deep semantic features with higher-resolution features, producing semantically enriched feature maps at multiple scales. Its strength lies in efficient and scale-consistent representation, making it well suited to road extraction where target structures vary substantially in width.

MA-Net augments the encoder–decoder paradigm with explicit attention mechanisms. The PAB introduces global spatial dependency modeling, while MFAB guides multi-scale fusion through channel-wise selectivity. These modules are intended to integrate global dependencies with local cues, which can be advantageous for maintaining continuity across occlusions and suppressing false positives in complex backgrounds. Relative to U-Net and FPN, MA-Net typically introduces additional architectural complexity; however, this complexity is explicitly directed toward improving contextual coherence and discriminative feature fusion.

4. Results

Model performance was evaluated using standard pixel-level segmentation criteria, namely Accuracy, Intersection over Union (IoU), F1-score, Precision, and Recall. In the context of road segmentation, where road pixels typically represent a small fraction of the image area, Accuracy alone

can be less diagnostic because it may remain high even when thin road structures are partially missed. For this reason, overlap-based and error-balance measures (IoU and F1-score) provide a more informative reflection of segmentation quality, while Precision and Recall clarify whether errors are dominated by false positives (spurious road detections) or false negatives (missed road pixels).

In addition to conventional pixel-level metrics, a tolerance-based topology-aware metric was computed to evaluate road-centerline continuity. Road extraction is not only a binary mask-generation task, because small gaps, disconnected links, or fragmented junctions can reduce the usefulness of the extracted road network for GIS and routing applications. Therefore, a Buffered Skeleton F1-score (BS-F1) was calculated. First, both the predicted road mask and the reference road mask were skeletonized to approximate their centerline structures. Then, a small spatial buffer was applied around each skeleton to account for minor rasterization and alignment differences. Buffered Skeleton Precision measures the proportion of the predicted skeleton located within the buffered reference skeleton, while Buffered Skeleton Recall measures the proportion of the reference skeleton recovered by the buffered predicted skeleton. The harmonic mean of these two values gives BS-F1. This metric provides a practical raster-based approximation of road-continuity preservation without requiring full conversion of segmentation masks into road graphs.

The quantitative results are reported in Table 1, which compares MA-Net, FPN, and U-Net under the same evaluation protocol. As shown in Table 1, U-Net achieves the strongest overall performance with Accuracy = 0.99, IoU = 0.89, and F1-score = 0.94, accompanied by Precision = 0.97 and Recall = 0.92. FPN follows with Accuracy = 0.99, IoU = 0.85, and F1-score = 0.92 (Precision = 0.95, Recall = 0.90), while MA-Net yields Accuracy = 0.98, IoU = 0.82, and F1-score = 0.90 (Precision = 0.92, Recall = 0.87).

The relative improvements of U-Net in IoU and F1-score indicate a closer overall match to the ground-truth road footprint, while its simultaneous gains in Precision and Recall suggest that it reduces both over-detection and missed road segments compared to the other architectures. In contrast, MA-Net exhibits the lowest Precision–Recall pair, consistent with a higher frequency of both spurious activations and omissions, whereas FPN provides a more balanced performance between these extremes.

Table 2. Metrics of the U-Net, MA-Net ,and FPN models.

Method	Accuracy	IoU	F1-Score	Precision	Recall
MA-Net	0.92	0.82	0.90	0.98	0.87
FPN	0.95	0.85	0.92	0.99	0.90
U-Net	0.97	0.89	0.94	0.99	0.92

A Beyond scalar metrics, Figure 4 provides qualitative visual comparisons that juxtapose representative scenes with the corresponding Ground Truth and model predictions, enabling a structural assessment of road continuity, boundary alignment, and background suppression. Cross the illustrated examples, the dominant road layout is generally recovered by all methods; however, differences emerge most clearly in difficult regions such as thin residential streets, intersections, and areas with visually confusing textures. In these cases, the predictions that more faithfully preserve narrow segments and reduce isolated non-road responses are also those that tend to achieve stronger overlap-based metrics, reinforcing the quantitative ranking observed in Table 1. To further characterize failure modes, Figure 5 presents the error maps for each model, which localize disagreement with the reference mask and make it possible to identify where errors concentrate spatially. The error maps reveal that disagreements are not uniformly distributed: they intensify around connectivity-critical structures (junctions and thin links) and in cluttered regions where road-like patterns can trigger false positives. Comparing the three methods, the error patterns are most widespread for MA-Net, more localized for FPN, and most limited for U-Net, which is consistent with the relative Precision/Recall balance and the superior IoU and F1-score reported for U-Net in Table 1.

The comparative evaluation indicates a consistent performance ordering across quantitative metrics and spatial error diagnostics. As reported in Table 2, U-Net achieves the strongest overall scores (Accuracy = 0.99, IoU = 0.89, F1 = 0.94, Precision = 0.97, Recall = 0.92), followed by FPN (Accuracy = 0.99, IoU = 0.85, F1 = 0.92, Precision = 0.95,

Recall = 0.90), while MA-Net attains the lowest values among the three (Accuracy = 0.98, IoU = 0.82, F1 = 0.90, Precision = 0.92, Recall = 0.87).

To further validate the comparative results, statistical analysis was conducted using per-image performance scores (Table 3). For each evaluated image, IoU and F1-score were computed separately for U-Net, FPN, and MA-Net. The mean, standard deviation, and 95% confidence intervals were then calculated for each model.

Table 3. Statistical validation of per-image segmentation performance.

Model	IoU Mean	IoU SD	IoU 95% CI	F1-score Mean	F1-score SD	F1-score 95% CI
U-Net	0.864	0.055	0.837-0.892	0.926	0.032	0.910-0.942
FPN	0.685	0.043	0.663-0.705	0.812	0.031	0.796-0.827
MA-Net	0.726	0.072	0.686-0.759	0.839	0.050	0.811-0.862

In addition, paired Wilcoxon signed-rank tests were used to compare U-Net with FPN and MA-Net, since all models were evaluated on the same image samples (Table 4). A Holm-Bonferroni correction was applied to account for multiple pairwise comparisons. The statistical results confirmed that U-Net achieved the highest mean IoU and F1-score. U-Net obtained an IoU of 0.864 with a 95% confidence interval of 0.837–0.892 and an F1-score of 0.926 with a 95% confidence interval of 0.910–0.942. MA-Net achieved an IoU of 0.726 and an F1-score of 0.839, while FPN achieved an IoU of 0.685 and an F1-score of 0.812. The Wilcoxon signed-rank tests showed that U-Net significantly outperformed both FPN and MA-Net for IoU and F1-score after Holm-Bonferroni correction. These results provide statistical support for the superiority of U-Net under the present evaluation setting.

Table 4. Paired Wilcoxon signed-rank test results.

Metric	Comparison	Wilcoxon Statistic	p-value	Holm Threshold	Significant After Correction
IoU	U-Net > FPN	105.0	0.000061	0.012500	Yes
F1-score	U-Net > FPN	105.0	0.000061	0.016667	Yes
IoU	U-Net > MA-Net	104.0	0.000122	0.025000	Yes
F1-score	U-Net > MA-Net	104.0	0.000122	0.050000	Yes

The topology-aware results provide a complementary interpretation of the segmentation outputs (Table 5). U-Net achieved the highest Buffered Skeleton F1-score (0.596; 95% CI: 0.535–0.656), followed by FPN (0.498; 95% CI: 0.416–0.582) and MA-Net (0.494; 95% CI: 0.413–0.576). This indicates that U-Net provided the best balance between centerline precision and centerline recall. FPN and MA-Net achieved very high BS-Recall values, suggesting that they recovered a large proportion of the reference skeleton; however, their lower BS-Precision values indicate more extra or fragmented skeleton responses. Pairwise Wilcoxon signed-rank tests showed that U-Net obtained higher BS-F1 values than FPN and MA-Net, but the differences were not statistically significant after Holm correction. Therefore, the topology-aware analysis is interpreted as complementary evidence rather than definitive proof of topological superiority. Overall, these results confirm that road-connectivity preservation remains challenging even when pixel-level segmentation scores are strong.

Table 5. Topology-aware evaluation using Buffered Skeleton F1-score

Model	BS-Precision	BS-Recall	BS-F1	BS-F1 95% CI
U-Net	0.443	0.950	0.596	0.535-0.656
FPN	0.348	0.994	0.498	0.416-0.582
MA-Net	0.346	0.980	0.494	0.413-0.576

The qualitative outputs in Figure 4 and the corresponding error maps in Figure 5 further reinforce this ranking by revealing how each architecture distributes errors across road corridors and structurally complex regions.

U-Net's superior performance can be attributed primarily to its explicit skip-connected decoding pathway, which transfers high-resolution spatial detail from early encoder stages directly to the decoder. For road extraction, this mechanism is particularly consequential because roads are thin, elongated, and topology-sensitive: small localization errors at the pixel level may translate into discontinuities that reduce IoU/F1 and undermine network connectivity. The skip connections provide the decoder with fine-grained cues (edges and local geometry) while deeper features provide semantic context, enabling U-Net to maintain a favorable balance between false positives and false negatives—reflected in its highest Precision and Recall (0.97 and 0.92, respectively).

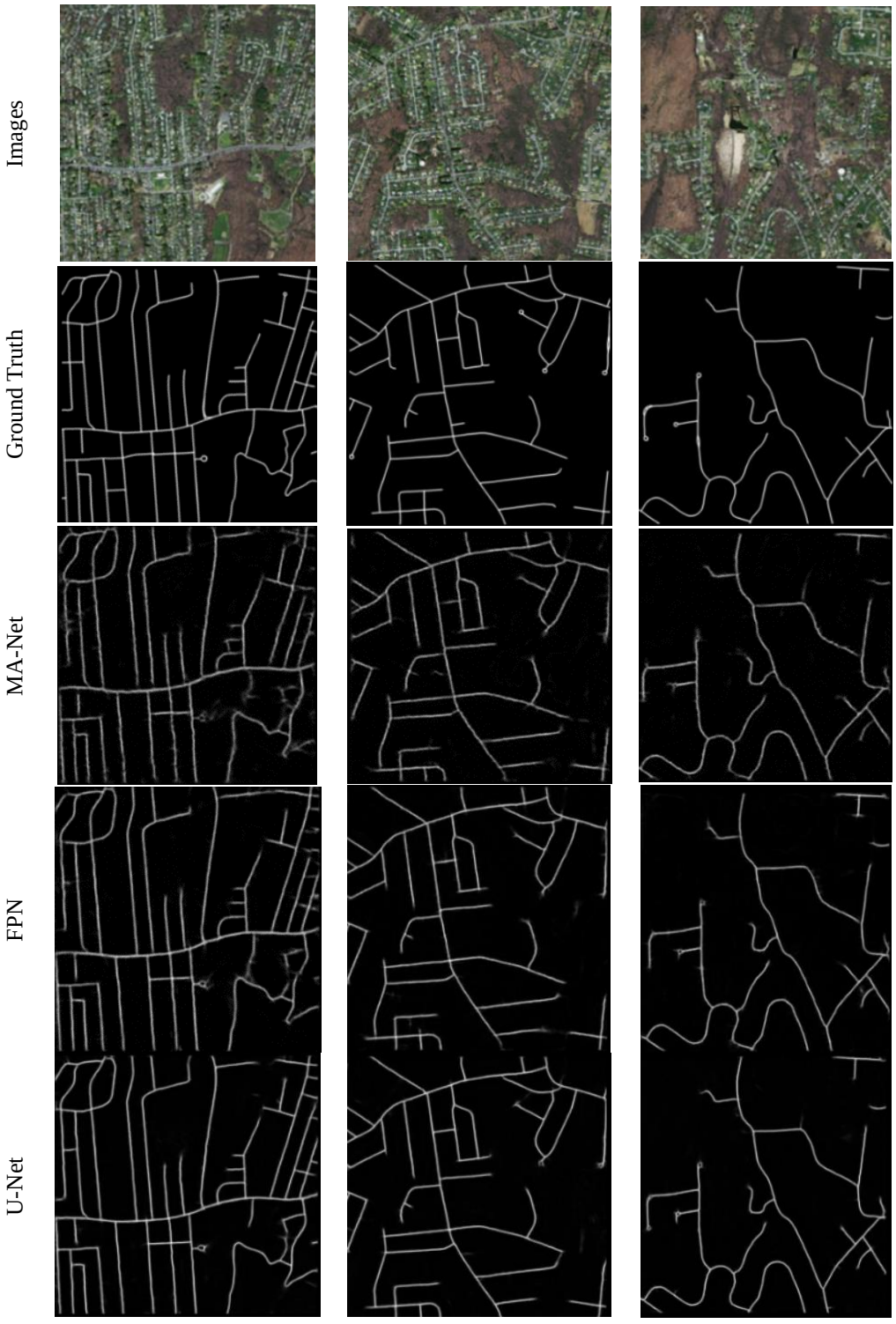

Figure 4. Qualitative comparison of road segmentation results on representative test scenes.

The error maps offer an informative perspective on why this may occur. Compared to U-Net and FPN, MA-Net exhibits the most extensive disagreement patterns, indicating more frequent deviations along road structures and a higher propensity for background-related errors.

A likely explanation is that attention-driven aggregation can be more sensitive to confounders common in overhead imagery (linear shadows, rooftops, parking boundaries, and other road-like textures). If attention selectively amplifies such cues, the model may produce additional false positives or weaken marginal road evidence, which would jointly reduce Precision and Recall. Furthermore, attention-enhanced architectures often introduce additional complexity; without strong regularization and sufficiently diverse training examples, this added flexibility can translate into less stable generalization relative to architectures with simpler, well-conditioned fusion pathways.

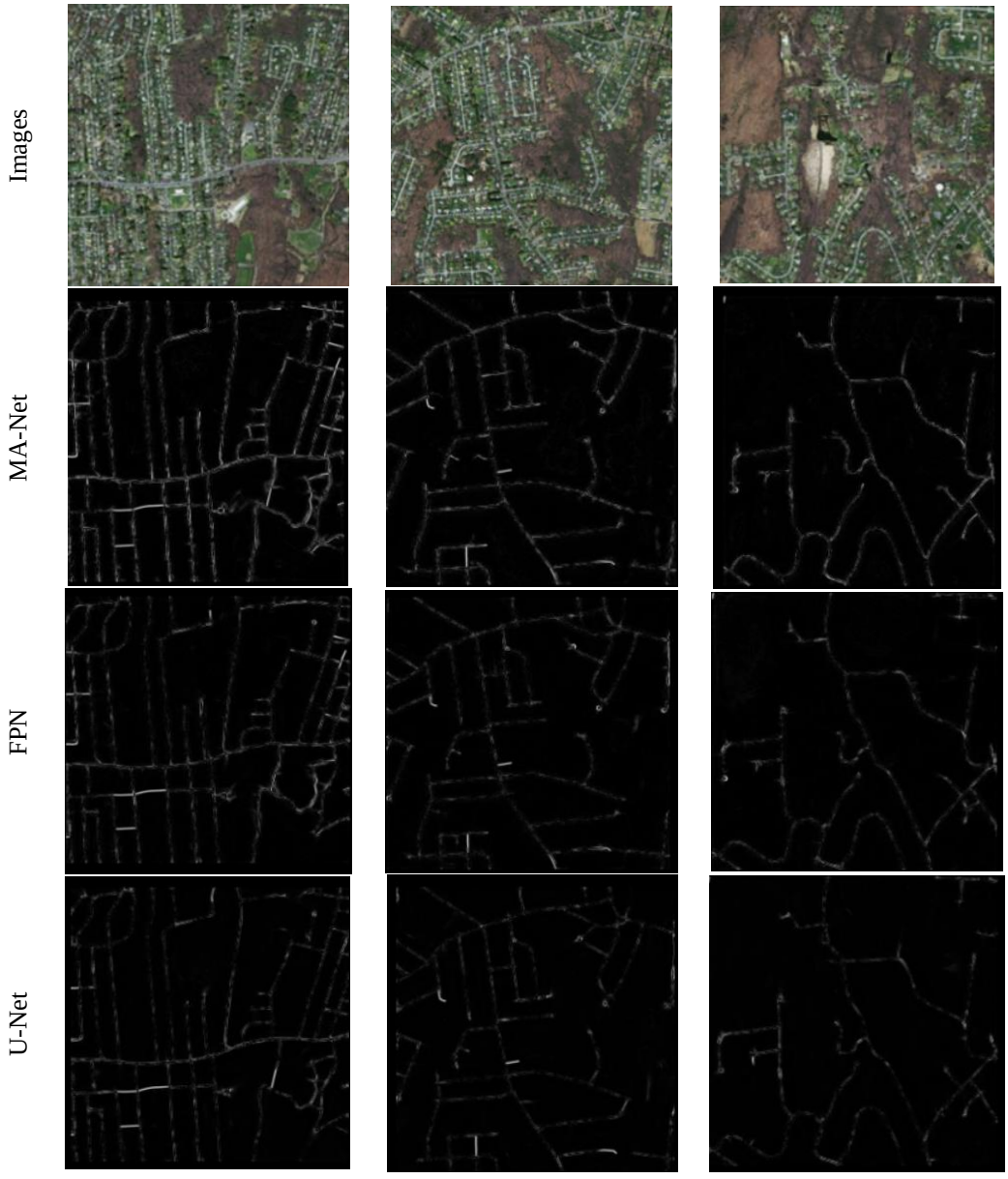

Figure 5. Error-map visualization of pixel-level disagreement between predictions and ground truth.

Taken together, the results suggest that the decisive factor in this study is not backbone capacity (which is controlled), but the inductive bias of the decoding/fusion design. U-Net's dense skip connections supply the decoder with direct access to high-frequency spatial detail, which is crucial for producing continuous, well-aligned road masks. This architectural property explains both its superior overlap-based metrics and its reduced error intensity in Figure 5.

FPN offers strong multi-scale semantics and performs competitively, but its pyramid fusion appears less effective than U-Net at recovering the finest structures with maximal geometric fidelity. MA-Net's attention mechanisms, while theoretically advantageous for global coherence, appear to be more vulnerable to clutter and road-like distractors in this dataset, leading to higher disagreement patterns and lower precision–recall balance.

From an operational standpoint, these findings are important because road segmentation quality is not solely determined by pixel-wise overlap, but also by connectivity preservation, particularly around junctions and thin links. Architectures that strongly support detail-aware reconstruction (as U-Net does) are therefore likely to yield more reliable downstream outcomes for vectorization and network analysis. At the same time, FPN's strong performance highlights the value of explicit multi-scale fusion, suggesting that hybrid strategies (e.g., stronger high-resolution reconstruction combined with pyramid context) may be a productive direction for further improving robustness in heterogeneous overhead scenes.

5. Discussion and Conclusion

This study investigated automated road extraction from high-resolution overhead imagery by benchmarking three widely used deep segmentation architectures—U-Net, FPN, and MA-Net—within a controlled experimental design. The experiments were conducted on the Massachusetts Roads Dataset, a standard benchmark consisting of 1,171 aerial image tiles (1500 × 1500 pixels) with labels generated as fixed-width rasterized road centerlines, using the conventional train/validation/test split (1108/14/49).

The dataset characteristics are practically important for interpretation: because target maps represent a buffered centerline footprint (rather than precise road-surface polygons), model performance reflects the ability to recover functional road masks aligned with the benchmark's centerline definition.

Quantitative results (Table 1) show a consistent ranking across overlap-based and error-balance measures. U-Net achieved the best overall scores (IoU = 0.89, F1 = 0.94) and the strongest precision–recall balance (0.97/0.92), indicating fewer false activations and fewer missed road pixels relative to the competing methods.

FPN yielded competitive performance (IoU = 0.85, F1 = 0.92), confirming that explicit multi-scale feature fusion is well suited to road structures that vary substantially in width within the same scene.

MA-Net obtained lower scores (IoU = 0.82, F1 = 0.90), suggesting that attention-enhanced fusion, while theoretically beneficial for contextual reasoning, did not translate into improved overall segmentation performance under the present dataset and training conditions.

The qualitative analysis reinforces these findings. The visual comparisons (Figure 4) indicate that all models generally reconstruct the dominant road layout, but differences emerge most clearly in difficult regions such as thin residential streets, intersections, and areas with visually confusing textures.

The error-map assessment (Figure 5) provides a complementary diagnostic view by localizing disagreements with the reference mask; errors intensify around connectivity-critical structures (junctions and thin links) and cluttered regions that can induce road-like false positives.

In this spatial perspective, U-Net exhibits the most confined disagreement patterns, FPN shows moderate residual errors in challenging thin structures, and MA-Net demonstrates the most widespread error responses—consistent with the quantitative precision–recall differences.

From an architectural standpoint, the results suggest that the decisive factor is not backbone capacity (held constant), but rather the inductive bias of the decoding and feature-integration mechanism. U-Net's skip-connected reconstruction provides strong access to high-frequency spatial detail, which is particularly advantageous for thin, topology-sensitive road structures.

FPN's multi-scale pyramid design effectively propagates semantic context across resolutions and therefore remains highly competitive, but appears slightly less effective than U-Net in recovering the finest structures with maximal geometric fidelity.

MA-Net's attention-driven refinement introduces additional flexibility intended to strengthen contextual reasoning, yet the observed error maps indicate that it is more prone to disagreement in cluttered backgrounds, potentially reflecting sensitivity to confounding linear patterns in overhead scenes.

Finally, several directions emerge for future work. First, because road utility depends strongly on **connectivity**, integrating topology-aware objectives and evaluation (beyond pixel overlap) may better align training with downstream road-graph extraction needs.

Although BS-F1 provides a practical raster-based approximation of road-continuity preservation, full graph-based evaluation remains an important direction for future work. Metrics such as Average Path Length Similarity (APLS), TOPO, and shortest-path-based evaluation require converting predicted road masks into vector road graphs and can more directly assess routing-level connectivity. APLS, for example, compares ground-truth and predicted road graphs using differences in shortest-path lengths, while TOPO-type metrics evaluate reachable subgraph similarity.

Second, dataset-specific label characteristics—centerline buffering and potential residual misregistration—motivate strategies for robustness, such as boundary-tolerant losses, uncertainty-aware training, or weakly supervised refinement.

Third, extending evaluation across additional geographic regions and sensor conditions would strengthen conclusions about generalization, particularly for operational deployments where illumination, occlusion, and background composition differ substantially.

The findings support U-Net as the most reliable architecture among those evaluated for this benchmark, while also emphasizing that multi-scale fusion (FPN) remains a strong competitive alternative and that attention-based designs warrant further investigation under broader training regimes and robustness-oriented settings.

References

BASTANI, F., HE, S., ABBAR, S., ALIZADEH, M., BALAKRISHNAN, H., CHAWLA, S., MADDEN, S. & DEWITT, D. Roadtracer: Automatic extraction of road networks from aerial images. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 4720–4728. https://doi.org/10.48550/arXiv.1802.03680

CANNY, J. J. I. T. O. P. A. & INTELLIGENCE, M. 2009. A computational approach to edge detection. 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

CHANG, C.-W., SANTRA, S., HSIEH, J.-W., HENDRI, P., LIN, C.-F. J. M. T. & APPLICATIONS 2022. Multi-fusion feature pyramid for real-time hand detection. 81**,** 11917–11929. https://doi.org/10.1007/s11042-021-11897-7

FAN, T., WANG, G., LI, Y. & WANG, H. 2020a. Ma-net: A multi-scale attention network for liver and tumor segmentation. *Ieee Access,* 8**,** 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372

FAN, T., WANG, G., LI, Y. & WANG, H. J. I. A. 2020b. Ma-net: A multi-scale attention network for liver and tumor segmentation. 8**,** 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372

KRÄHENBÜHL, P. & KOLTUN, V. J. A. I. N. I. P. S. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. 24. https://doi.org/10.48550/arXiv.1210.5644

LAFFERTY, J., MCCALLUM, A. & PEREIRA, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. https://dl.acm.org/doi/10.1145/3765612.3767302

LIN, T.-Y., DOLLÁR, P., GIRSHICK, R., HE, K., HARIHARAN, B. & BELONGIE, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 2117–2125. https://doi.org/10.48550/arXiv.1612.03144

MNIH, V. & HINTON, G. E. Learning to detect roads in high-resolution aerial images. European conference on computer vision, 2010. Springer, 210–223. https://doi.org/10.1007/978-3-642-15567-3_16

RONNEBERGER, O., FISCHER, P. & BROX, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015. Springer, 234–241. https://doi.org/10.48550/arXiv.1505.04597

TAN, M. & LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, 2019. PMLR, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

ZHOU, L., ZHANG, C. & WU, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018. 182–186. https://doi.org/10.1109/CVPRW.2018.00034