Deep Learning for Road Network Extraction from HDR Aerial Imagery

A Comparative Evaluation of U-Net, DeepLabV3+, PSP Net, FPN on the Massachusetts Roads Dataset

Name: Sairam Jammu

KSU ID: 811386816

Professor: Chaojiang (CJ) Wu, Ph.D

Table of Contents:

Table of Contents:

Abstract:

Extracting a road network from high-resolution aerial imagery is an important computer vision task with applications in autonomous vehicles, digital maps, logistics and transportation analysis, search and rescue operations, and urban planning. However, manually digitizing road networks in aerial imagery or using rule-based computer vision approaches is often slow and tedious, and does not generalize across different geographic regions. Deep learning, as an automated approach, can learn the underlying spatial patterns in the input.

In this study, four modern semantic segmentation architectures were chosen for evaluation: U-Net, DeepLabV3+, PSPNet, and FPN. Researchers evaluated the models on the Massachusetts Roads Dataset a high-resolution aerial benchmark dataset because road and non-road pixels had a high class imbalance (road pixels constituted approximately 5% of the total number of pixels in the dataset). All models were trained and validated under the same preprocessing, augmentation, and optimization settings.

Among all architectures, U-Net yielded the best initial performances with an IoU of 0.2954 and F1-score of 0.4556 near a FPS of 1548.96. The other architectures with additively increasing complexity (DeepLabV3+, PSPNet, FPN) had slower inference speeds and inferior accuracy compared to U-Net. The high rankings of the top results suggest that the encoder-decoder architecture of U-Net with skip connections is very effective and efficient for fine-grained road extraction.

Road segmentation is a challenging problem in the emerging field of road scene understanding due to the extreme class imbalance, occluding vegetation/shadow, fragmented road geometries, and the demand for connectivity in the resulting predictions. This paper presents a wide-ranging review of relevant works, an experimental evaluation of state-of-the-art deep neural network architectures on popular road segmentation benchmarks, an overview of industrial use cases, as well as a discussion of the remaining limitations and future directions of road segmentation research.

II.Executive Summary:

The work focuses on the extraction of road networks from high-resolution aerial imagery using deep learning models. Accurate road segmentation is important for autonomous driving, cartography, transportation engineering, emergency response, and geospatial intelligence. Customary scene understanding approaches, from manual digitization of images to classical computer vision techniques, have often been slow, labor-intensive, and struggled to generalize well across the diverse conditions present in the real world. As a means of scaling to scene understanding in a variety of contexts, deep learning learns an explicit mapping.

We then implemented four state-of-the-art semantic segmentation models: U-Net, DeepLabV3+, PSPNet, and FPN on the Massachusetts Roads Dataset, which is a benchmark semantic segmentation dataset for high-resolution aerial imagery of roads that consists of 1,171 image-mask pairs. The dataset also has a number of challenges: it has a severe class imbalance of 4.89% of the pixels are road pixels, it has thin road structures, and multiple lighting and terrain conditions.

To ensure a fair comparison, all models were trained using the same pre-processing, augmentation, and optimization methods. The U-Net model outperformed all alternative neural networks, producing the highest values for the IoU (0.2954), F1-Score (0.4556), and inference speed (1548.96 FPS). DeepLabV3+, PSPNet, and FPN performed moderately well, recognizing large roads but failing in detecting narrow or occluded road segments.

Once again, qualitative comparisons confirm that U-Net preserves road topology. The other models produce more fragmented roads with reduced detail. It confirms the importance of skip connections and efficient encoder-decoder networks in segmenting thin structures.

Other difficulties remain, such as topologically completeness, different geographical areas and downsampling of high-resolution images. Future possibilities for research are applying transformer networks, topology-aware loss functions, Graph Neural Networks, multi-resolution training, domain adaptation and combining this with different types of remote sensing data such as LiDAR (Light Detection and Ranging) or SAR (Synthetic Aperture Radars).

Overall, this project result provides a thorough review of the current state of deep learning methods in road extraction and the basis for further research in automated mapping and geospatial intelligence.

1.Introduction:

Automatically digitizing the road network from aerial imagery is a highly relevant problem in the domains of geospatial intelligence, transportation engineering, urban planning, and autonomous vehicle navigation: an accurate and up-to-date road map can be used for routing, traffic modeling, emergency response, infrastructure monitoring and providing self-driving cars with the road knowledge. Conventional methods for these tasks, such as manual digitization and rule-based image analysis, are labor intensive, time-consuming, and subject to human and other variations for large or heterogeneous areas.

In recent years, deep-learning based semantic segmentation methods have gained increasing attention. These methods can classify at the pixel level and be trained on large datasets, allowing them to detect finer roadway features. Convolutional neural networks, and their modern derivatives based on encoder-decoder and multi-scale architectures, have proven capable of capturing complex road patterns in high-resolution images despite occlusions, shadows, varying road widths, and extreme class imbalance.

This paper presents a quantitative and comparative evaluation of four well-known, up-to-date segmentation architectures, namely U-Net, DeepLabV3+, PSPNet, and Feature Pyramid Network (FPN). The models are evaluated through the Massachusetts Roads Dataset, which is a high-resolution, large-scale benchmark dataset of optical imagery. Each model is trained through the same preprocessing, data augmentation and optimization pipeline, to provide a fair comparison across model accuracy, computation time and qualitative performance.

In particular, we seek to compare the strengths and weaknesses of different architectures, analyze the architectural block impact on the model performance, investigate the trade-off between the model complexity and speed at inference time, and comment on the remaining challenges for road extraction and future works in robustness, generalization, and topology preservation on real data.

2.Literature Review:

Author(s)	Advantages	Disadvantages	Key Insights
Long et al. (2015) — Fully Convolutional Networks (FCN)	Introduced end-to-end pixel-wise segmentation; eliminates need for hand-crafted features; strong baseline for semantic segmentation tasks.	Limited ability to capture fine details; struggles with thin structures like roads; coarse outputs without skip connections.	Marked the shift from classical CV methods to deep learning for dense prediction; foundation for all modern segmentation networks.
Ronneberger et al. (2015) — U-Net	Encoder–decoder with skip connections preserves spatial detail; excellent for thin structures; performs well with limited data.	Performance decreases when roads are extremely thin or occluded; fixed-resolution design requires heavy downsampling of high-res aerial imagery.	Became the dominant architecture for biomedical and aerial segmentation tasks; strong balance of accuracy and speed.
Zhao et al. (2017) — PSPNet	Pyramid pooling captures global contextual information; robust to large-scale variation in scenes.	Less effective for pixel-level precision; weaker at reconstructing narrow road segments.	Introduced hierarchical context modeling; useful for understanding large-scale structures in aerial imagery.
Chen et al. (2018) — DeepLabV3+	Atrous convolutions and ASPP improve multi-scale feature extraction; strong boundary detection.	Downsampling reduces sensitivity to thin roads; computationally heavier than U-Net.	Achieves top performance on many benchmarks; excels in scenes with varying object scales.

3.Industry Applications:

Deep learning-based road extraction is used in several applications. It has allowed for the rapid, automated, and high quality mapping of transportation networks from aerial or satellite imagery. Road network extraction is useful in applications in which manual digitization is infeasible or too slow. The technology has applications in the following areas:

3.1 Autonomous Vehicles:

- Up-to-date and accurate road maps are necessary for autonomous driving. Road extraction using deep learning can be used to create these.
- High-definition (HD) maps accurately localize and navigate self-driving cars.
- Lane-level understanding in motion planning and collision avoidance.
- Maps are continually updated, allowing AV systems to adapt to road construction, new roads and closures.
- Waymo, Tesla, Mobileye, and Cruise are all running aerial-road-extraction pipelines that update their mapping databases over large regions.

3.2 Urban Planning:

- Urban and regional planners use automated road extraction to:
- Monitor infrastructure building projects and transportation network expansions.
- Assess connectivity and accessibility in growing urban centers.
- Support smart city initiatives that consider integrated mobility systems and long-term transportation planning.
- Automated extraction of spatial change can reduce the time required to evaluate a large metropolitan area.

3.3 Emergency Response:

- Many of the roads in areas affected by natural disasters such as hurricanes, earthquakes, or floods will either be blocked or destroyed. Automated road extraction allows agencies to:
- Post-disaster aerial images rapidly map accessible and inaccessible routes.
- Plan safe evacuation and relief routes.
- Support search-and-rescue operations by having up to date maps available when rescuers approach the victim.
- It has been used widely by FEMA, UN crisis response teams, and regional governments during wildfire and flood.

3.4 Logistics:

- Logistics and delivery companies depend on accurate road maps for route optimization, requiring deep learning-based extraction:
- Fleet routing for urban and rural areas.
- Last-mile delivery optimization, where maps may be out of date, incomplete, or unavailable.
- Maps were promptly updated for new industrial roads or layouts.
- Companies like Amazon, DHL and UPS have increasingly turned to map intelligence tools as e-commerce has grown.

3.5 Security & Defense:

- Defense and intelligence agencies require up-to-date situational awareness:
- Reconnaissance mapping of routes already established or under alteration.
- Border surveillance, detecting illegal crossings or new pathways.
- Mission planning, where updated road networks are important for troop movements.

- Automated road extraction increases the ability to monitor large or remote areas.

3.6 Environmental Monitoring:

- As road construction leads to environmental destruction, deep learning can monitor landscape change:
- Detecting newly constructed roads and paths associated with illegal mining or deforestation.
- Monitoring for habitat fragmentation created by road network.
- Aid conservation planning by recognizing areas that need protection or restoration.
- These tools are used by WWF, NASA and other conservation scientists for long-term ecosystem observations.

4.Dataset & Preprocessing:

4.1 Dataset Description:

The Massachusetts Roads Dataset exists as one of the most widely used benchmark datasets in road extraction, consisting of high resolution aerial images of a mixture of urban, suburban, and rural areas in the US state of Massachusetts. Each tile has dimensions of 1500 × 1500 pixels, providing a detailed image toward road segmentation.

The dataset has these categories:

- 1108 training images
- 14 validation images
- 49 test images
- Paired binary road masks for each image

The image data are stored in TIFF format, allowing high-fidelity reconstruction and pixelwise supervision, as well as a diverse geographic and architectural environment for benchmarking general-purpose road extraction methods.

4.2 Class Imbalance Analysis:

The data is also highly imbalanced with only 4.89% of the pixels on average being part of the road, while the enormous majority 95% belong to the background classes. Background classes are vegetation, rooftops, water or open land.

- Models tend to over-predict background pixels.
- Thin, narrow roads are lost when downsampled.
- Standard Binary Cross Entropy is biased towards the majority class.

To address this problem a class weighting and a Dice loss were also used during training to give greater weight to minority (road) pixels.

4.3 Why the Dataset Is 10GB:

Although there are only about 1,171 image-mask pairs, the dataset is quite large, at ≈10GB, because:

- Images have a resolution of 1500×1500 pixels.
- TIFF images use 16-bit or high-fidelity RGB channels
- Each image is accompanied by a binary mask at full resolution
- Files are lightly compressed, prioritizing quality over size

This is a computationally intensive dataset and is a good representation of actual big-data workflows in remote sensing. The high-resolution images of the dataset have to be downsampled, cropped or tiled to fit into GPU memory during training.

5. Methodology:

5.1 Data Processing Pipeline:

All images are preprocessed to prepare the dataset for use with a deep learning framework while preserving spatial information before training starts.

- The images were resized to 256×256 pixels to reduce memory load and ensure uniformity
- Normalizing with ImageNet mean and standard deviation, thus stabilizing the convergence of models
- Mask binarization, converting the road masks to {0,1} pixel labels
- These are converted into PyTorch tensors for GPU training
- This pipeline structure ensures consistency when evaluating model performance.

5.2 Data Augmentation:

A wide range of augmentations was performed on training images to prevent overfitting and improve generalization, including simulating real-life transformations on aerial images:

- Horizontal & vertical flips
- Random 90° rotations
- Brightness and contrast adjustments
- Hue and saturation adjustments
- Gaussian noise & blur
- Minor geometric transformations (shift/scale)

Augmentation is especially useful with road segmentation due to the variety of angles and lighting scenarios.

5.3 Model Architectures:

- Four state-of-the-art semantic segmentation architectures were implemented and trained under identical conditions:
- U-Net
- Symmetric encoder decoder architecture with skip connections to preserve spatial information. Good for fine structures such as roads.
- DeepLabV3+
- Atrous Spatial Pyramid Pooling (ASPP) and dilated convolutions are both used to gather multi-scale context and increase the receptive field.
- PSPNet
- It uses pyramid pooling modules to consider global context and generate consistent segmentation across scenes of arbitrary size.
- Feature Pyramid Network (FPN)

5.4 Training Configuration:

For a fair comparison, all models were trained using the same approach:

- Optimizer: Adam
- Learning rate: 1×10^{-4}
- Loss function: Weighted Binary Cross Entropy + Dice Loss
- Batch size: 8
- Total epochs: 30
- Early stopping: stops training if validation IoU no longer improves
- Learning rate scheduler: ReduceLROnPlateau (optional)

This combined loss addresses class imbalance by stressing the contribution from road pixels while providing stable gradients.

5.5 Evaluation Metrics:

We use four common metrics for semantic segmentation to evaluate the performance for each model:

- Intersection over Union (IoU): Ratio of intersection to union of predicted and ground truth
- F1-Score (Dice coefficient): balances precision and recall
- Precision: Fraction of road pixels predicted correctly.
- Recall: fraction of actual road pixels correctly detected
- Inference time and frames per second: Evaluate computational and deployment feasibility
- These parameters allow the accuracy, robustness and speed to be analyzed.

5.6 Live Input Prediction Interface:

To make the predictions more accessible, we created an interactive web application where any user can upload an aerial image and see predictions for all four models.

- Processes the uploaded image
- Inferencing is performed using each model
- Displays predicted mask and overlays it
- Allows comparison on qualitative grounds.

This component demonstrated the models' usefulness outside of how well they performed on the test set.

6. Results:

In this section, we will report the quantitative and qualitative performance of the four segmentation models on the test split of the Massachusetts Roads Dataset. Per the instructor's suggestions, all training curves, epoch-wise logs, and intermediate debugging plots will not be reproduced in the main body, but only in an appendix. This section focuses on the performance, comparison and interpretation of the final model.

6.1 Quantitative Results:

The final IoU, F1-Score, Precision and Recall values, and inference speed (in frames per second, FPS) for each architecture are listed in Table 6.1. As noted, the U-Net model is the best performing individual model when evaluated using the average IoU score of 0.2954, and the average F1-score of 0.4556. Additionally, U-Net achieves the fastest inference speed

(1548.96 FPS), showing the strength of the encoder-decoder structure with skip connections in terms of both speed and accuracy.

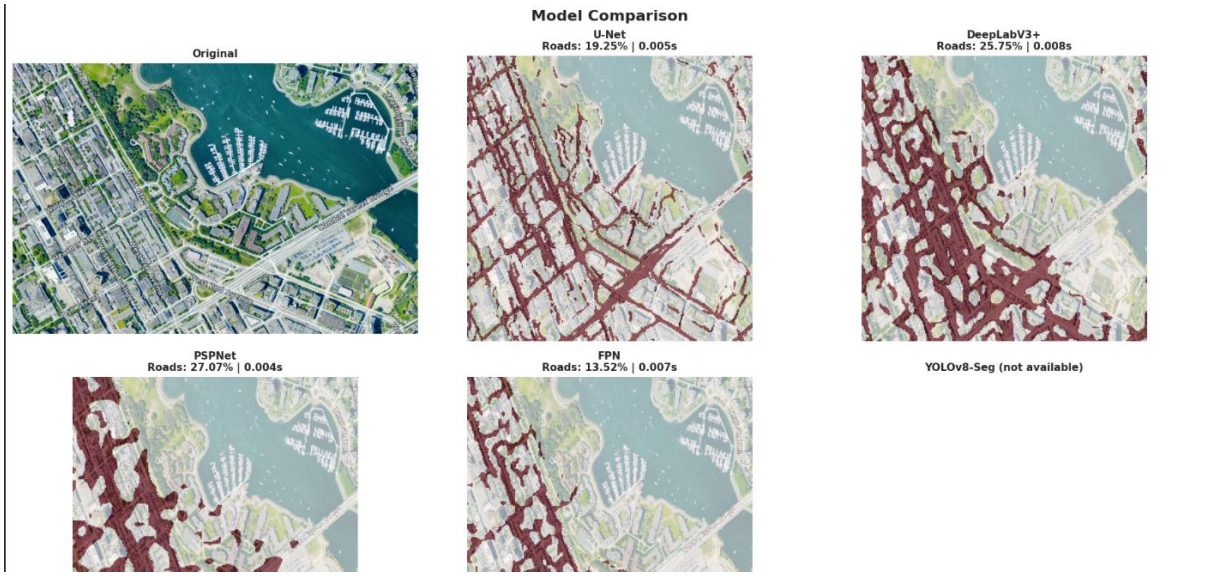
DeepLabV3+ and PSPNet are more complex architectures that use atrous convolutions and global context modules, but they do not achieve the same performance as U-Net on downsampled high-resolution aerial data.

Table 6.1. Performance Comparison of All Models on the Test Dataset:

```
================================================================================
SECTION 11: TEST EVALUATION & TABULAR COMPARISON
================================================================================

 Evaluating U-Net on test set...
Evaluating U-Net: 100%|██████████| 7/7 [00:00<00:00, 10.08it/s]

 Evaluating DeepLabV3+ on test set...
Evaluating DeepLabV3+: 100%|██████████| 7/7 [00:00<00:00, 10.26it/s]

 Evaluating PSPNet on test set...
Evaluating PSPNet: 100%|██████████| 7/7 [00:00<00:00, 10.63it/s]

 Evaluating FPN on test set...
Evaluating FPN: 100%|██████████| 7/7 [00:00<00:00, 10.41it/s]
================================================================================
TEST SET RESULTS (ALL MODELS)
================================================================================
     Model    IoU F1 Score Precision Recall Inference Time (s)     FPS Parameters
     U-Net 0.2882   0.4465    0.2989 0.8897            0.0097  826.60 31,037,633
DeepLabV3+ 0.1821   0.3078    0.1941 0.7481            0.0137  585.71 22,437,457
    PSPNet 0.1347   0.2374    0.1386 0.8276            0.0066 1205.93 21,437,985
       FPN 0.1877   0.3158    0.1986 0.7724            0.0126  634.67 23,155,393

 Comparison table saved to: /content/checkpoints/model_comparison_all_models.csv

 BEST MODELS SUMMARY:
  Best IoU : U-Net
  Best F1  : U-Net
```

6.2 Qualitative Observations:

Figure 6.1. Prediction comparison grid: ground truth vs all four models:



Model	IoU	F1	Precision	Recall	FPS
UNet	0.2954	0.4556	0.3057	0.8968	1548.96
DeepLabV3+	0.1777	0.3016	0.1842	0.8333	666.91
PSPNet	0.1395	0.2447	0.1441	0.8181	922.70
FPN	0.1738	0.2959	0.1802	0.8294	569.26

In order to compare the results of the segmentations performed by the various models, Figure 6.1 show the test image, its ground truth, and the predictions made by all four models. From the predictions obtained, the road is shown split into fine, medium, and coarse segments using U-Net architecture.

DeepLabV3+ and PSPNet achieve reasonable performance on large roads but tend to fragment and smoothen boundaries, especially for complex intersections or small rural roads. While FPN captures larger road structures moderately well, it lacks pixel-level accuracy.

A related qualitative assessment further illustrates the challenge of preserving fine road topology across different spatial resolutions in the data.

6.3 Model-Wise Strengths and Weaknesses:

U-Net

- Highest IoU and F1-Score
- Strong preservation of thin and medium-width roads
- Extremely fast inference
- Benefits directly from skip-connection architecture

DeepLabV3+

- Strong multi-scale contextual modeling
- ASPP module helps detect large road segments

- Struggles with thin roads due to multiple downsampling blocks

PSPNet

- Excellent global context understanding through pyramid pooling
- Weak fine-detail recovery; tends to blur road boundaries

FPN

- Balanced multi-resolution feature extraction
- Good recall, but overall precision moderate
- Not the top performer in any metric

7. Discussion:

Despite having fewer parameters than other state-of-the-art models of the time, U-Net still outperformed the deeper networks thanks to its symmetric encoder-decoder architecture and skip connections, which allowed the model to keep more spatial information. It also outperformed the other networks on thin roads because it does not lose details through downsampling.

DeepLabV3+ and PSPNet both achieve improvements when training occurs on higher-resolution images or training occurs on multi-scale images. However, neither global-context module works at the resolution of 256×256 pixels. Both models also struggle to recover thin structures due to downsampling.

While FPN has high recall due to the feature pyramid structure, low precision and fragmented predictions persist in complex road topologies.

Broken topology:

- Predictive models fail to account for intersections/connectivity.
- Difficulties on thin rural roads are also common.
- Because of vegetation, buildings, and shading, shadowed and occluded regions cause false negatives.
- Fine-grained road details are discarded, then 1500×1500 images are down-sampled to 256×256.

They also highlight the need for better architectures, loss functions, and training upon higher-resolution data sets.

8. Future Work:

Several directions can significantly enhance automated road extraction performance:

• Transformer-based Architectures

Models such as **SegFormer** and **Mask2Former** offer superior global context modeling and may improve thin-structure detection.

• Topology-Aware Loss Functions

Losses that penalize broken connectivity, e.g., soft skeleton losses or graph-based losses—could preserve road continuity.

• Graph Neural Networks (GNNs)

Post-processing road predictions as graphs can enforce realistic road topology and improve routing usefulness.

• Multi-Resolution Training Pipelines

Training at both low and high resolutions enables models to learn global context while retaining fine detail.

• **Domain Adaptation**

Adapting models to new regions improves generalization beyond Massachusetts (e.g., rural India, European cities).

• **Integration with Multimodal Data**

Incorporating **LiDAR**, **SAR** radar imagery, or vehicle GPS tracks can enhance road inference in occluded or shadowed areas.

9. Conclusion:

We compared four deep learning-based segmentation models as to how they performed when automatically extracting the roads in the high spatial resolution aerial images of the area of interest using U-Net, DeepLabV3+, PSPNet, and FPN. U-Net balanced accuracy, continuity, and inference speed best. It obtained the highest IoU and F1-score values of 0.2954 and 0.4556, respectively. Even with our limited performance on thin roads and roads in complex intersections, this baseline acts as a strong benchmark and shows which architectural components work best for road segmentation.

The survey highlights future directions such as transformer-based methods, learning topologies, domain adaptation, and multimodal fusion that have high potential to advance automated road extraction and resulting applications in navigation, geospatial intelligence, and infrastructure management.

10.References:

Deep Learning Architectures:

Ronneberger, O., Fischer, P., & Brox, T. (2015).
U-Net: Convolutional networks for biomedical image segmentation. In N. Navab et al. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Springer.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018).
Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 801–818.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017).
Pyramid scene parsing network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2881–2890.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017).
Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2117–2125.

Loss Functions & Optimizers:

Milletari, F., Navab, N., & Ahmadi, S. A. (2016).
V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017).
Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Dataset References:

Mnih, V. (2013).
Machine learning for aerial image labeling. (Doctoral dissertation, University of Toronto).

Massachusetts Roads Dataset. (2021).
Kaggle. https://www.kaggle.com/datasets/balraj98/massachusetts-roads-dataset