Road Extraction from Aerial Imagery Using Deep Learning

Frank Chen

Introduction

Accurate extraction of road networks from satellite imagery is essential for applications such as navigation, urban planning, and disaster response. Manual annotation of such data is labor-intensive and infeasible at scale, making automatic methods highly desirable. Deep learning, particularly convolutional neural networks (CNNs), has become the standard approach for semantic segmentation tasks, including road detection.

This project builds an enhanced UNet-based model for road segmentation, integrating these architectural improvements to better capture roads with varying scales and appearances in noisy satellite images.

Literature Review

The task of road extraction from satellite imagery is a subset of semantic segmentation, where each pixel in an image is classified into a semantic category, in this case, road or background. Among the many models developed for this purpose, UNet (Ronneberger et al., 2015) has become a foundational architecture due to its encoder-decoder structure with skip connections, which enables precise localization and efficient training even on limited datasets.

Subsequent developments have sought to improve upon UNet by enhancing its depth, contextual awareness, and focus. He et al. (2016) introduced residual connections in ResNet, which help neural networks train deeper architectures by mitigating vanishing gradient issues.

When integrated into UNet, these residual blocks allow the model to learn more complex feature hierarchies, which is particularly useful for distinguishing thin or partially occluded road segments.

Capturing contextual information at different spatial scales is another challenge in semantic segmentation. Chen et al. (2017) proposed Atrous Spatial Pyramid Pooling (ASPP), a module that applies parallel convolutions with various dilation rates to extract multi-scale features. ASPP is especially beneficial in remote sensing tasks, where objects such as roads may appear at varying sizes and orientations across an image.

Lastly, attention mechanisms have been incorporated into encoder-decoder models to enable spatially selective feature fusion. Oktay et al. (2018) introduced Attention Gates for medical image segmentation, which allow the network to focus on relevant structures and suppress background noise. This mechanism has since been applied in geospatial tasks, improving performance on segmentation of small or ambiguous features.

Together, these advancements form the basis of many successful segmentation models, including the architecture used in this project.

Methodology

The architecture used in this project is a modified version of the UNet convolutional neural network, designed to improve the quality of road segmentation from satellite imagery. The model is implemented in PyTorch and builds upon the traditional encoder-decoder structure by integrating several enhancements: Residual Blocks, ASPP, Attention Gates, and UpConvolution layers.

The encoder and decoder use residual convolution blocks to improve gradient flow and feature extraction, while maintaining the UNet’s symmetric structure. In the decoder, upsampling is performed via transposed convolutions, with attention gates enhancing the contribution of encoder features by focusing on relevant spatial regions. The final layer of the network is a 1x1 convolution that reduces the number of channels to one to produce a binary segmentation mask.

The training process uses a custom loss function that combines binary cross-entropy and the Tversky loss. The Tversky loss is a generalization of the Dice loss, with tunable parameters to control the penalties for false positives and false negatives. This choice of loss is well-suited for road extraction tasks, where class imbalance is significant and most of the pixels in the image are background, and only a small fraction corresponds to roads.

The model is trained using the Adam optimizer for 20 epochs on images resized to 256×256 pixels. A batch size of 8 is used, and training is accelerated using a CUDA-enabled GPU. During training and validation, the Intersection over Union (IoU) metric is computed to evaluate the segmentation accuracy. IoU is a standard metric in semantic segmentation that measures the overlap between the predicted and true masks.

The pipeline also includes visualization routines to display the original input image, the ground truth mask, and the predicted segmentation side by side. These visualizations are useful for qualitatively assessing the model’s performance and identifying typical failure cases.

Data Processing

This project uses the DeepGlobe Road Extraction dataset, which contains 6,226 RGB satellite images paired with binary masks indicating road regions. Each image has a resolution of 1024×1024 pixels, and the masks label road pixels as foreground and all other areas as background.

To facilitate training and evaluation, the dataset was split into training and test subsets in a 70:30 ratio. The images were resized to 256×256 pixels to reduce computational load while preserving essential structural information. Before training, all images were normalized using standard ImageNet statistics to ensure stable gradient flow during optimization.

A custom PyTorch dataset class was implemented to handle image loading, transformation, and pairing with the corresponding binary masks. Data augmentation techniques such as random flipping and rotation were considered but not extensively applied in the current version due to time constraints. The images were batched using a DataLoader with a batch size of 8 and shuffled during training to improve generalization.

No separate validation set was used. Instead, performance was monitored on the held-out test set. All training and inference were conducted on an NVIDIA RTX 3070 GPU.

Results

The training process took approximately 70 minutes on an NVIDIA RTX 3070 GPU. The model's segmentation performance was evaluated on the remaining 30% of the dataset using the Intersection over Union (IoU) metric.

The final model achieved a mean IoU score of 0.5259, indicating moderate overlap between predicted and ground truth road masks. This score represents a noticeable improvement over the baseline UNet model, particularly in the continuity and clarity of extracted road structures.

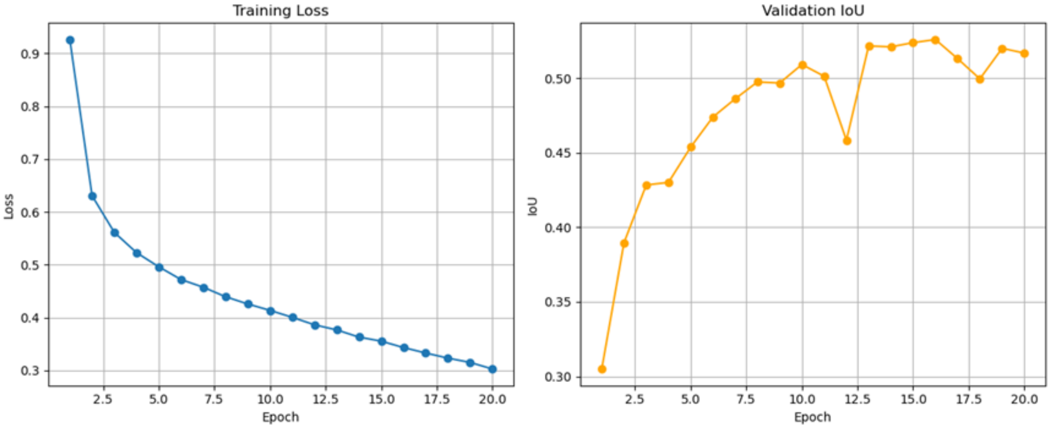

Fig 1. Training loss and IoU

Qualitative evaluation of the predicted masks showed that the enhanced model was more successful at identifying thin, curving, or partially occluded roads. The inclusion of Attention Gates contributed to more focused predictions, while ASPP helped capture road features across various spatial scales. UpConvolution layers provided smoother and more accurate reconstructions, especially near road boundaries.

Figure examples from the testing phase showed cleaner and more continuous road extractions compared to earlier versions of the model, with fewer false positives in non-road regions. These results suggest that the architectural modifications had a tangible impact on both quantitative performance and visual quality.

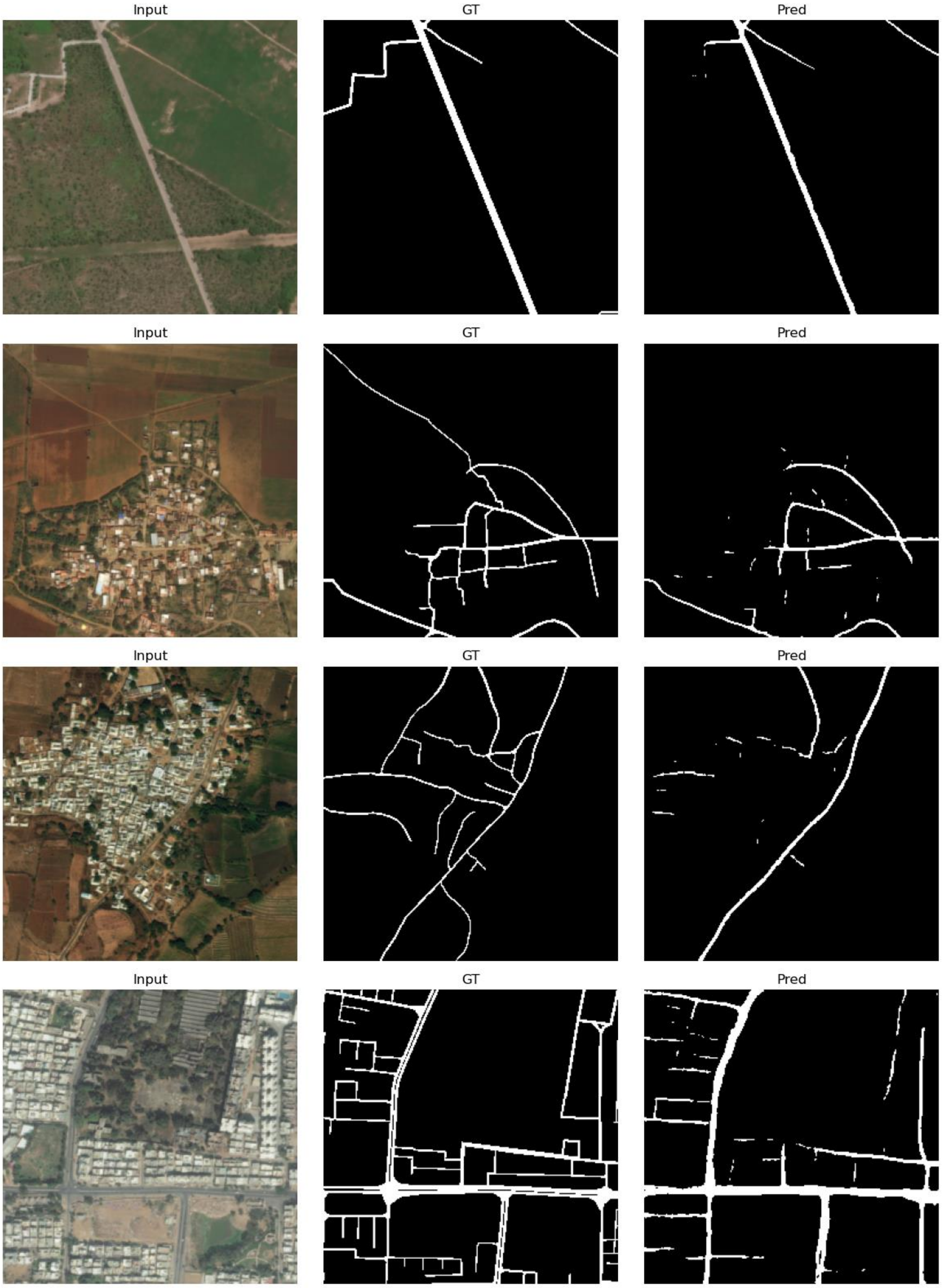

Fig 2. Comparison of true mask and predicted mask

Discussion and Limitations

The enhanced UNet model demonstrated improved performance in road segmentation tasks compared to a baseline architecture. By incorporating residual blocks, ASPP, attention mechanisms, and learnable upsampling, the model achieved a higher IoU score and produced qualitatively cleaner segmentation maps. These improvements validate the use of architectural enhancements commonly seen in advanced semantic segmentation literature.

Despite these gains, several limitations remain. The IoU score of 0.5259, while improved, indicates that the model still misses a significant portion of the road network. Visual inspection revealed that thin roads or roads that are partially covered by shadows, trees, or noises, were occasionally misclassified as non-roads. These errors may be due to the lack of contextual cues or limitations in the training data.

Additionally, the model was trained for only 20 epochs due to time and hardware constraints. With more extensive hyperparameter tuning, additional data augmentation, or longer training, performance could likely be improved. Furthermore, no post-processing techniques such as conditional random fields or morphological smoothing were applied, which might help improve road connectivity in the segmentation outputs.

Another limitation is the reliance on a single dataset. While the DeepGlobe dataset is representative, evaluating the model on other road extraction benchmarks would be necessary to assess its generalizability. Future work could also explore integrating other modalities, such as elevation or multispectral data, to provide additional context for better segmentation.

References

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer international publishing.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.