

Haul Road Extraction in Open-Pit Mines via Dual-Encoder RGB–DSM Transformer Fusion

Loghman Moradi ¹, Kamran Esmacili ¹

¹ Dept. of Civil and Mineral Engineering, University of Toronto, Toronto, - loghman.moradi@mail.utoronto.ca,
kamran.esmacili@utoronto.ca

Keywords: Haul Road Extraction, Road Mapping, Open-Pit Mine, UAV Imagery, Digital Surface Model, Intelligent Mining.

Abstract:

Haul roads are essential to open-pit mines, acting like the mine's circulatory system. Keeping accurate, up-to-date maps of these roads is critical for maintenance, safety, and efficient material handling, yet automating this task is challenging. Traditional deep learning models that rely only on RGB images often fail in mining environments, where road surfaces resemble bare earth, dusty terrain, or shadowed areas. To address this, we propose a dual-encoder transformer that combines UAV-captured RGB images with DSM data using stage-wise cross-attention, leveraging both visual and topographic information. Two SegFormer encoders process each data type separately, creating detailed feature representations that are fused at each stage. This allows the model to learn specialized information while sharing knowledge between modalities. A lightweight All-MLP decoder produces the final segmentation map. We tested our method on a high-resolution dataset of 12,000 tiles from the Mildred Lake open-pit mine in Fort McMurray, Canada. Our model achieves 80.8% mIoU, 88.7% F1-score, and 73.7% road accuracy, outperforming an RGB-only baseline by 3.3%, 2.4%, and 7.8 points, respectively. Ablation studies demonstrate that including DSM data consistently improves recall and road detection, especially in areas where RGB information alone is ambiguous or terrain is complex.

1. Introduction

The mining operations are mainly of two types: underground mining and surface mining. Surface mining includes open-pit mining, strip mining, and quarrying, which together dominate the mining industry, accounting for more than two-thirds of total mineral production annually (Ramani, 2012). Material haulage is one of the most resource-intensive aspects of the surface mining process. Haul roads are the backbone of the entire material-handling process. These roads are designed to facilitate the continuous transportation of ore and waste from the mining areas to processing plants and waste-dumping sites. As the cost of haulage alone may account for 30 to 50 percent of total open-pit mining costs, haul roads are key determinants of the economic viability and safe operation of the mining site (de Werk et al., 2017).

The increasing size of surface mining operations has created a need for a more demanding haul road infrastructure. Additionally, the increasing prevalence of ultra-class haul trucks with payloads over 400 tonnes poses several structural issues, including pavement degradation rates, rolling resistance, and unpredictable road performance. Inadequate maintenance of haul roads not only increases truck operating costs and fuel consumption rates but may also cause tires and machinery to fail. In addition, poor maintenance of haul roads contributes to increased dust generation, road instability, and greater environmental impact, all of which are subject to increasing levels of regulation in modern mining jurisdictions (Ozdemir and Kumral, 2019). These factors highlight the need to continuously monitor and maintain mine haul roads, and to document the road network in both temporal and spatial terms.

A foundational requirement for effective road monitoring and maintenance is precise haul road mapping. This information can be very useful in assessing and evaluating road geometry, identifying areas where roads are deteriorating, and improving route planning. Yet, there are significant challenges in keeping information up to date in active open-pit mining areas. Haul roads, in

fact, are constantly changing due to mining operations, with new roads being developed, existing roads being extended, modified or abandoned, and roads being frequently regraded. Moreover, unlike public roads, mining roads are not marked with standard signs and boundaries and have irregular, unstructured, and site-specific shapes and configurations, usually dictated by the ore-body's shape (Meneses and Sepúlveda, 2023). This has made conventional methods of road mapping labor-intensive, time-consuming, and practically difficult in active open-pit mining areas, leading to a significant difference between the actual state of the road network and available spatial data.

Extracting road data from active mining environments is inherently challenging. The approaches currently employed in road mapping in mining environments include a variety of methodologies. These include ground-based field surveys, GPS tracking of vehicles on the roads, and remotely sensed data, including optical satellite and Unmanned Aerial Vehicle (UAV) images, as well as Light Detection And Ranging (LiDAR) point cloud data. Although all the above approaches have their individual merits, they are limited by trade-offs in terms of cost-effectiveness, operational safety, and time. For example, field surveying is highly accurate in terms of position. However, it is labor-intensive and operationally hazardous. LiDAR provides rich 3D structural data. However, the cost of acquisition is high. With the increasing availability of high-resolution UAV and satellite images, along with the rapid development of deep learning techniques for the analysis of geospatial data, the possibilities for the automatic extraction of road data from remotely sensed images are now considerable, and this has significant implications in cost reduction and improving safety in surface mining operations.

The problem of road extraction from aerial and satellite imagery has attracted significant research attention, with most existing studies focusing on structured urban road networks with clear lane markings, road widths, and geometry (Liu et al., 2024; Yu et al., 2025). Unstructured roads, especially haul roads in open-pit mines as shown in Figure 1, pose a distinct set of problems:

irregular and dynamic geometry, lack of boundary markings, varied surface conditions, and high spectral similarity to the environment (Wang et al., 2026). These factors make existing research on urban road extraction unsuitable for addressing haul road extraction problems in mines, leading to an increased interest in domain-specific research in the field.



Figure 1- Aerial imagery samples of open-pit mine haul roads with extracted boundaries shown in red.

Previous work on deep learning-based mining road extraction utilized existing Convolutional Neural Network (CNN) architectures for UAV images and compared different architectures such as FCN (Long et al., 2015), UNet (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), and DeepLabV3+ (Chen et al., 2018). In particular, UNet-based architectures have shown robust performance; subsequent work has further improved results by incorporating attention mechanisms such as ECA (Wang et al., 2019) and contrast enhancement preprocessing with CLAHE (Yang et al., 2024). A two-stage approach extending this work used UNet-based segmentation and geometric post-processing to calculate road safety parameters, such as width, gradient, and turning radius (Medinac et al., 2020; Saputra et al., 2023). Other work has focused on road assessment instead of extraction; UAV data was utilized to calculate superelevation and gradient metrics from Digital Elevation Model and orthophotos (Medinac et al., 2020), while point cloud-based methods utilized optical detection along with vehicle vibration signals to calculate road roughness and cross-slope (Douglas et al., 2024).

More recent studies have incorporated novel architectural features to improve the delineation of road boundaries and connectivity. For example, the MD-LinkNeSt model used a ResNet50 network as a feature-encoding module to delineate fine road features (Gu et al., 2021). Similarly, the RATT-UNet model utilized residual attention units to suppress unnecessary activations of the features in delineating roads (Xiao et al., 2022). To address the complexities of the heterogeneous terrain in open-pit mines, modified U-Net architectures have incorporated multi-scale convolutional blocks with channel attention modules to jointly extract texture features and context information (Cui et al., 2024). Similarly, a D-SegNeXt model used a novel convolutional module, the "hybrid dilated convolutional module," specifically suited to handle the complexities of road networks in open-pit mines (Sunwen et al., 2024). Other novel architectures have utilized the transformer module to delineate the areas of traversal in

geometrically complex terrains. Similarly, modified DeepLabV3+ architectures have incorporated novel structures to improve the coverage of receptive fields in delineating dynamic road networks (Gu et al., 2022). In addition to supervised approaches for delineating haul roads, a semi-supervised framework, "UniMatch," was used to extract natural dirt roads from GaoFen-2 satellite images (Wang and Wang, 2024). The satellite images used in the framework exhibit texture characteristics similar to those of haul roads.

An important shortcoming common to most of the studies is the use of a single input modality. Monomodal segmentations are generally limited in their ability to transcend the boundaries of their input. In an open-pit mine environment, where haul roads tend to have spectral characteristics similar to those of the surrounding rock and exposed subgrade, a single modality may not provide adequate discriminative power. This spectral ambiguity motivates the incorporation of complementary modalities that capture structurally distinct scene properties.

The broader community of remote sensing and computer vision researchers has addressed the aforementioned limitation by increasingly publishing research on multimodal fusion. CMX (Zhang et al., 2023) was an early and influential contribution in this space — a cross-modal fusion framework for RGB-X segmentation built around a Feature Rectification Module that bidirectionally calibrates paired modality features and notably demonstrated that the same architecture generalizes across depth, thermal, polarization, event, and LiDAR inputs. Since then, transformer-based designs have largely become the architecture of choice for multimodal fusion. FTransUNet (Ma et al., 2024) combined CNN-based shallow feature extraction with a Fusion Vision Transformer that alternates between cross-attention and self-attention layers, progressively building up a joint multimodal representation. FTransDeepLab (Feng et al., 2025) took a different angle, grafting a multimodal rectification module and cross-attention fusion block onto a multiscale SegFormer encoder within the DeepLabv3+ framework. CFFormer (Zhao et al., 2024) went a step further by explicitly suppressing inter-modal noise before the cross-attention step — a design choice that reflects growing awareness that not all cross-modal signals are equally useful. On the scalability side, MultiSenseSeg (Wang et al., 2024) addressed the practical problem of handling arbitrary modality combinations through modality-specific lightweight experts and an adaptive matching module, making it possible to train a single model across both mono- and multi-modal configurations. MFTransNet (He et al., 2023) pursued a more efficiency-oriented direction, blending CNN and transformer components to keep computational costs manageable while still supporting multi-scale feature fusion for high-resolution remote sensing imagery.

The UAV platform is particularly well-suited to fill the identified gap. Its ability to capture high-resolution images at flexible time intervals enables frequent, cost-effective monitoring of road surfaces and changes typical of active mining operations. Beyond RGB orthophotos, UAV photogrammetric workflows can also produce a Digital Surface Model (DSM) as a byproduct of image

reconstruction, providing co-registered RGB and elevation data without extra sensor cost. These two modalities encode fundamentally different information: RGB captures surface spectral properties such as colour, texture, and albedo, while DSM represents three-dimensional topography, including road gradients and the sizes and shapes of berms around the road. The information from these modalities is complementary rather than redundant—conditions that complicate RGB interpretation, such as shadows or dust, often still reveal strong topographic signatures, and vice versa. Leveraging this complementarity through a carefully designed fusion architecture is a promising, yet underexplored, approach for robust haul road extraction.

To address this gap, this paper proposes a dual-encoder transformer architecture that performs stage-wise cross-attention fusion between parallel RGB and DSM encoder streams for automated haul road segmentation from UAV imagery in open-pit mining environments. By allowing each encoder stage to attend to the complementary representations of the other modality, the proposed method enables deep, hierarchical integration of spectral and topographic features without sacrificing the independent representational capacity of each stream.

2. Methodology

The proposed architecture, illustrated in Figure 2, is a dual-encoder transformer framework designed to jointly process RGB images and DSM data for haul road segmentation in open-pit mining environments. It comprises three main components: two parallel SegFormer (Xie et al., 2021) encoder branches that independently extract hierarchical features from each modality, a stage-wise cross-modal fusion module (Zhang et al., 2023) connecting the two branches at each encoder stage, and a lightweight All-MLP decoder that aggregates the fused multi-scale features to produce the final segmentation map. By keeping separate encoder pathways for each modality instead of concatenating inputs at the network entry, the design preserves the unique statistical characteristics of spectral and topographic data while enabling deep, hierarchical cross-modal interactions.

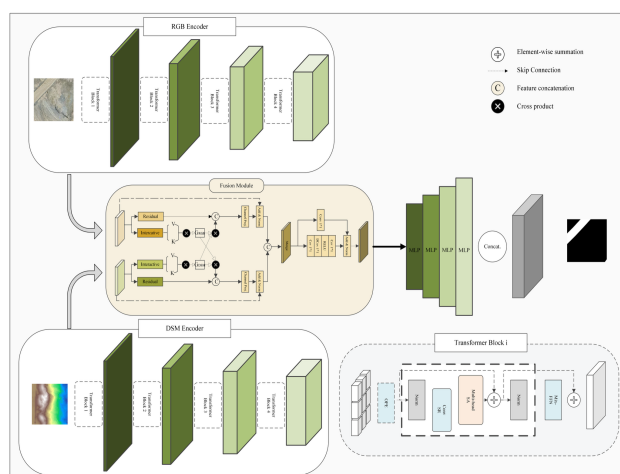


Figure 2- The proposed dual-encoder transformer framework for mine haul road segmentation, featuring parallel RGB and DSM

SegFormer encoders connected through stage-wise cross-attention fusion and an All-MLP decoder.

Both encoder branches use the Mix Transformer (MiT) backbone from SegFormer, chosen for its hierarchical multi-scale design and avoidance of fixed positional encodings—an advantage for handling UAV imagery captured at varying resolutions. Each MiT encoder generates four feature maps at successive down-sampling scales of $4\times$, $8\times$, $16\times$, and $32\times$ relative to the input, producing a feature pyramid that captures road-relevant cues across a broad range of spatial frequencies.

Each encoder stage stacks a series of transformer blocks, and each block follows the same three-part structure. First, an overlapping patch embedding layer tokenizes the feature map using a stride smaller than the patch size, preserving local continuity at patch boundaries and enhancing segmentation accuracy along edges. Second, an Efficient Self-Attention (ESA) module applies sequence reduction to the key and value projections, lowering the computational complexity of self-attention from quadratic to sub-quadratic relative to spatial resolution—an important efficiency gain for processing high-resolution UAV tiles. Third, a Mix Feed-Forward Network (Mix-FFN) introduces spatial awareness via a 3×3 depthwise convolution within the FFN, implicitly encoding local positional context through zero-padding, without requiring explicit positional encodings that would need interpolation at inference.

The RGB encoder branch processes three-channel optical images to capture surface spectral properties, including colour, texture, and albedo variations. The DSM encoder branch handles the single-channel elevation input, representing topographic structures such as road gradients, cross-sectional relief, and elevation discontinuities at excavation boundaries. Both branches use the same MiT architecture but are trained independently with separate weights, allowing each to specialize in the unique statistical characteristics of its respective modality. At each of the four encoder stages, the output feature maps from the RGB and DSM branches are passed through a cross-modal fusion module, shown in detail in Figure 3. The fusion module operates in two sequential parts.

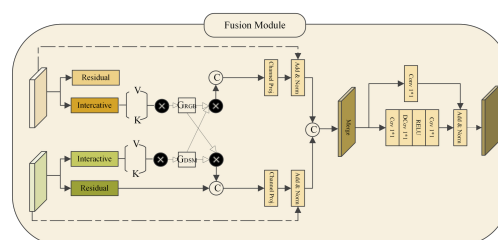


Figure 3- Architecture of the cross-modal fusion module with cross-attention between RGB and DSM feature streams.

In the first section, a cross-attention mechanism enables global information exchange between the two modality streams. Feature maps from each branch are flattened and projected through linear embeddings to create two vectors of equal size: a residual vector that preserves the modality's own representation, and an

interactive vector that participates in cross-modal attention. The interactive vector from one branch serves as the query, while the interactive vector from the other branch provides the key-value pairs, allowing each modality to selectively attend to and be influenced by the other. The resulting cross-attention context vector is then element-wise multiplied with the branch’s interactive vector, producing a modality-specific representation enriched with complementary cross-modal information. This design allows the RGB branch to incorporate topographic cues at each scale and the DSM branch to leverage spectral and textural cues, without merging the two streams into a single undifferentiated representation.

In the second part, the outputs from the cross-attention modules of both branches are concatenated and processed by a mixed-channel embedding layer. This layer projects the combined representation back to the original feature map dimensionality, producing a single fused feature map at each encoder stage that preserves spatial resolution while integrating information from both modalities.

The fused multi-scale feature maps from all four stages are fed into the All-MLP decoder head inherited from SegFormer. Each stage’s fused feature map first passes through a lightweight MLP layer to unify channel dimensionality across scales, then is upsampled to one-quarter of the original input resolution. The four upsampled feature maps are concatenated along the channel dimension and passed through a final MLP fusion layer that combines multi-scale context into a single feature volume. A linear projection followed by bilinear upsampling then produces the final full-resolution binary segmentation map, distinguishing haul road from non-road terrain.

3. Experiments

The dataset was collected across multiple acquisition campaigns at the Mildred Lake open-pit mine in Fort McMurray, Alberta, Canada, spanning different seasons to capture varying surface conditions. Each campaign used a DJI M350 drone equipped with a 50-megapixel camera, flying at 100 m altitude with 80% forward overlap and 60% sidelap, with a minimum of four flight strips per mission. The collected imagery was processed in Agisoft Metashape (“Agisoft Metashape,” 2024) to generate orthophotos, DSMs, and point clouds. Both the orthophotos and DSMs as shown in Figure 4, have a ground sampling distance of less than 10 cm, providing fine-grained detail of road surfaces, boundaries, and topography, essential for haul road segmentation.

Road annotations were manually created in ArcGIS Pro, with all haul road surfaces marked as binary masks. The complete dataset contains roughly 12,000 image tiles at 1024×1024 pixels. To maintain spatial and temporal independence between subsets and prevent data leakage, the dataset was split based on geographic location and acquisition date rather than random sampling, resulting in training, validation, and test sets of 80%, 10%, and 10%, respectively. This approach ensures the model is evaluated on spatially and temporally distinct tiles, offering a more realistic

measure of generalization performance in active mining environments.

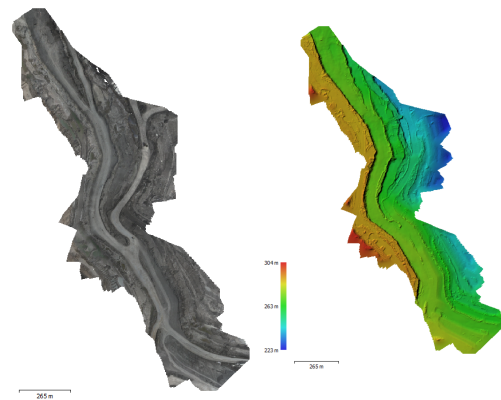


Figure 4- Sample orthophoto and corresponding DSM generated from UAV data.

The model’s performance is evaluated using five standard semantic segmentation metrics. Overall Accuracy measures the proportion of correctly classified pixels across all classes. Precision assesses the reliability of positive predictions, calculating the fraction of predicted road pixels that are truly road. Recall measures the model’s sensitivity to road pixels, representing the proportion of actual road pixels correctly identified—a critical metric in mining where missed roads can have operational consequences. The F1-score, as the harmonic mean of precision and recall, provides a balanced assessment, especially under class imbalance conditions where background pixels dominate. Mean Intersection over Union (mIoU) serves as the primary metric, computing the overlap-to-union ratio between predicted and ground-truth maps, averaged across all classes, offering a comprehensive measure of segmentation quality. Additionally, Road Accuracy (R. Acc) is reported separately to specifically quantify prediction quality for the road class, independent of background performance.

All experiments were carried out on a system equipped with four NVIDIA H100 GPUs with 80GB of memory. Models were implemented in Python 3.11 using PyTorch and CUDA 12.2. Each model variant was trained from scratch under identical conditions to ensure a fair and controlled comparison. Models were trained for at least 120 epochs to ensure convergence, using the AdamW optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 0.2. A cosine annealing learning rate scheduler gradually reduced the learning rate throughout training. To maintain consistency across experiments, all five model variants used the same hyperparameters, data pipeline, and augmentation strategy.

To systematically assess the impact of DSM integration and encoder backbone capacity, five model configurations are compared as shown in Table 1. Models 1 and 2 serve as RGB-only baselines, using SegFormer-B0 and SegFormer-B2 backbones, respectively, to establish the upper bound of monomodal performance. Models 3 through 5 introduce DSM as a second input modality via the proposed dual-encoder cross-attention fusion framework. DSM encoder backbone capacity is progressively increased from B0 to B1 to B2, while Models 4 and 5 pair these

with a fixed B2 RGB encoder. This setup allows controlled evaluation of three factors: the benefit of adding DSM over RGB-only input, the influence of RGB backbone size, and the effect of DSM encoder capacity on fusion quality.

Table 1- Quantitative comparison of five model configurations across all evaluation metrics (%).

#	RGB	DSM	Acc	mIoU	F1	Recall	Precision	R. Acc
1	B0	-	93.1	72.1	82.1	76.4	89.8	58.4
2	B2	-	94.7	77.5	86.3	80.6	93.7	65.9
3	B0	B0	93.8	76.7	85.8	85.4	86.2	74.5
4	B2	B1	94.6	78.4	87	84.8	89.4	73.3
5	B2	B2	95.4	80.8	88.7	85	92.9	73.7

Among the RGB-only baselines, the SegFormer-B2 backbone (Model 2) achieves 77.5% mIoU and 86.3% F1, outperforming the B0 variant (Model 1) by 5.4 and 4.2 percentage points, respectively. This demonstrates that a larger backbone improves road segmentation using spectral data alone. However, road-class accuracy remains relatively low at 65.9%, suggesting that appearance-based features alone often struggle to reliably distinguish road surfaces from spectrally similar terrain.

Incorporating DSM data through the proposed stage-wise cross-attention fusion consistently improves performance across all configurations. For example, Model 3, which uses two B0 encoders, achieves a road accuracy of 74.5%, an increase of 16.1 percentage points over the RGB-only baseline, while recall rises from 76.4% to 85.4%. This demonstrates that elevation-derived features help recover road pixels that spectral information alone might miss. The improvement in recall is particularly valuable in open-pit mines, where haul roads can be geometrically distinct but spectrally ambiguous due to dust, shadows, and varying surface conditions.

The full B2+B2 configuration in Model 5 is the strongest result across the board — 95.4% overall accuracy, 80.8% mIoU, 88.7% F1, and 73.7% road accuracy — and the gains over the RGB-only B2 baseline are consistent rather than concentrated in a single metric. The +7.8 percentage point jump in road accuracy is the most practically significant of these, but the mIoU and F1 improvements hold up as well. One result worth examining more closely is precision: Model 5 recovers to 92.9%, nearly matching the RGB-only B2 baseline at 93.7%, after a noticeable drop in the smaller B0+B0 configuration (Model 3: 86.2%). This pattern suggests that when the DSM encoder lacks sufficient capacity, it introduces spurious activations, likely from elevation noise or non-road topographic features, that inflate false positives. Scaling the DSM encoder up to B2 appears to give the network sufficient representational capacity to distinguish genuinely road-relevant topographic cues from background elevation variation, restoring precision without sacrificing the recall gains from DSM fusion.

Comparing Models 4 and 5 highlights the impact of the DSM encoder capacity on fusion quality. Model 4, which pairs a lighter B1 DSM encoder with the B2 RGB encoder, achieves 78.4% mIoU and 87.0% F1, already outperforming the RGB-only B2 baseline. Upgrading the DSM encoder to B2 in Model 5 provides an additional +2.4% mIoU and +1.7% F1, showing that richer DSM feature representations at each encoder stage produce more discriminative cross-modal attention maps and enhance the decoder’s ability to accurately delineate road boundaries.

Across all configurations, the pattern is consistent: DSM integration helps, cross-attention fusion handles it well, and the gains become more reliable as backbone capacity increases. The stage-wise design, in particular, seems to be earning its complexity; improvements show up not just in aggregate metrics like mIoU but also in road-class recall specifically, where the practical value of elevation data in spectrally ambiguous mining terrain would be expected to show up most.

Beyond the numbers, Figure 5 provides a qualitative view of how the five model configurations behave on representative tiles. Green, red, and yellow overlays mark true positives, false positives, and false negatives, respectively. Row 1 shows a particularly difficult scenario with dark, shadowed terrain that closely resembles road surfaces in terms of spectral reflectance. Both RGB-only models (B0 and B2) generate significant false positives, misclassifying shadowed areas as road. Bringing in DSM progressively cleans this up across the other configurations. It is a fairly direct illustration of what elevation data adds here: shadows and surface texture variation confound spectral discrimination, but the topographic signature of non-road terrain remains distinct regardless of lighting conditions.

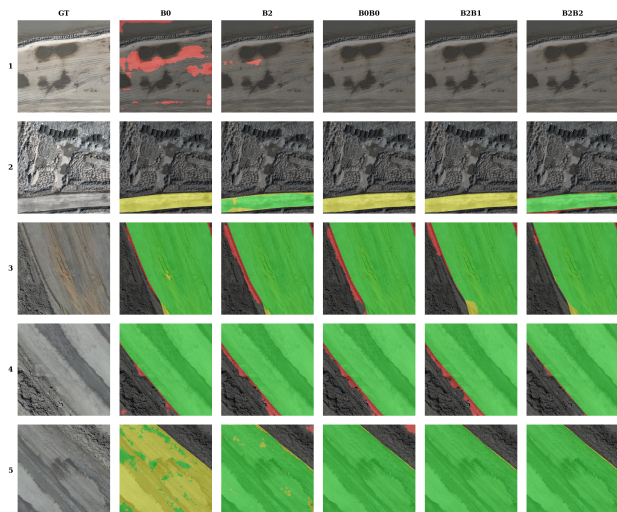


Figure 5- Qualitative comparison of segmentation outputs across five representative tiles for all model configurations (green: TP, red: FP, yellow: FN). GT is ground truth, and B is encoder version for both RGB and RGB-DSM.

Row 2 shows a narrow road segment on steep, structured terrain. The RGB-only, especially the B0 model, produces significant

false negatives, failing to capture the full extent of the road. The dual-encoder RGB-DSM models recover much of this missing coverage, with the B2+B2 configuration producing the most complete prediction. This indicates that the stage-wise cross-attention mechanism effectively uses topographic gradients to detect geometrically distinct road surfaces, even under challenging visual conditions.

Rows 3 and 4 display diagonal road segments, where all models capture the general road coverage reasonably well. However, the RGB-only models show boundary errors, visible as red fringes along the road edges. In contrast, the models with DSM data generate sharper and more accurate boundaries, highlighting the contribution of DSM features to precise boundary localization at each encoder stage since road edges in open-pit mines correspond to real elevation discontinuities, and having DSM features available at each encoder stage gives the network a geometrically grounded reference for where the road surface actually ends.

Row 5 presents a different kind of challenge — the road occupies a large fraction of the tile, and spectral appearance shifts significantly across the scene, making it hard for appearance-based models to maintain consistent predictions. The B0 RGB-only model has the most difficulty here, missing substantial road areas entirely. The B2 RGB-only model does better but still produces a fragmented output — the predictions are riddled with holes where the model loses confidence mid-road, likely where surface colour or texture changes abruptly. The B2+B2 model handles variability much more gracefully, producing a coherent, largely complete segmentation with few missed regions or spurious detections. When road appearance is this inconsistent across a single tile, having elevation data as a stable secondary reference appears to keep the predictions from fragmenting.

Across all five rows, the qualitative picture aligns well with what the numbers showed in Table 1. DSM integration reliably reduces missed detections in geometrically complex terrain and reins in false positives where spectral ambiguity would otherwise mislead an RGB-only model. The B2+B2 configuration is the most consistent performer throughout, and the variety of scenes in the Figure 5 — shadowed terrain, steep cuts, high road coverage, spectral variability — provides reasonable confidence that these improvements are not scene-specific artifacts but reflect a genuine, generalizable benefit of the proposed fusion approach.

The computational complexity and inference time of all five model configurations are summarized in Table 2. Although Model 3, which is the lightweight B0+B0 dual encoder model, increases the parameter count by only 4.94M over the RGB-only B0 model, it also increases the inference time approximately three times over to 163.9 ms from 52.8 ms. This is a reasonable trade-off for the accuracy gains that this model offers. On the other hand, the top-performing model, Model 5, which is the B2+B2 model, achieves 58.6M parameters and 684.26 GFLOPs with a mean inference time of 382 ms per tile. This is approximately double the computational cost of the RGB-only B2 model (Model 2) but is within the acceptable range for a near-real-time mapping pipeline where processing occurs post-flight rather than

on-board. Overall, the computational overhead introduced by DSM fusion is justified by the segmentation gains, and the framework remains practical for post-flight haul road mapping workflows.

Table 2- Computational complexity of evaluated models (M = million, ms = milliseconds).

#	RGB	DSM	Params (M)	GFLOPs	Inference Mean (ms)
1	B0	-	3.71	54.15	52.8
2	B2	-	27.35	453.71	192.4
3	B0	B0	8.65	101.34	163.9
4	B2	B1	47.56	620.1	323.5
5	B2	B2	58.6	684.26	382

4. Conclusion

This paper set out to address a specific and practical limitation of existing haul-road extraction methods: their reliance on RGB imagery alone in environments where spectral information is frequently unreliable. The proposed dual-encoder architecture addresses this by running parallel SegFormer branches for RGB and DSM data, allowing each modality to develop its own feature representations before they are brought together through stage-wise cross-attention fusion. Rather than a late concatenation or a simple channel stack at the input, the integration happens progressively at every encoder stage, and based on the results, it appears to matter.

The experiments, conducted on an ultra-high-resolution multi-season dataset from the Mildred Lake open-pit mine, consistently showed that bringing DSM into the picture improves segmentation in ways that scaling the RGB backbone alone cannot replicate. The full B2+B2 configuration reached 80.8% mIoU, 88.7% F1-score, and 73.7% road accuracy, gains of 3.3%, 2.4%, and 7.8 percentage points over the strongest RGB-only baseline. The road accuracy gap is particularly telling, since that is precisely the metric most sensitive to whether the model actually understands what a haul road looks like versus what merely resembles one spectrally. The qualitative results reinforce this: the fusion model holds up in shadowed terrain and dust-covered surfaces where RGB-only predictions fall apart, and recovers road detections in topographically distinct areas that spectral features alone would miss entirely.

A few directions stand out as natural next steps. The most immediate is rethinking what gets fed into the DSM encoder. Raw elevation captures absolute height but says relatively little about local surface geometry, and for road detection, it is arguably the geometry that matters most. Slope maps and Laplacian-filtered surfaces encode gradient discontinuities and local curvature, respectively, and both are straightforward to compute from the DSM before it ever reaches the network. Whether substituting or supplementing the raw elevation channel with these derived representations improves boundary delineation and reduces false

positives is an open question worth investigating directly. On a related note, frequency-domain representations of both RGB and DSM inputs may surface periodic structural patterns in road geometry that spatial-domain features tend to miss. This is a less-explored but potentially useful direction.

Outside of these methodological considerations, the operational limitations of UAV-based data collection in an open pit mine setting should also be acknowledged. For one, flight windows are often constrained by blasting operations and heavy equipment movements. Moreover, data collection is not possible at night, which is a major drawback considering that mining operations are not restricted to day shifts. Finally, processing is not instantaneous; that is, from the end of flight to orthophoto and DSM generation using Metashape, there is a non-trivial process that must be undertaken before any segmentation can be performed. However, this process is significantly faster than traditional approaches and can be considered near-real-time.

The other open question is generalizability. The Mildred Lake dataset covers multiple seasons, which helps, but it represents a single mine in a specific geological and climatic setting. Testing the framework across sites with different surface materials, vegetation cover, and road construction practices would provide a clearer picture of where the approach holds up and where it needs adaptation. Alongside this, a rigorous head-to-head comparison against other multimodal segmentation methods under a unified evaluation protocol remains to be done. The results here are promising, but properly situating the approach within the broader fusion literature will require that comparison to be made explicit.

Acknowledgement

The authors would like to thank the Natural Science and Engineering Research Council of Canada (NSERC) and Suncor for their financial support.

References

- Agisoft Metashape, 2024.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818.
- Cui, P., Meng, X., Zhang, W., 2024. Road Extraction From High-Resolution Remote Sensing Images of Open-Pit Mine Using D-SegNeXt. *IEEE Geosci. Remote Sens. Lett.* 21, 1–5. <https://doi.org/10.1109/LGRS.2024.3397949>
- de Werk, M., Ozdemir, B., Ragoub, B., Dun-brack, T., Kumral, M., 2017. Cost analysis of material handling systems in open pit mining: Case study on an iron ore prefeasibility study. *Eng. Econ.* 62, 369–386.
- Douglas, A., Langenderfer, M., Johnson, C., 2024. Road Condition Monitoring Utilizing UAV Photogrammetry Aligned to Principal Curve of Mine Haul Truck Path. *Min. Metall. Explor.* 41, 61–72. <https://doi.org/10.1007/s42461-023-00877-0>
- Feng, H., Hu, Q., Zhao, P., Wang, S., Ai, M., Zheng, D., Liu, T., 2025. FTrans-DeepLab: Multimodal fusion transform-er-based DeepLabv3+ for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.*
- Gu, Q., Xue, B., Ruan, S., Li, X., 2021. A road extraction method for intelligent dispatch-ing based on MD-LinkNeSt network in open-pit mine. *Int. J. Min. Reclam. Envi-ron.* 35, 656–669. <https://doi.org/10.1080/17480930.2021.1949800>
- Gu, Q., Xue, B., Song, J., Li, X., Wang, Q., 2022. A High-Precision Road Network Construction Method Based on Deep Learning for Unmanned Vehicle in Open Pit. *Min. Metall. Explor.* 39, 397–411. <https://doi.org/10.1007/s42461-022-00548-6>
- He, S., Yang, H., Zhang, X., Li, X., 2023. MFTransNet: A multimodal fusion with CNN-transformer network for semantic segmentation of HSR remote sensing im-ages. *Mathematics* 11, 722.
- Liu, Z., Wu, W., Wang, D., Cui, B., Gu, X., 2024. Automatic extraction and 3D modeling of real road scenes using UAV imagery and deep learning semantic seg-mentation. *Int. J. Digit. Earth* 17, 2365970. <https://doi.org/10.1080/17538947.2024.2365970>
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.
- Ma, X., Zhang, X., Pun, M.-O., Liu, M., 2024. A multilevel multimodal fusion trans-former for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* 62, 1–15.
- Medinac, F., Bamford, T., Hart, M., Kowalczyk, M., Esmaeili, K., 2020. Haul Road Monitoring in Open Pit Mines Using Unmanned Aerial Vehicles: a Case Study at Bald Mountain Mine Site. *Min. Metall. Explor.* 37, 1877–1883. <https://doi.org/10.1007/s42461-020-00291-w>
- Meneses, D., Sepúlveda, F.D., 2023. Modeling productivity reduction and fuel consumption in open-pit mining trucks by considering the temporary deterioration of mining roads through discrete-event simulation. *Mining* 3, 96–105.
- Ozdemir, B., Kumral, M., 2019. Simulation-based optimization of truck-shovel material handling systems in multi-pit surface mines. *Simul. Model. Pract. Theory* 95, 36–48. <https://doi.org/10.1016/j.simpat.2019.04.006>
- Ramani, R.V., 2012. Surface mining technology: progress and prospects. *Procedia Eng.* 46, 9–21.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

- Saputra, Z., Sakti, A.D., Firmana, A., Ignatius, M., Hede, A.N.H., Saepuloh, A., 2023. Enhanced mine road monitoring using unmanned aerial vehicles and deep-learning approach. *Remote Sens. Appl. Soc. Environ.* 32, 101080. <https://doi.org/10.1016/j.rsase.2023.101080>
- Sunwen, D., Yanwei, C., Guorui, F., 2024. Road Detection in Open-Pit Mining Areas Using D-DMCATNet Network Based on UAV Images. *IEEE Geosci. Remote Sens. Lett.* 1–1. <https://doi.org/10.1109/LGRS.2024.3477496>
- Wang, M., Wang, J., 2024. Dirty road extraction from GF-2 images by semi-supervised deep learning method for arid and semi-arid regions of southern Mongolia. *Int. J. Digit. Earth* 17, 2384631.
- Wang, N., Wang, X., Pan, Y., Lei, L., Yao, W., Zhong, Y., 2026. Identifying rural roads in remote sensing imagery: From benchmark dataset to coarse-to-fine extraction network—A case study in China. *ISPRS J. Photogramm. Remote Sens.* 231, 487–506.
- Wang, Q., Chen, W., Huang, Z., Tang, H., Yang, L., 2024. Multi-SenseSeg: A cost-effective unified multimodal semantic segmentation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* 62, 1–24.
- Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2019. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks.
- Xiao, D., Yin, L., Fu, Y., 2022. Open-Pit Mine Road Extraction From High-Resolution Remote Sensing Images Using RATT-UNet. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. <https://doi.org/10.1109/LGRS.2021.3065148>
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Al-varez, J.M., Luo, P., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers.
- Yang, Y., Zhou, W., Jiskani, I.M., Wang, Z., 2024. Extracting unstructured roads for smart Open-Pit mines based on computer vision: Implications for intelligent mining. *Expert Syst. Appl.* 249, 123628. <https://doi.org/10.1016/j.eswa.2024.123628>
- Yu, Z., Chen, Z., Xiao, K., Lei, X., Tang, R., He, Q., Sun, Z., Guo, H., 2025. SegRoadv2: a hybrid deformable self-attention and convolutional network for road extraction with connectivity structure. *Int. J. Digit. Earth* 18, 2480267. <https://doi.org/10.1080/17538947.2025.2480267>
- Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelwagen, R., 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Trans. Intell. Transp. Syst.* 24, 14679–14694.
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2881–2890.
- Zhao, J., Zhang, M., Zhou, Z., Wang, Z., Lang, F., Shi, H., Zheng, N., 2024. CFFormer: A cross-fusion transformer framework for the semantic segmentation of multi-source remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 63, 1–17.