CENTERLINENET: PATCH-ALIGNED SUPERVISION FOR THIN ROAD CENTERLINE EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Road networks evolve over time, requiring frequent map updates. AI tools can assist with this task; however, methods based on raster segmentation followed by thinning, skeletonization, or automatic tracing may fail to capture the local structure of road networks, increasing the burden on human annotators. Our goal is to directly predict thin centerline representations that reflect structural patterns used by annotators, particularly at intersections. A secondary goal is to scale training by learning from variable-quality vector data, such as OpenStreetMap, rather than relying on precisely aligned segmentation masks that are difficult to produce at scale. A key challenge is spatial misalignment in training data: while minor for thick segmentation masks, even small shifts become a major obstacle when learning thin centerlines, as pixel-wise losses are disproportionately affected. We propose CenterlineNet, a weakly supervised model that addresses this challenge with a patch alignment loss that compares local neighborhoods instead of individual pixels. This loss matches each predicted neighborhood to its nearest annotated centerline, enabling flexible alignment within a distance tolerance. We present two variants, basic and reciprocal, with the latter handling many-to-one mappings via softmax-in-group weighting, and introduce an intersection-aware component that specifically targets road junctions to improve connectivity.

1 INTRODUCTION

The extraction of road networks from remote sensing imagery is a fundamental task in computer vision with applications including autonomous driving, urban planning, emergency response, and geographic information systems. Despite advances in deep learning for semantic segmentation, road extraction remains challenging due to the thin, elongated nature of road structures, complex structure relationships, and spatial uncertainties in both imagery and ground truth annotations.

Traditional approaches rely on encoder–decoder architectures such as U-Net (Ronneberger et al., 2015) and DeepLabV3+ (Chen et al., 2018) trained with pixel-wise loss functions like binary cross-entropy or Dice loss. More recent work proposes stronger backbones such as CoANet (Liu et al., 2022) and MSMDFF-Net (Zhang et al., 2023), which incorporate multi-scale context and feature fusion to better handle thin and complex road structures. However, these methods still train with pixel-wise losses that implicitly assume perfect spatial alignment between predictions and labels—an assumption often violated in real-world deployments. Misalignment arises from (1) registration errors between imagery and vector annotations, (2) subjectivity and inconsistency in how annotators interpret road geometry, (3) temporal differences between image acquisition and ground-truth creation, and (4) visual ambiguity or occlusion. Fig. 1 illustrates how accurate centerline predictions can disagree with annotated roads under pixel-wise comparison despite capturing the correct structure.

Our contributions are threefold: (1) a patch alignment loss that compares local patches under a label-derived offset field to tolerate spatial offsets while preserving road structure; (2) a reciprocal formulation with softmax-in-group weighting to resolve many-to-one mappings between predictions and ground truth; and (3) an intersection-aware loss component to improve connectivity at road junctions. We demonstrate that CenterlineNet achieves competitive performance on road extraction tasks while remaining robust to annotation noise and spatial misalignments.

Figure 1: Four common road misalignments. **(1)** Ground-truth (GT) pixels displaced from the true road centerline due to registration error; **(2)** annotation variability, GT centerlines include extra edge-line segments; **(3)** temporal mismatches, new roads appear to have been added to forested areas after imagery acquisition; and **(4)** visual ambiguities, roads are missing from GT.

2 Related Work

We organize related work into three areas relevant to our approach: (1) road extraction using traditional and deep learning-based methods, (2) segmentation techniques designed to handle spatial misalignment, and (3) approaches that enforce structural consistency in thin, connected structures. While these lines of research address different aspects of the problem, none fully resolve the challenge of learning from spatially uncertain labels while preserving structural fidelity.

Early methods for road extraction relied on handcrafted features, edge detection, and mathematical morphology (Mena, 2003). Among geometric approaches, Hu et al. (2007) introduced a spoke wheel operator and Fourier-based shape classification for road tracking. Hu et al. (2007) expanded this into a road network extraction system using spoke-based tracking with pruning to eliminate false roads. Deep learning drastically improved road extraction performance. U-Net (Ronneberger et al., 2015) and DeepLab (Chen et al., 2018) became popular for semantic segmentation of geospatial imagery. Modern architectures have been tailored to road extraction: Zhang et al. (2018) introduced Deep Residual U-Net, Zhou et al. (2018) proposed D-LinkNet with dilated convolutions, and He et al. (2019) integrated ASPP with U-Net using structural similarity loss. However, these advances still rely on pixel-wise loss functions that assume perfect spatial alignment. While achieving good performance on aligned benchmarks, they suffer degradation when one attempts to learn misaligned vector maps, such as crowd-sourced annotations from OpenStreetMap.

Despite these advances, spatial misalignment remains a key obstacle for road network learning. One early attempt to address this was the work of (Batra et al., 2019), who proposed a multi-branch CNN that predicts both per-pixel segmentation and local orientation cues, using orientation information to encourage structurally coherent roads. Their framework also included a connectivity refinement module applied after inference, combining orientation-guided training with post-processing refinement. While effective, their approach remains fundamentally tied to pixel-wise loss over thicker representations and does not directly address misalignment at fine granularity.

Other strategies have attempted to tolerate misalignment indirectly. Some works apply relaxed evaluation metrics, buffered ground truth, or morphological post-processing (Sun et al., 2019), but these adjustments only occur after training and do not solve the underlying objective mismatch. The STEAL framework (Acuna et al., 2019) addressed weakly aligned labels through iterative self-training with an explicit alignment loss, gradually refining noisy annotations. Related work in remote sensing has also sought to improve the quality of crowdsourced supervision by aligning OpenStreetMap-derived labels to imagery before training (Zhang et al., 2020). While these methods reduce the effects of noisy or shifted labels, they still depend on heuristics or pre-alignment steps. They do not address misalignment tolerance within the training loss itself.

A third line of research emphasizes preserving connectivity in thin structures, since road networks are typically represented as centerlines forming graph-like structures where broken connections severely impact usability. Mosinska et al. (2018) pioneered structure-aware loss functions informed by persistent homology concepts. Building on this, Yuan & Xu (2022) developed GapLoss to explicitly reduce gaps in predicted road networks.

Some multi-task approaches explicitly extract road centerlines. Wei et al. (2020) presented a two-step CNN that first segments roads and then traces centerlines. Similarly, Lu et al. (2021) proposed MRENet, which performs simultaneous road surface segmentation and centerline extraction. However, in contrast to these methods, our approach enables robust centerline extraction directly through its loss formulation, greatly simplifying downstream post-processing or tracing steps should they be employed. Recent innovations like Skeleton Recall Loss Kirchhoff et al. (2024) focus on maximizing overlap between predicted and ground-truth skeletons, thereby preserving connectivity in thin structure segmentation. However, these structure-aware approaches still assume precise spatial alignment. When misalignment occurs, they may penalize structural correct predictions that are spatially offset. However, existing methods either assume spatial alignment, address misalignment only with post-processing or evaluation adjustments, or fail to resolve many-to-one ambiguities along thin centerlines. None directly preserve connectivity under misaligned supervision.

3 METHODOLOGY

The accurate extraction of road centerlines from remote sensing imagery remains a complex challenge due to the presence of spatial misalignments and the intricate structure of road networks. To address these issues, we introduce a loss formulation tailored for singleton centerline localization, where "singleton" denotes the nearly one-pixel-wide centerline representation used as ground truth rather than thick road polygons. The following section outlines the methodological framework, presenting the network design, loss functions, and training strategies employed in our approach.

3.1 DEEPLABUNETPRECISE ARCHITECTURE

CenterlineNet uses a hybrid backbone, which we call DeepLabUNetPrecise. The encoder follows DeepLabV3+, using a ResNet-101 with Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context. Unlike prior hybrid models (He et al., 2019; Zhou et al., 2018) that pool to $1/16$ or $1/32$ scale, our encoder stops at $1/8$ resolution and substitutes atrous convolutions for deeper pooling. This preserves finer feature maps while still expanding the receptive field. The decoder then upsamples through U-Net–style skip connections to restore predictions at full input resolution, which is essential for reconstructing one-pixel-wide road centerlines.

The architecture itself follows established hybrid patterns; our main contribution lies in the loss design. The role of DeepLabUNetPrecise is to provide sufficient resolution and feature detail so that the proposed patch alignment and intersection-aware losses can operate effectively.

3.2 LOSS FUNCTIONS

Our central idea is a patch alignment loss that tolerates spatial misalignment by comparing small neighborhoods rather than individual pixels. For each prediction location, we align its predicted logit-patch (logits of the K nearest pixels centered on a prediction) to the most relevant ground-truth neighborhood and measure per-patch cross-entropy. This preserves thin structures while allowing small spatial shifts.

For every pixel $\mathbf{x} = (x, y)$, we compute an offset vector $\mathbf{v}(\mathbf{x})$ that points to the nearest ground-truth centerline pixel. This offset field serves two roles: (i) it defines a tolerance band around the centerline (used later via a binary mask to include/exclude locations), and (ii) it provides the correspondence needed to extract and compare a predicted patch $\widehat{\mathbf{p}}(\mathbf{x})$ with the ground-truth patch $\mathbf{p}(\mathbf{x}+\mathbf{v}(\mathbf{x}))$. The exact loss expressions and masks are given in the subsections that follow.

We train CenterlineNet with a weighted sum of four terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{patch}} + \alpha\,\mathcal{L}_{\text{fp}} + \beta\,\mathcal{L}_{\text{singleton}} + \gamma\,\mathcal{L}_{\text{intersection}}, \tag{1}$$

where $\alpha = 5.0$, $\beta = 0.5$, and $\gamma = 1.0$ in our experiments. Each term is defined below.

3.2.1 RECIPROCAL SOFTMAX-IN-GROUP WEIGHTING

To resolve many-to-one correspondences, each pixel $\mathbf{x}$ that maps to the same ground-truth location as others is assigned a reciprocal weight $w_{\mathbf{x}}$. Let $G(\mathbf{x})$ denote this group of pixels and τ a tempera-

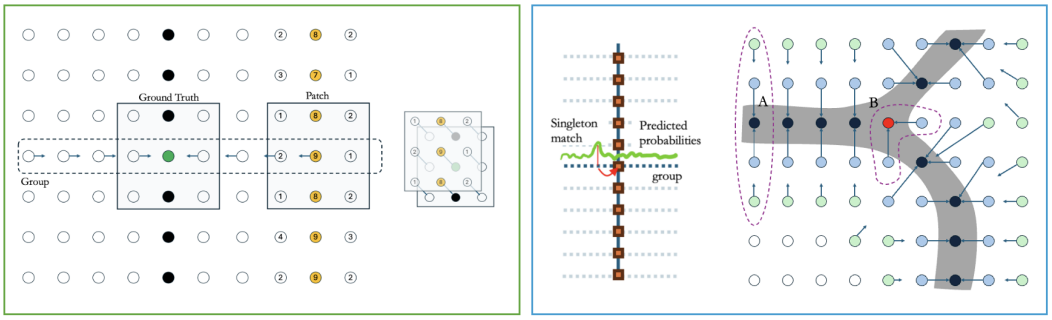

Figure 2: Patch alignment and reciprocal grouping. Left: patch alignment—arrows indicate offsets from predicted pixels to their nearest ground-truth centerline; patches extracted at corresponding locations are compared with per-patch cross-entropy. Right: many-to-one mappings are handled by softmax-in-group weighting, which concentrates loss on the most confident pixel within each group.

ture parameter that controls the sharpness of the distribution. The weights are defined by a softmax over logits within each group:

$$w_{\mathbf{x}} = \frac{\exp\left(\frac{z(\mathbf{x})}{\tau}\right)}{\sum_{\mathbf{y}\in G(\mathbf{x})} \exp\left(\frac{z(\mathbf{y})}{\tau}\right)}. \tag{2}$$

This ensures that only the most confident prediction in each group receives high weight, reducing ambiguity when multiple pixels compete for the same ground-truth location.

3.2.2 Weighted Patch Alignment Loss

First, we define the patch alignment loss

$$\mathcal{L}_{\text{patch}} = \frac{\sum_{\mathbf{x}} m(\mathbf{x})\, w_{\mathbf{x}}\, \mathrm{CE}\big(\widehat{\mathbf{p}}(\mathbf{x}), \mathbf{p}(\mathbf{x}+\mathbf{v}(\mathbf{x}))\big)}{\big(\sum_{\mathbf{x}} m(\mathbf{x})\, w_{\mathbf{x}}\big)\cdot K}. \tag{3}$$

Here $CE(\cdot,\cdot)$ is the binary cross-entropy on logits, $m(\mathbf{x})$ is a binary mask indicating whether pixel $\mathbf{x}$ lies within distance d_{max} (a hyperparameter set to double the average road-width) of a ground-truth centerline, $w_{\mathbf{x}}$ is the reciprocal softmax weight (defined in eq 2), $\mathbf{v}(\mathbf{x})$ is the offset vector at pixel $\mathbf{x}$, $\widehat{\mathbf{p}}(\mathbf{x})$ is the set of logits in the predicted patch, $\mathbf{p}(\mathbf{x}+\mathbf{v}(\mathbf{x}))$ the binary labels of the corresponding ground-truth patch, and K the number of pixels in a patch. The loss averages binary cross-entropy errors across masked patches, with weights emphasizing high-confidence predictions and normalization ensuring comparability across patches.

3.2.3 False Positive Loss

The false-positive loss penalizes road predictions farther than d_{max} from any ground-truth centerline:

$$\mathcal{L}_{\text{fp}} = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\big(1-m(\mathbf{x})\big)\,\mathrm{CE}(z(\mathbf{x}),0), \tag{4}$$

where Ω is the image domain, $m(\mathbf{x}) \in \{0,1\}$ is the validity mask ($m(\mathbf{x})=1$ inside the tolerance band and 0 otherwise), and $z(\mathbf{x})$ is the logit at pixel $\mathbf{x}$. This loss suppresses predictions outside the valid road region.

3.2.4 Singleton Loss

The singleton loss $\mathcal{L}_{\text{singleton}}$ ensures that only one pixel in each group of predictions is activated. For each pixel $\mathbf{x}$, let $G(\mathbf{x})$ denote the set of predictions whose offset vectors map them to the same

ground-truth location. We then assign binary targets $t_{\mathbf{x}}$ so that exactly one pixel in each group receives $t_{\mathbf{x}} = 1$ (the one with the highest logit, ties broken by proximity to the ground truth) and all others receive $t_{\mathbf{x}} = 0$. These targets are derived deterministically and do not need to be differentiable. Then

$$\mathcal{L}_{\text{singleton}} = \frac{1}{\sum_{\mathbf{x}} m(\mathbf{x})} \sum_{\mathbf{x}} m(\mathbf{x}) \, \text{CE}(z(\mathbf{x}), t_{\mathbf{x}}), \tag{5}$$

which encourages the network to concentrate confidence on a single representative pixel per ground-truth location, reducing redundant predictions and focusing on a single, thin line prediction.

3.3 Intersection-Aware Loss

The intersection loss $\mathcal{L}_{\text{intersection}}$ enforces structural connectivity at road junctions, where prediction errors are especially costly. We detect intersections as road pixels that have more than two neighbors in a designated window. Let N_{int} denote the number of such ground-truth intersections, and let $\mathbf{p}(\mathbf{x}_i)$ be the ground-truth patch centered at intersection i. For each intersection, we search within a radius r for the predicted patch $\widehat{\mathbf{p}}(\mathbf{x}_j)$ that minimizes cross-entropy with the ground truth:

$$\mathcal{L}_{\text{intersection}} = \frac{1}{N_{\text{int}}} \sum_{i=1}^{N_{\text{int}}} \min_{\|\mathbf{x}_j - \mathbf{x}_i\| \le r} \text{CE}(\widehat{\mathbf{p}}(\mathbf{x}_j), \mathbf{p}(\mathbf{x}_i)) \, . \tag{6}$$

At junctions, multiple road branches meet, and the offset field becomes unstable, so nearest-neighbor assignment based on the vector field $\mathbf{v}(x)$ may select the wrong prediction. To avoid this, we treat all predicted points within a search radius as potential matches to an intersection, and supervise them jointly against the ground-truth patch, ignoring the vector field. This collective supervision maintains connectivity and yields more reliable learning at intersections.

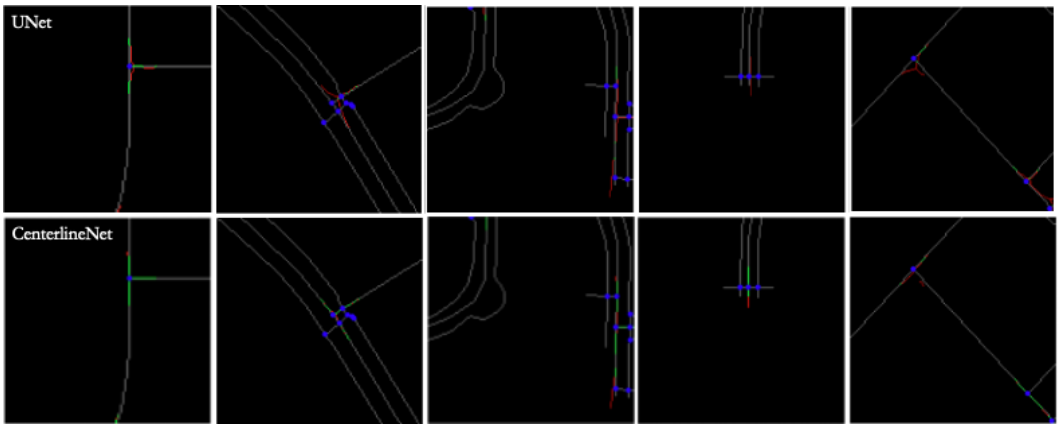

Figure 3: Intersection detection and targeted supervision. Intersections (blue dot) with U-Net+CE (top row) vs. CenterlineNet (bottom row). Improved intersection awareness is shown with green overlays (prediction aligned with ground truth) for CenterlineNet, versus red overlays (prediction outside tolerance) for the baseline.

4 Experimental Evaluation

We conduct comprehensive experiments to evaluate CenterlineNet's performance on road extraction and intersection topology preservation. Our evaluation addresses the fundamental challenges of assessing methods designed to handle spatial misalignment by utilizing specialized metrics that focus on topological accuracy rather than pixel-perfect correspondence.

4.1 Evaluation Methodology

The evaluation of road extraction methods poses fundamental challenges when spatial misalignment is present. Traditional pixel-wise metrics such as Intersection over Union (IoU), precision, and recall operate under the assumption of perfect spatial correspondence between predictions and ground truth annotations. This assumption directly contradicts our method's core premise of handling spatial uncertainties and misalignments that are inherent in real-world remote sensing data.

To address this contradiction, we develop a comprehensive evaluation framework that measures accuracy and structural preservation while accounting for reasonable spatial tolerances. Our approach recognizes that for road network extraction, the preservation of connectivity, intersection structure, and overall network structure is more critical than pixel-perfect alignment.

Our primary evaluation approach employs skeleton-based bipartite matching to assess road network quality independent of minor spatial offsets. Both predicted road networks and ground truth annotations undergo morphological thinning operations to produce nearly one-pixel-wide centerline representations. This preprocessing step removes large variations in road width prediction and focuses evaluation on the fundamental network structure and connectivity patterns.

We establish correspondences between predicted and ground truth skeleton pixels using the Hungarian algorithm for optimal bipartite matching. The matching process operates within varied distance threshold of pixels, which reflects realistic spatial tolerances for satellite imagery while accounting for typical registration errors, annotation inconsistencies, and the inherent uncertainties in manual road labeling. Table **??** highlights performance differences across method types.

Table 1: Quantitative Results (Centerline1M): Combined evaluation metrics on skeletonized predictions and ground truth centerlines. Combined Loss is combination of CE, Patch Alignment, Reciprocal, and Intersection Aware Loss.

Dataset	Models	Bipartite Matching (3-px tolerance)			Intersection Structure
		Precision (↑)	Recall (↑)	F1 (↑)	Structural IoU (↑)
Centerline1M	U-Net + CE Loss	0.276	**0.695**	0.395	0.108
	DeepLabV3 + CE Loss	0.329	<u>0.647</u>	0.436	0.111
	DeepLabV3 + Dice Loss	0.379	0.626	0.472	0.102
	CenterlineNet + CE Loss	0.612	0.463	0.527	0.114
	CenterlineNet + CE + Patch Alignment Loss	<u>0.822</u>	0.497	<u>0.619</u>	<u>0.151</u>
	CenterlineNet + CE + Patch Alignment + Reciprocal Loss	0.668	0.537	0.596	**0.157**
	CenterlineNet + Combined Loss	**0.841**	0.524	**0.646**	<u>0.151</u>

4.2 Intersection Structure Evaluation

Road intersections represent critical structural features that define network connectivity and determine practical utility for navigation applications. Unlike road segments, intersections present unique evaluation challenges due to their small spatial extent, complex geometric structure, and high importance for overall network functionality. We detect intersections in skeletonized road networks as pixels in the ground truth having more than two neighbors within their local morphological neighborhood. For evaluation, we employ a patch-based approach that directly compares the local road structure around intersection locations.

For each ground truth intersection, we extract a local pixel patch centered at the intersection location from the ground truth skeleton. We then find the spatially closest predicted intersection within a predetermined radius and extract a corresponding patch from the predicted skeleton. The matching is based purely on Euclidean distance between intersection centers, ensuring that the subsequent patch IoU evaluation provides an unbiased measure of structural similarity. If no predicted intersection exists within the specified radius, the ground truth intersection is considered unmatched for recall calculation.

The intersection quality is evaluated using the Intersection over Union (IoU) between these spatially matched patches, which captures both the spatial accuracy of the intersection location and the preservation of the local road network structure. This patch-based evaluation measures fundamentally different aspects of intersection quality compared to the bipartite matching approach used for road segments.

While bipartite matching focuses on whether individual skeleton pixels can be paired within a distance threshold, the patch-based IoU captures the local geometric shape and structure of intersections. It evaluates whether the predicted intersection preserves the correct angular relationships between incident road segments, maintains proper connectivity patterns, and reproduces the overall geometric configuration of the intersection. For example, a T-junction with specific arm angles and lengths will have a characteristic patch pattern that must be preserved to achieve high IoU scores.

The predetermined patch size is large enough to capture the full intersection structure including the immediate approach segments, while remaining focused on the local neighborhood. This approach naturally handles variations in intersection geometry, missing or extra road segments, and small spatial offsets while providing a direct measure of how well the predicted network preserves the local connectivity structure and geometric shape around intersections. The evaluation encompasses structural IoU (average IoU of matched intersection patches), results shown in Table **??**.

4.3 Experimental Setup and Datasets

All experiments were run on two NVIDIA TITAN RTX GPUs (24 GB each). We trained CenterlineNet on **Centerline1M**, our 1 m-resolution dataset of U.S. road centerlines automatically derived from USGS imagery (M2M API) and OpenStreetMap. Centerline1M intentionally retains the noisy, misaligned nature of crowd-sourced labels: OpenStreetMap vectors were rasterized into nearly one-pixel binary masks (anti-aliasing off), keeping only drivable road types (e.g., `highway=primary`, `secondary`, `residential`, `tertiary`) and ignoring service and non-drivable classes. No manual correction or alignment refinement was applied. The dataset comprises 10,845 tiles (8,579 training, 2,266 validation).

To test generalization, we evaluated on a well-known dataset. **SpaceNet Roads**[1] offers multi-city, multi-resolution satellite imagery together with road vector annotations (centerline graphs), providing a complementary test of geographic and sensor transfer (Van Etten et al., 2018). We rasterized and skeletonized these vector road networks in order to align with our own evaluation pipeline.

Additionally, we tested against **RoadTracer**, which combines aerial imagery with OpenStreetMap road graphs and evaluates performance using a graph-based junction metric (Bastani et al., 2018).

Table 2: Quantitative Results (SpaceNet and RoadTracer): Bipartite evaluation metrics on SpaceNet and RoadTracer dataset with CoANet and CenterlineNet.

Dataset	Models	Bipartite Matching (3-px tolerance)		
		Precision (↑)	Recall (↑)	F1 (↑)
SpaceNet	CoANet	0.521	**0.501**	0.492
	CenterlineNet	**0.546**	0.465	**0.624**
RoadTracer	CoANet	0.189	0.195	0.167
	CenterlineNet	**0.603**	**0.419**	**0.744**

We had also planned to evaluate on **DeepGlobe Road Extraction** (8,570 RGB images of size 1024×1024, with per-pixel road vs. background masks at 50 cm resolution) (Demir et al., 2018), but due to lack of availability of pretrained CoANet weights, we opted to use SpaceNet and RoadTracer instead.

4.4 Occlusions

A critical challenge in road centerline extraction from satellite imagery is the presence of occlusions caused by trees, buildings, shadows, or cloud cover, which obscure parts of the road network. These occlusions introduce gaps in visual continuity and spatial misalignment between predictions and ground truth. To improve resilience under these conditions, we augment our training data with synthetic occlusions in addition to naturally occluded scenes.

For each training tile we randomly select a road pixel from the ground-truth mask (or a random location if no road pixel exists) and center a rectangular occlusion over that point. The occlusion

[1] `https://spacenet.ai/datasets/`

size is sampled between 0.5× and 1.0× of a base fraction of the image dimensions (0.25 in our experiments), producing variable occlusion shapes. Within this rectangle we replace the original pixel values with the mean color of the entire image rather than black or random noise. The same occlusion rectangle is applied to all input channels, and we also generate a binary occlusion mask to record the affected area. Ground-truth road masks themselves are left unaltered so that the model still receives supervision at occluded locations.

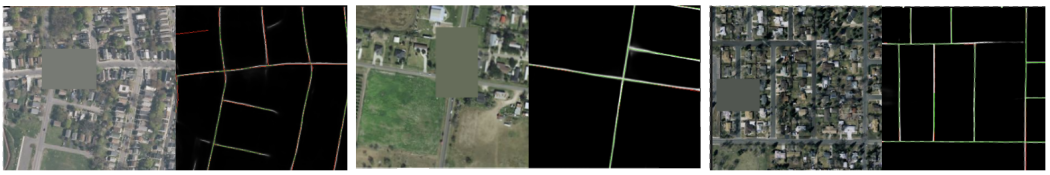

Figure 4: Examples of CenterlineNet predictions on synthetically occluded satellite imagery. For each scene, the left panel shows the original satellite image with an artificial occlusion (gray rectangle), and the right panel shows the corresponding CenterlineNet prediction overlaid on a black background. Green pixels indicate predicted road centerlines correctly matching ground truth within the spatial tolerance, while red pixels mark mismatches. These examples illustrate CenterlineNet's ability to maintain road connectivity and infer centerlines across missing or obscured regions.

This augmentation forces CenterlineNet to infer road connectivity across missing visual segments, reduces false negatives in occluded regions, and improves generalization to deployment scenarios where occlusion is the norm rather than the exception.

4.5 Results and Analysis

CenterlineNet achieves competitive performance compared to baseline methods across overall road extraction metrics, with improved Precision, F1, and Structural IoU metric scores. Not only quantitatively we can see improvements qualitatively see Fig. 5. In UNet and DeepLab predictions we can see not structurally correct prediction branches, even though DeepLab with Dice loss seems to have success over its predecessors CenterlineNet still shows optimal improvements over the later.

We conduct an ablation studies to understand the contribution of each loss component in our approach. We evaluate a baseline configuration using standard binary cross-entropy loss with pixel-wise alignment assumptions, then systematically add the patch alignment loss while maintaining standard assumptions for other components. We further examine the contribution of softmax-in-group weighting to handle many-to-one mapping scenarios, and finally integrate specialized supervision for intersection detection and preservation.

As shown in Table 1, the inclusion of different components leads to varied improvements across Precision and F1 scores. Notably, the patch alignment loss substantially boosts Precision and F1, while the reciprocal formulation improves Recall. The addition of intersection supervision further enhances precision and F1, demonstrating the complementary benefits of various loss combinations. Despite its improvements, CenterlineNet sometimes struggles with occluded roads, very thin or low-contrast rural roads, and complex highway interchanges, see supplementary materials.

5 Conclusion

We presented CenterlineNet, a weakly supervised approach for road centerline extraction that addresses spatial misalignment in remote sensing applications. Our patch alignment loss provides spatial tolerance while maintaining topological accuracy, suitable for real-world scenarios where perfect annotation alignment cannot be guaranteed. Our contributions include: (1) a patch alignment loss using vector fields to establish flexible correspondences, (2) a reciprocal formulation handling many-to-one mappings through softmax-in-group weighting, and (3) an intersection-aware component improving network connectivity. Results demonstrate competitive performance with improved robustness to spatial noise and annotation inconsistencies. The approach enables practical deployment where spatial uncertainties are inherent. Future work will explore extension to other linear infrastructure extraction tasks and integration with vector post-processing methods.

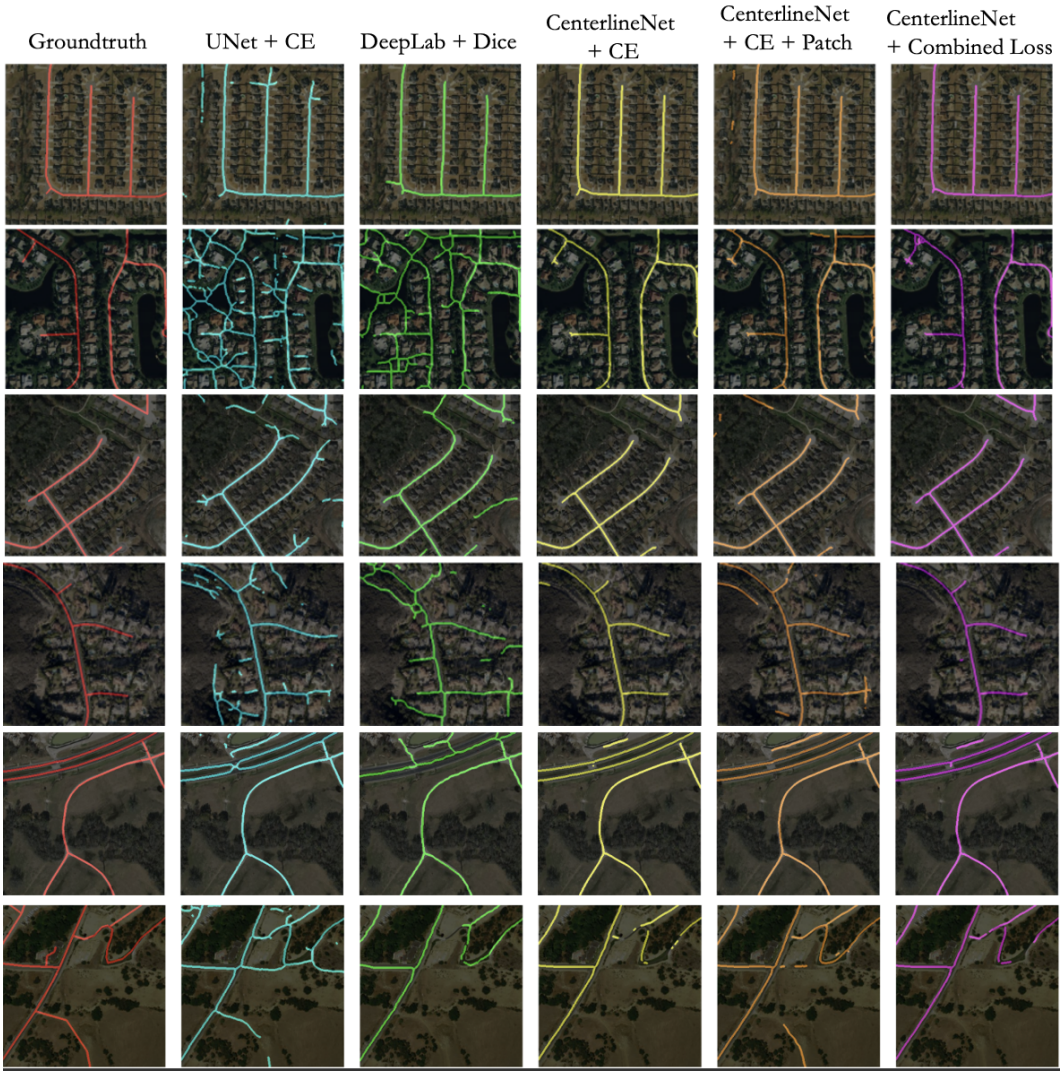

Figure 5: Qualitative Results: Aerial images overlayed with thickened groundtruth and various model thickened predictions masks showing improved structure accuracy with CenterlineNet.

ACKNOWLEDGMENTS

AI tools, ChatGPT and GitHub Copilot, were used during polishing writing and researching similar model publications for prior art. The authors directed the work, made all research decisions, and carried out the analysis with substantial human effort.

REFERENCES

D. Acuna, A. Kar, and S. Fidler. Self-training with noisy labels for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6952–6961, 2019.

Fahad Bastani, Songtao He, and Mohammad Alizadeh. Roadtracer: Automatic extraction of road networks from aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, C. V. Jawahar, and Manohar Paluri. Improved road connectivity by joint learning of orientation and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision (ECCV), 2018.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.

Hao He, Dongfang Yang, Shicheng Wang, Shuyang Wang, and Yongfei Li. Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. Remote Sensing, 11(9):1015, 2019.

Jiuxiang Hu, Anshuman Razdan, John C. Femiani, Ming Cui, and Peter Wonka. Road network extraction and intersection detection from aerial images by tracking road footprints. In IEEE International Conference on Image Processing (ICIP), pp. IV–9–IV–12, 2007. doi: 10.1109/ICIP.2007.4379742.

Yannick Kirchhoff, Maximilian R. Rokuss, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Philipp Vollmuth, Jens Kleesiek, Fabian Isensee, and Klaus Maier-Hein. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In European Conference on Computer Vision (ECCV), 2024. to appear.

Yu Liu, Wen Zhang, Xiaoming Wang, et al. Coanet: Connectivity-aware network for road extraction from remote sensing imagery. pp. 1234–1243, 2022. doi: 10.1109/CVPRW12345.2022.00123.

Xiaomei Lu, Yanfei Zhong, Zhuo Zheng, Dong Chen, Yu Su, Ailong Ma, and Liangpei Zhang. MRENet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images. Remote Sensing, 13(2):239, 2021.

J B. Mena. State of the art on automatic road extraction for GIS update: a novel classification. Pattern Recognition Letters, 24(16):3037–3058, 2003.

Agata Mosinska, Pablo Marquez-Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3136–3145, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Y. Sun, H. Liu, S. Bambach, and W. Liu. Leveraging crowdsourced data for road extraction from aerial imagery via deep learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 750–759, 2019.

Adam Van Etten, David Lindenbaum, Todd Bacastow, et al. Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232, 2018. Open Data on AWS, includes building and road extraction tasks.

Yulong Wei, Kefei Zhang, and Shunping Ji. Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing. IEEE Transactions on Geoscience and Remote Sensing, 58(12):8919–8931, 2020.

Wei Yuan and Wenbo Xu. GapLoss: A loss function for semantic segmentation of roads in remote sensing images. Remote Sensing, 14(10):2422, 2022.

C. Zhang, W. Li, D. Tuia, and Y. Wang. Iterative alignment of openstreetmap road annotations to aerial imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 162:155–167, 2020. doi: 10.1016/j.isprsjprs.2020.02.005.

Li Zhang, Hao Chen, Xin Liu, et al. Msmdff-net: Multi-scale multi-directional feature fusion network for road extraction. Remote Sensing, 15(4):987–1002, 2023. doi: 10.3390/rs15040987.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters, 15:749–753, 2018. URL https://api.semanticscholar.org/CorpusID:206437632.

Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018.

A Appendix