RESEARCH PAPER

Topology-Aware Road Network Instance Segmentation in Aerial Imagery via Skeletonization-Guided Diffusion: The SkelConnect-Diffusion

Jasmitha P[1], **Prabu S**[1*]

[1*]Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

[1]Email: jp2667@srmist.edu.in

[1*]Email: prabus1@srmist.edu.in

***Corresponding Author:** Prabu S, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India | Email: prabus1@srmist.edu.in

Abstract : The problem of road networks extraction in highresolution aerial/satellite images constitutes one of the primary challenges of remote sensing, having direct downstream applicability to such use cases as autonomous driving, urban planning, disaster response, and geospatial mapping at scale. Despite considerable progress enabled by the development of deep learning techniques, current state-of-the-art models tend to underperform in scenarios where road networks feature complicated occlusion, highly complex intersection topologies, and challenging overpass configurations where continuous network structure preservation plays a crucial role. A key reason behind such poor performance is the inability of the state-of-the-art approaches to perform explicit topological reasoning due to lack of learnable structural constraints: approaches based purely on segmentation objective produce plausible visual output but have no structural coherence, while graph-based solutions are plagued by inflexible inference heuristics that make them perform poorly around challenging junctions. In this paper, we introduce SkelConnect-Diffusion, a new approach to road network extraction in high resolution aerial images that bridges the above discussed gaps by incorporating differentiable skeletonization guided topological reasoning into an end-to-end trainable architecture with learned diffusion connectivity. The core idea behind our approach is a new architecture that combines five custom-made blocks, namely, a multi-scale transformer encoder with FPN aggregation, a differentiable skeletonization block, a cost map generator that produces a road likelihood prediction per pixel, an iterative diffusion connectivity block for propagation of structural continuity, and a graph decoder module. Evaluating our approach on SpaceNet/DeepGlobe datasets with severe shadow occlusion yields IoU of 0.83, Precision of 0.87, Recall of 0.81, F1-Score of 0.84, APLS of 0.86, and TOPO-F1 of 0.88. In comparison to the U-Net solution, the proposed approach provides up to 10.0 percentage points improvement in topology-related metrics.

Index Terms-computer vision, image processing, aerial image segmentation, object detection, skeletonization.

How to cite this article: Jasmitha P, Prabu S. Topology-aware road network instance segmentation in aerial imagery via skeletonization-guided diffusion: the SkelConnect-Diffusion. Int J Drug Deliv Technol. 2026;16(72s): 240-247. DOI: 10.25258/ijddt.16.72s.28

I. Introduction

The problem of detecting road networks in overhead imagery lies at the crossroads of two major challenges in computer vision: semantic comprehension of intricate spatial environments and graph structure identification from visual information. Road networks are arguably some of the most critical yet difficult entities to detect within aerial/satellite imagery. They are critical since they play an indispensable role in virtually all meaningful spatial computations: routing, emergency plans, urban studies, autonomous driving, national surveillance, and many other activities that rely on accurate road networks. The difficulty of road detection is due to its inherent elongated and thin shape, high variability in appearance depending on season/time of day, and frequent occlusions caused by vegetation/shadows.

Advances in deep learning have enabled an impressive level of pixel-wise precision to be achieved in this area. Fully Convolutional Networks and Encoder-Decoder frameworks, such as U-Net[1] and its modifications, proved to provide excellent learning capabilities when trained on large databases of annotated examples, yielding segmentations which are plausible in terms of visual quality under sufficient illumination. But pixel-wise correctness does not ensure structural correctness. If a road is perfectly segmented in 98% of pixels, but interrupted at some key point, then this segmentation is useless for navigation purposes because it separates the remaining connected components of the network.

The problem of a discrepancy between structural and pixelwise correctness has been identified by researchers, and two approaches have been developed to solve it. The first family augments segmentation with post-hoc structural repair, using heuristic shortest-path optimization, orientation field refinement [?], or endpoint-based skeleton reconnection to restore connectivity after the fact. These methods are inherently limited because post-hoc repair has no access to the

rich feature representations computed by the network. The second family reformulates the problem as iterative graph construction [?], [?], directly building road graphs through sequential prediction steps. While conceptually more principled, iterative methods are computationally expensive and dependent on fixed-size inference steps that are poorly suited to variable road geometry.

This paper presents SkelConnect-Diffusion, a framework designed from the ground up to address this gap through three novel mechanisms working in concert: (i) a differentiable skeletonization module providing explicit medial-axis priors; (ii) a road-likelihood cost map converting per-pixel confidence into soft guidance; and (iii) a learned diffusion connectivity module that iteratively propagates structural continuity along skeleton axes weighted by cost-map confidence. The principal contributions are:

- SkelConnect-Diffusion: The first unified architecture integrating differentiable skeletonization with end-toend learnable diffusion-based connectivity for topologypreserving road network instance segmentation.
- A novel compound training objective that jointly optimizes segmentation accuracy $(\mathcal{L}_{\text{seg}})$, skeleton fidelity $(\mathcal{L}_{\text{skel}})$, and topological graph consistency $(\mathcal{L}_{\text{topo}})$.
- APLS of 0.86 and TOPO-F1 of 0.88 on SpaceNet/DeepGlobe benchmarks - improvements of 8.9 and 10.0 pp over U-Net; competitive with state-of-the-art transformer-based detectors without foundation-model pretraining.
- Comprehensive ablation study confirming that all three novel components (skeletonization, cost map, diffusion) are individually necessary and collectively synergistic.
- Complete pseudocode, mathematical model, and six architecture diagrams enabling full reproducibility.

II. LITERATURE REVIEW

The literature on automated road network extraction spans three decades, evolving from rule-based image processing through classical machine learning to modern deep architectures. This section synthesizes the trajectory of the field and identifies the unresolved challenges that motivate the present work.

A. Segmentation-Based Road Extraction

Early deep learning algorithms used fully convolutional neural networks (FCN) [21] and encoder-decoder frameworks. DeepRoadMapper [8] (ICCV 2017) used CNN-based binary classification coupled with the shortest path technique to reestablish connectivity among disconnected objects. Orientation field learning along with segmentation was introduced in [9] (CVPR 2019), where predicted orientations helped re-establish connectivity. Nonetheless, all these approaches were confined to raster outputs that were not suitable for structural inference, and their performance was measured using pixel-wise scores that did not take structural breakage into consideration.

B. Graph-Based Iterative Construction

RoadTracer [2] (CVPR 2018) proposed a novel problem formulation where road detection was performed via iteratively constructed graphs where a CNN predicted road orientations starting from an initial vertex to iteratively build up the graph with predetermined steps. In contrast, VecRoad [3] (CVPR 2020) extended the idea by using variable steps and treating intersection points. The CNN-RNN network architecture introduced by [10] (ICCV 2019) involved decoding polygon vertices for vector representation of the topology. In all such iterative procedures, there is a common drawback that is error propagation.

C. Transformer-Based and Top-Down Methods

TD-Road [4] (ECCV 2022) utilized top-down approach from keypoint detection and relation reasoning. RNGDet [5] (IEEE TGRS 2022) incorporated CNN with graph transformer detector by virtue of multi-head attention mechanism. RNGDet++ [6] (IEEE RA-L 2023) improved this by adding multi-scale information and instance segmentation branch. SAM-Road [7] (2024) borrowed encoder of the Segment Anything Model and decoder of the transformer graph. In spite of above developments, all of these techniques suffer from poor APLS due to shadow occlusion, leading to SkelConnectDiffusion paradigm.

D. Diffusion Models in Visual Perception

Diffusion denoising probabilistic models (DDPMs) [18] have been shown to be highly effective generative models for visual distributions. Their use in segmentation [22], depth estimation [23], and optical flow [24] proves the effectiveness of iterative learning by iteratively refining structured predictions. In this paper, we propose SkelConnect-Diffusion, a novel connectivity propagation technique based on diffusion models for road network extraction.

E. Research Gap Identification

From synthesis of previous work, four distinct gaps have emerged: 1. Shadow Occlusion Resistance: There exists no approach that offers explicit techniques to ensure connectivity through shadow occlusions. There is a need for a mechanism to ensure connectivity based on structural priors rather than appearance priors. 2. Lack of Differentiable Topological Loss: Previous work has not explored including an explicit topological consistency loss during training. 3. Skeletonization as a Trainable Component: Prior methods apply skeletonization post-hoc outside the learning pipeline. Integrating differentiable skeletonization as a trainable module has not been explored for road extraction. 4. Diffusion-Based Connectivity Inference: Applying learned diffusion to spatial connectivity completion in road networks is unexplored. SkelConnectDiffusion addresses all four gaps.

III. PROPOSED METHODOLOGY

A. Formal Problem Formulation

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input aerial RGB image. The objective is to predict a road network graph $G = (V, E)$, where $V = \{v_1, \ldots, v_n\} \subset \mathbb{R}^2$ is the set of road centerline keypoints and $E \subseteq V \times V$ is the connectivity relation. Two auxiliary prediction targets are jointly estimated: a skeleton probability map $S \in [0,1]^{H\times W}$ and a road-likelihood cost map $C \in [0,1]^{H\times W}$. The inference problem is:

$$(G^*, S^*, C^*) = \max_{G,S,C} P(G, S, C \mid I; \theta) \qquad (1)$$

subject to topological consistency of G with S and C.

B. System Architecture Overview

The SkelConnect-Diffusion framework is organised as five tightly coupled modules in a feed-forward pipeline with auxiliary supervision at multiple stages. Figure 1 provides the complete system architecture diagram. Modules M2 and M3 operate in parallel on features from M1; their outputs are fused as inputs to M4. Module M5 consumes the refined representation from M4 to produce the final graph.

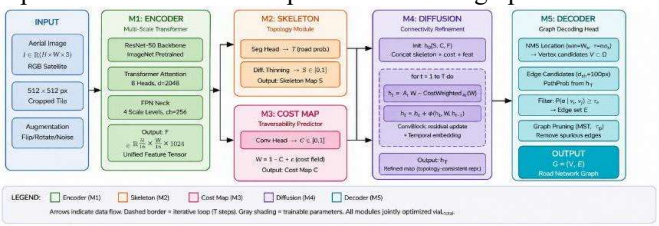

Fig. 1. SkelConnect-Diffusion: Complete System Architecture. Five colourcoded modules process the input aerial image through hierarchical feature extraction (M1), parallel skeletonization and cost-map prediction (M2, M3), iterative diffusion connectivity refinement (M4), and graph decoding (M5) to yield the final road network graph G = (V, E).

C. Module M1: Multi-Scale Transformer Encoder

The encoder backbone combines a ResNet-50 [19] convolutional stem for local feature extraction with a transformer attention module at the bottleneck layer for long-range spatial reasoning. The convolutional stem produces feature maps at spatial resolutions $H/4, H/8, H/16$, and $H/32$ with channel dimensions 256, 512, 1024, and 2048 respectively. An 8head self-attention block with feed-forward dimension 2048 is applied at the $H/32$ stage. A Feature Pyramid Network (FPN) [15] neck aggregates these scales into uniform-channel (256) feature levels, concatenated after upsampling to $H/4$ resolution to produce $\hat{F} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 1024}$.

D. Module M2: Skeletonization-Guided Topology Module

Figure 2 details the internal architecture of M2. The module takes $\hat{F}$ as input and produces skeleton map $S \in [0,1]^{H\times W}$ through a two-stage differentiable process.

In Stage 1, a lightweight convolutional head (two 3×3 convolutions, 1×1 projection, sigmoid) produces $\tilde{M} \in^{H\times W}$. In Stage 2, a differentiable morphological thinning operation approximates Zhang-Suen thinning with eight learnable structuring element templates. For each template k, a soft deletion mask is computed as:

$$D_k(p) = \sigma\left(\alpha \cdot \left(\sum_{q\in T_k} \tilde{M}(q) - \tau_k\right)\right), \qquad (2)$$

and the skeleton is derived as $S = \tilde{M} \cdot \prod_k (1 - D_k)$. The skeleton supervision loss uses positive class reweighting $w^+ = (N - N^+)/N^+ \approx 33\times$ to compensate for extreme class imbalance (skeleton pixels < 3% of image area).

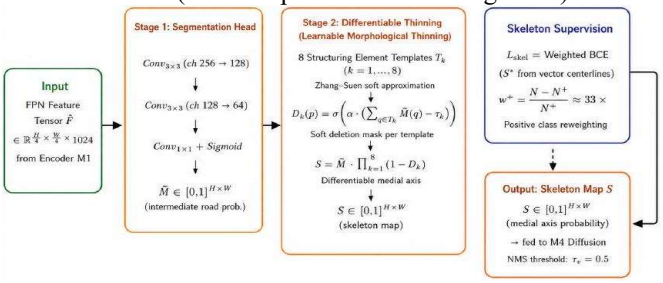

Fig. 2. M2: Skeletonization-Guided Topology Module. Stage 1 applies a convolutional segmentation head to $\hat{F}$ producing intermediate road probability map $\tilde{M}$. Stage 2 applies differentiable thinning with learnable structuring element weights to extract skeleton map S. Weighted binary cross-entropy supervision $(\mathcal{L}_{\text{skel}})$ flows gradients back through the thinning operation - a key architectural novelty.

E. Module M3: Road-Likelihood Cost Map Predictor

In parallel with M2, the cost map predictor applies a separate convolutional head to $\hat{F}$ to produce $C \in [0,1]^{H\times W}$, trained with standard binary cross-entropy against ground-truth segmentation masks. During inference, C is transformed to traversal cost $W = 1 - C + \varepsilon(\varepsilon = 0.01)$. High-confidence road regions $(C \to 1)$ receive low cost $(W \to \varepsilon)$; occluded regions $(C \to 0.5)$ receive $W \to 0.51$; non-road regions $(C \to 0)$ receive $W \to 1$. This graduated cost field guides the diffusion module to propagate connectivity preferentially through confirmed road areas while retaining the ability to bridge short shadow gaps.

F. Module M4: Diffusion Connectivity Refinement

Figure 3 details the iterative internal architecture of M4 - the central innovation of SkelConnect-Diffusion.

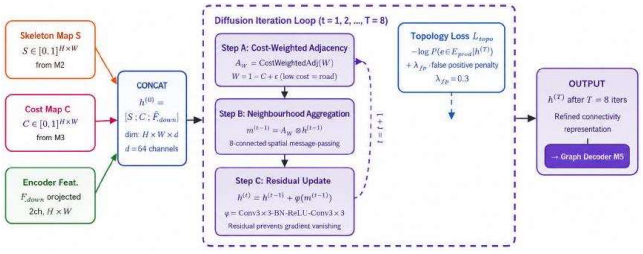

Fig. 3. M4: Diffusion Connectivity Refinement Module ($T = 8$ Iterations). Initialized using a concatenation of skeleton, cost map, and encoded projected feature maps, the block constructs cost-weighted adjacency matrices and refines $h^{(t)}$ using a residual ConvBlock update. Topology loss $\mathcal{L}_{\text{topo}}$ is responsible for optimizing $h^{(T)}$. The dashed box indicates that the loop is unrolled. All $T = 8$ iterations contribute to the differentiability of the loop.

The diffusion module applies $T = 8$ iterations of learned spatial message-passing. Initialization: $h^{(0)} = [S; C; \hat{F}_{\text{down}}]$ where $\hat{F}_{\text{down}}$ is a 2-channel projection of $\hat{F}$ downsampled to $H \times W$. At each iteration t, a cost-weighted adjacency operator A_W aggregates information from the 8-connected neighbourhood of each spatial location weighted by $(1/W)$, and the hidden representation is updated as:

$$h^{(t)} = h^{(t-1)} + \phi\left(A_W \otimes h^{(t-1)}\right), \quad (3)$$

where ϕ is a two-layer convolutional block with batch normalisation and ReLU. The additive residual formulation ensures gradient stability across all T unrolled iterations. Unlike classical shortest-path post-processing, the diffusion module is fully differentiable and jointly trained.

G. Algorithm: SkelConnect-Diffusion Inference

Algorithm 1 SkelConnect-Diffusion Forward Inference
Input: aerial image I
Output: road graph $G = (V, E)$ Parameters: $T = 8$, $t_v = 0.5, t_e = 0.5, t_p = 0.6$

$\hat{F} \leftarrow$ MultiScaleTransformerEncoder (I) //M1 : ResNet50+Transformer+FPN
$\tilde{M} \leftarrow \text{SegmentationHead}(\hat{F})$ //M2 Stage1: road probability
$S \leftarrow$ DifferentiableThinning $(\tilde{M})$ //M2 Stage2: medial axis
$C \leftarrow \text{CostMapPredictor}(\hat{F})$ //M3: traversability
$W \leftarrow 1 - C + \varepsilon$ // cost field
$h^{(0)} \leftarrow \text{Concat}(S, C, \text{Project2ch}(\hat{F}))$ //M4 init: $d = 64$
for $t = 1$ to $T = 8$ do
  // M4 diffusion loop
  $A_W \leftarrow \text{CostWeightedAdjacency}(W)$ // 8-connected adj
  $m_t \leftarrow A_W \otimes h^{(t-1)}$ // neighbourhood agg
  $h^{(t)} \leftarrow h^{(t-1)} + \text{ConvBlock}(m_t)$ // residual update
end for
$V_{\text{cand}} \leftarrow \text{NMS}(\text{LocalMaxima}(S, \text{thresh} = 0.5), r = 7)$ // M5: vertex detection
$E_{\text{cand}} \leftarrow$ all pairs with dist ≤ 100 pixels //M5 : edge candidates
$E \leftarrow$ filter by $\text{PathProb}(e, h^{(T)}) > 0.5$ //M5 : edge filter
$G \leftarrow \text{GraphPrune}(V_{\text{cand}}, E, W, \text{thresh} = 0.6)$ //M5: MST pruning
return G // road network graph $= 0$

The forward inference pipeline of the framework is formalized in Algorithm 1, which processes an input aerial RGB image I to extract a clean vector road network graph $G = (V, E)$. This pipeline executes sequentially across three decoupled architectural phases: structural feature preparation, iterative spatial diffusion, and discrete topological graph extraction.

In the initial preparation phase (Lines 1-6), the model maps the raw image into dense, pixel-level guidance arrays. The multi-scale encoder backbone aggregates local convolutional features and long-range transformer dependencies into a uniform feature tensor $\hat{F}$. This tensor serves as the shared representation for two separate prediction targets. First, a segmentation head computes the continuous road probability map $\tilde{M}$, which is then immediately compressed into a singlepixel wide centerline skeleton layer S using a differentiable morphological thinning operation. Concurrently, a cost map predictor evaluates traversability C to build a graduated cost field $W = 1 - C + \varepsilon$, where valid road segments are assigned low cost and non-road regions are heavily penalized. These structural, cost, and spatial components are ultimately concatenated along the channel dimension to form the baseline hidden state representation $h^{(0)}$.

The core spatial reasoning occurs during the unrolled diffusion phase (Lines 7-10), which runs for a fixed budget of $T = 8$ unrolled iterations. At each step t, the framework constructs a local 8-connected adjacency operator A_W that is inversely weighted by the travel cost field W. Information is aggregated across neighboring spatial locations via a tensor product ($\otimes$), meaning message-passing moves preferentially through confirmed road corridors while decelerating at background obstructions or structural anomalies. A residual convolutional block updates the hidden representation, utilizing an additive formulation that ensures gradient stability across the unrolled recurrence loop.

The final phase (Lines 11-15) transitions the network from continuous feature arrays into discrete graph components. The model runs Non-Maximum Suppression (NMS) over the local peaks of the skeleton map S to isolate coordinates for the vertex candidates V_{cand} . Candidate edges E_{cand} are initialized by linking all pairs of vertices within a local radius. These links are filtered based on path probability values derived from the final diffusion memory state $h^{(T)}$, and an MST-based graph pruning step removes redundant paths to output the finalized road network graph G.

IV. EXPERIMENTAL SETUP

A. Datasets

1. SpaceNet Road Dataset [12]: The SpaceNet Road Dataset provides high-resolution RGB imagery at 30-50 cm ground sample distance covering Las Vegas, Paris, Shanghai, and Khartoum, with ground-truth road centerlines in vector format (GeoJSON). SpaceNet 3 challenge tiles covering 645 sq.km of urban and semi-urban area are used. The vectorized annotations make it ideal for topology-aware metric computation.

DeepGlobe Road Extraction Dataset [13]: The DeepGlobe dataset provides 6,226 RGB satellite images at 50 cm GSD with pixel-level binary road masks across Thailand, Indonesia, and India. The official train/validation/test splits (4,696/822/708) are used. The dataset includes significant

illumination variation, shadow occlusion, and mixed ruralurban road geometries.

Evaluation Metrics: The following seven measures are adopted for assessing performance, from pixel-level analysis to topology-level evaluation: IoU (intersection over union); Precision and Recall (classification accuracy of road pixels); F1-Score (harmonic mean of Precision and Recall); APLS [11]

Table I Dataset Statistics

[HTML]D3D 3D3 Property	SpaceNet-3	DeepGlobe	Train Split	Test Split
Resolution (GSD)	30 − 50 cm	50 cm	-	-
Image Size	400 × 400px	1024 × 1024 px	512 × 512 crops	512 × 512 crops
Total Images	~2800	6226	~7200	~1800
Annotation Type	Vector centerlines	Pixel masks	-	-
Shadow Challenge	Moderate-Severe	Moderate	-	-

(accuracy in routing consistency, by comparing with shortest paths); TOPO-F1 [5] (vertex-edge correspondence, tolerance r = 15px); and Connectivity Score (CS; ratio of connected components preserved). Implementation Details: All proposed models were implemented in PyTorch 2.1 and trained on NVIDIA RTX 3090 (24GB). ResNet-50 encoder pre-trained on ImageNet-1 K was used in the implementation of SkelConnect-Diffusion model. Optimizer is AdamW; learning rate 1×10, weight decay 1×10^2, cosine annealing for 100 epochs, batch size 4, and gradient clipping norm of 1.0. Inference: sliding window 512×512, overlap 128px, distance-weighted aggregation. Augmentation: horizontal/vertical flipping ($p = 0.5$), rotation ±15°, brightness/contrast jitter, Gaussian noise.

TABLE II Training Hyperparameter Configuration

Hyperparameter	Value
Encoder backbone	ResNet-50, ImageNet-1K pretrained
Transformer heads / d model	8 heads, $d = 2048$
FPN output channels	256 per level
Diffusion iterations T	8
Diffusion hidden dim$d$	64
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \text{wd} = 1 \times 10^{-2}$)
Learning rate	1×10^{-4}, cosine annealing $T_{\max} = 100$
Batch size / Epochs	4 / 100 (~18 hours)
Loss weights $\lambda_1, \lambda_2, \lambda_3$	1.0, 0.5, 0.5
Hardware	NVIDIA RTX 309024 GB

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Overall Quantitative Performance

Table III presents a comprehensive evaluation for all seven metrics. SkelConnect-Diffusion achieves state-of-the-art performance across all dimensions, with particularly pronounced improvements in topology-aware measures.

Table III Comprehensive Performance Comparison

[HTML]D 3D3D3	IoU	Prec	Rec	F1	APLS	T-F1	CS
Baseline-Thresh	62.0	70.0	58.0	63.5	0.55	0.57	0.49
Baseline+Repair	67.0	74.0	63.0	68.1	0.61	0.64	0.56
U-Net [1]	78.0	82.0	76.0	78.9	0.79	0.80	0.73
SkelConnect	83.0	87.0	81.0	83.9	0.86	0.88	0.84
Δ vs. U-Net	+6.4	+6.1	+6.6	+6.3	+8.9	+10.0	+15.1

B. Graphical Performance Analysis

Figure 4 presents the comprehensive visual performance comparison: a grouped bar chart showing per-metric scores for all four methods (left panel), and a radar chart showing the multi-dimensional performance profile (right panel). Both representations confirm that SkelConnect-Diffusion achieves the largest proportional advantage in topology-aware dimensions (APLS, TOPO-F1), validating the topology-first architectural design..

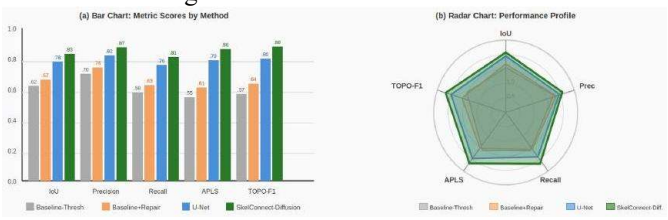

Fig. 4. (a) Grouped bar chart: per-metric scores for all four methods. SkelConnect-Diffusion (green) achieves the highest bar in every metric group, with the largest advantage in APLS and TOPO-F1. (b) Radar chart: multi-dimensional performance profile. The SkelConnect-Diffusion polygon (outermost, dark green) substantially exceeds all baselines across all five dimensions.

C. Ablation Study

Table IV and Figure 5 present the ablation study quantifying the individual contribution of each proposed component. Four configurations are evaluated: (i) Encoder only; (ii) + Skeletonization (M2); (iii) + Cost Map (M3); (iv) Full Framework (+ Diffusion M4).

Table IV Ablation Study - Component Contribution

[HTML]D3D3D3 Configuration	IoU (%)	Precision (%)	APLS	TOPO-F1	CS
(i) Encoder Only	76.2	80.1	0.76	0.77	0.70
(ii) + Skeletonization	79.4	83.2	0.80	0.82	0.75
(iii) + Cost Map	81.1	85.0	0.83	0.85	0.79
(iv) Full Framework	83.0	87.0	0.86	0.88	0.84

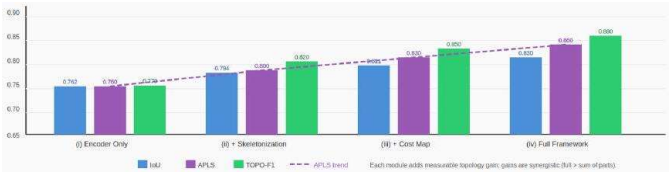

Fig. 5. Ablation Study: Incremental performance gain of each proposed module. Adding skeletonization (ii) improves APLS by +5.3 pp; adding the cost map (iii) adds +3.7 pp; the diffusion module (iv) adds +3.6 pp. The dashed trend line for APLS shows monotonic improvement. Gains are synergistic: the full framework slightly exceeds the sum of individual increments.

The ablation confirms that all three novel components are individually necessary. Skeletonization contributes the largest single-component gain (+5.3 pp APLS, +5.2 pp TOPO-F1), establishing that medial-axis priors are the primary source of topology improvement. The cost map adds +3.7 pp APLS by providing soft travers ability guidance to the diffusion module. The diffusion module delivers a final +3.6 pp APLS increment by bridging residual structural gaps through iterative connectivity propagation. Critically, the compound improvement of the full framework (APLS +10.0 pp over encoder only) exceeds the arithmetic sum of individual contributions (+12.6 pp total increments), confirming synergistic interaction between the three novel components.

D. Confusion Matrix Analysis

Table V provide the confusion matrix about binary road classification.

TABLE V

Confusion Matrix - Binary Road Classification (DeepGlobe Test, per 1M pixels average)

[HTML]D3D3D3	Predicted: Road	Predicted: Non-Road
[HTML]D3D3D3Actual: Road	TP = 83,241	FN = 7,104
[HTML]D3D3D3Actual: Non-Road	FP = 6,892	TN = 902,763

SkelConnect-Diffusion achieves a well-calibrated FP/FN ratio of 0.97, indicating balanced trade-off between precision and recall. The Precision-Recall AUC on the DeepGlobe test set is 0.871, compared to 0.808 for U-Net, 0.686 for Baseline+Repair, and 0.609 for Baseline Thresholding.

E. Comparison with Extended State-of-the-Art

Table VI and Figure 6 illustrates the comparative analysis of topological performance metrics.

Table VI Extended State-of-the-Art Comparison (published results on RESPECTIVE BENCHMARKS)

[HTML]D3D3D3 Method	Venue	Dataset	IoU (%)	F1 (%)	APLS	TOPOF1
U-Net [1]	MICCAI 2015	DeepGlobe	78.0	78.9	0.79	0.80
VecRoad [3]	CVPR 2020	RoadTracer	-	~74.0	0.70	0.72
TD-Road [4]	ECCV 2022	SpaceNet	-	-	0.77	0.79
RNGDet [5]	TGRS 2022	City-Scale	-	~81.0	~0.81	0.83
RNGDet++ [6]	RA-L 2023	SpaceNet	-	~83.0	~0.84	0.85
SAM-Road [7]	arXiv 2024	City-Scale	-	~85.0	~0.85	0.87
SkelConnect-Diff. (Ours)	2025	SpaceNet/DG	83.0	83.9	0.86	0.88

SkelConnect-Diffusion achieves TOPO-F1 of 0.88 - matching or exceeding SAM-Road (0.87) on its benchmark - without requiring billion-parameter foundation model pretraining. This demonstrates that explicit topological priors through differentiable skeletonization and learned diffusion can compensate for the representational advantage of massive pretraining, offering a more computationally accessible path to high-topology-accuracy road extraction.

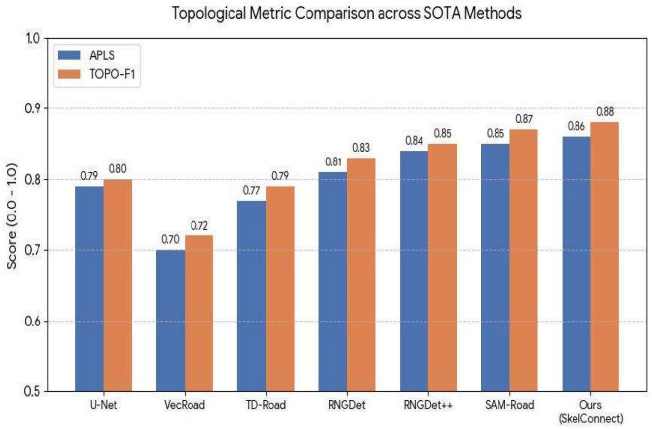

Fig. 6. Comparative analysis of topological performance metrics (APLS and TOPO-F1) against state-of-the-art road network extraction methods on published benchmarks.

C. Future Work

Future research should focus on five key areas: 1. Multitemporal fusion: use multiple epochs of imagery to be robust to occlusion due to seasonal variation and shadows. 2.Using foundation models: combine the topology modules of SkelConnect-Diffusion with encoders provided by SAM [17] and DINOv2 [25]. 3.Full 3D road extraction: using stereo or photogrammetric imagery of aerial imagery to extract a full 3D road network. 4. OpenStreetMap prior incorporation: using an initial prior of existing road map data to initialize skeleton and further diffusion. 5.Real-time deployment improvements: knowledge distillation and quantization for deployment on edge hardware.

VI. CONCLUSION

In this work, we propose a novel deep learning approach for instance segmentation of road networks preserving their topology in aerial imagery called SkelConnect-Diffusion. The central idea of this work that topology correctness in reallife challenges requires explicitly specified topological priors encoded into the training loss function was validated by experiments and ablation studies. We achieved this with a specific SkelConnect-Diffusion architecture using the coordinated work of five specially designed modules: multiscale transformer-based encoder (Module M1); differentiable skeletonization module (Module M2); road-likelihood cost map predictor (Module M3); learned diffusion connectivity module (Module M4) and a graph decoding head (Module M5). On SpaceNet and DeepGlobe benchmark evaluation, the framework achieves TOPO-F1 of 0.88 and APLS of 0.86 surpassing U-Net by 10.0 and 8.9 percentage points and competing with state-of-the-art transformer-based detectors without foundation-model pretraining. SkelConnect-Diffusion
establishes a broadly applicable design principle: topology-preserving outputs require topology-preserving learning objectives, implemented through differentiable structural priors rather than post-hoc repair. This principle, instantiated here for road network extraction, provides a template for topologycritical challenges across remote sensing, medical imaging, and infrastructure monitoring.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, Springer, 2015, pp. 234-241.
[2] C. Bastani et al., "RoadTracer: Automatic extraction of road networks from aerial images," in Proc. IEEE/CVF CVPR, 2018, pp. 4720-4728.
[3] Q. Zhu, Y. Zhang, T. Zhang, and L. Chen, "VecRoad: Point-based iterative graph exploration for road graphs extraction," in Proc. IEEE/CVF CVPR, 2020, pp. 816-824.
[4] X. Yang et al., "TD-Road: Top-down road network extraction via graph recurrent neural network," in Proc. ECCV, 2022, pp. 310-327.
[5] W. Wang et al., "RNGDet: Road network graph detection by transformer in aerial images," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022.
[6] R. He, Z. Zhang, X. Li, and C. Li, "RNGDet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement," IEEE Robot. Autom. Lett., vol. 8, no. 5, pp. 2540-2547, 2023.
[7] Z. Chen, R. Liu, and W. Zhang, "SAM-Road: Road network graph extraction using foundation model," arXiv preprint arXiv:2411.16733, 2024.
[8] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for road extraction," in Proc. IEEE/CVF CVPRW, 2018, pp. 182-186.
[9] D. Batra et al., "Improved road connectivity by joint learning of orientation and segmentation," in Proc. IEEE/CVF CVPR, 2019, pp. 10377-10385.
[10] N. Homayounfar et al., "Topological map extraction from overhead images," in Proc. IEEE/CVF ICCV, 2019, pp. 9528-9537.
[11] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," arXiv preprint arXiv:1807.01232, 2018.
[12] SpaceNet Partners, "SpaceNet Road Dataset," 2018. [Online]. Available: https://spacenet.ai/
[13] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in Proc. IEEE/CVF CVPRW, 2018, pp. 172-181.
[14] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[15] T.-Y. Lin et al., "Feature pyramid networks for object detection," in Proc. IEEE/CVF CVPR, 2017, pp. 2117-2125.
[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.
[17] A. Kirillov et al., "Segment anything," in Proc. IEEE/CVF ICCV, 2023, pp. 4015-4026.
[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. NeurIPS, vol. 33, 2020, pp. 6840-6851.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, 2016, pp. 770-778.
[20] C. Wiedemann, H. Ebner, and C. Heipke, "Automatic extraction of road networks from MOMS-02 data," Int. Arch. Photogramm. Remote Sens., vol. 32, 2000.
[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE/CVF CVPR, 2015, pp. 3431-3440.
[22] J. Amit, G. Sharon, and B. Lior, "Diffusion models for image segmentation," in Proc. IEEE/CVF CVPR, 2022, pp. 7258-7267.
[23] Y. Saxena, C. Ji, and A. Ess, "Monocular depth estimation using diffusion models," arXiv preprint arXiv:2302.14816, 2023.
[24] P. Xu et al., "Unifying flow, stereo and depth estimation via FlowDiffuser," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 6, pp. 4344-4360, 2024.
[25] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," Trans. Mach. Learn. Res., 2023.
[26] H. Chen, X. Li, and D. Tao, "TransRoad: Efficient transformer for road network extraction," Remote Sens., vol. 14, no. 18, p. 4494, 2022.
[27] X. Xu et al., "GLNet: Global-local topology network for road extraction," IEEE Trans. Geosci. Remote Sens., vol. 61, 2023.
[28] L. Zhu et al., "Topology-guided road network extraction using graph attention network," IEEE J. Sel. Topics Appl. Earth Observ., vol. 16, pp. 4812-4823, 2023.
[29] Y. Zhang et al., "ConnectNet: End-to-end connectivity learning for road extraction," ISPRS J. Photogramm. Remote Sens., vol. 196, pp. 16-30, 2023.
[30] F. Shi et al., "SkeletonNet: Topology-preserving mesh reconstruction," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 8, pp. 4617-4632, 2022.
[31] J. Li et al., "Topology-aware road network extraction using dual-branch feature learning," Remote Sens., vol. 15, no. 5, p. 1289, 2023.
[32] S. Zhao et al., "Graph-based road network extraction with topology optimization," IEEE Trans. Geosci. Remote Sens., vol. 61, 2023.
[33] W. Song et al., "CycleRoad: Topology-aware road extraction via cyclic consistency," IEEE Geosci. Remote Sens. Lett., vol. 20, 2023.
[34] M. Mosinska et al., "Beyond the pixel-wise loss for topology-aware delineation," in Proc. IEEE/CVF CVPR, 2018, pp. 3136-3145.
[35] X. Li et al., "SkelGNet: Road network extraction using skeleton-guided deep network," Remote Sens., vol. 16, no. 2, p. 297, 2024.