VERA-ROAD: VISUAL EVIDENCE REASONING AND ADAPTIVE REPAIR FOR ROAD NETWORK EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Road-network extraction from very-high-resolution remote sensing imagery is difficult not only because roads vary across scenes, but also because different regions within the same image fail for different reasons. A narrow road may require stronger geometric continuity, a low-contrast segment may benefit from fine-detail enhancement, and a visually ambiguous junction may require a wider spatial context. Most extraction pipelines nevertheless process all regions through a largely fixed computational path. We introduce VERA-Road, a region-wise visual decision framework that replaces this static processing pattern with an *observe–act–verify–repair* loop. A shared vision-language policy first interprets local, neighboring, and global evidence and issues an action for each region. The proposed **Contextual Action Composer (CAC)** realizes the action by combining complementary detail, geometry, and cross-band operators. Rather than assuming that the issued action is always sufficient, **Outcome-Guided Repair (OGR)** measures the resulting prediction state and selectively applies directional-alignment or multi-scale-context repair only to unresolved regions. Finally, **Topology-Rewarded Policy Learning (TRPL)** optimizes the action policy from downstream graph quality with a group-relative objective, avoiding the need for region-level action annotations. On the SpaceNet road benchmark, the resulting model reaches 86.79 F1, 93.47 precision, 82.51 recall, and 87.42 APLS, while the ablation and sensitivity studies show complementary gains from action composition, selective repair, and topology-aware policy optimization.

1 INTRODUCTION

Road networks are among the most structurally constrained objects in remote sensing imagery. Their appearance may be locally weak, partially occluded, spectrally similar to surrounding surfaces, or fragmented by shadows and buildings, yet the final prediction is expected to preserve long-range connectivity. These requirements make road extraction fundamentally different from ordinary pixel-wise segmentation: a locally plausible mask can still produce a topologically invalid network. Accurate automatic mapping is therefore important for transportation analysis, urban planning, navigation, and disaster response, while remaining challenging at very-high spatial resolution.

Early deep models largely addressed the problem through dense segmentation and post-processing, from patch-based road detection to encoder–decoder architectures, city-scale graph recovery, and connectivity refinement (Mnih & Hinton, 2010; Mattyus et al., 2017; Zhou et al., 2018; Van Etten, 2019; Batra et al., 2019; Mei et al., 2021; Liu et al., 2024). More recent graph-oriented methods predict vertices and edges directly or combine dense road likelihoods with explicit topology reasoning (Bastani et al., 2018; Tan et al., 2020; Li et al., 2019; He et al., 2020; Bahl et al., 2022; Xu et al., 2022; 2023; Yang et al., 2023). Foundation-model-based approaches further improve representation transfer. In particular, SAM and SAM2 have motivated road-specific graph extraction systems that reuse large-scale visual pretraining while adding lightweight topology heads (Kirillov et al., 2023; Ravi et al., 2024; Hetang et al., 2024; Yin et al., 2025). These advances improve the feature backbone and graph decoder, but the feature-processing path itself is still mostly predetermined once the network is built.

This static design is restrictive because road regions do not fail for the same reason. Consider three common cases. First, a thin or low-contrast segment may be visible only through fine local variations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Second, an elongated road crossing clutter may require stronger directional and structural cues. Third, a region whose RGB appearance is ambiguous may be better resolved by complementary cross-band evidence. The appropriate operation can also change after an intermediate prediction is produced: a transformation that seems reasonable from the input may not actually improve road continuity. Thus, the central question is not only *what features should be extracted*, but also *how each region should be processed, and whether that processing has worked*.

We formulate this question as region-wise visual decision making. Given an initial road prediction and its visual context, a controller selects a processing action for each region. The action is executed on the feature map, the resulting prediction is inspected, and only regions that remain unresolved are sent to a repair stage. This leads to a closed-loop pipeline in which the model can revise an earlier decision rather than committing to a single pass. We instantiate the controller with Qwen2.5-VL-7B-Instruct (Bai et al., 2025), while keeping the visual encoder and readout compatible with the original road-graph pipeline.

Based on this formulation, we propose VERA-Road. Its first component, Contextual Action Composer (CAC), converts region-level decisions into combinations of three complementary operators that emphasize detail, geometry, and cross-band evidence. Its second component, Outcome-Guided Repair (OGR), verifies the effect of the issued action using updated uncertainty and graph measurements, and selectively invokes either directional-alignment repair or multi-scale-context repair. Its third component, Topology-Rewarded Policy Learning (TRPL), learns the controller from final graph outcomes. Since road annotations do not specify which action should be selected in each region, we optimize the policy using TOPO-F1/APLS-based reward and a group-relative objective derived from GRPO (Shao et al., 2024). Only low-rank adaptation parameters (Hu et al., 2022) of the policy are updated during policy learning.

Our contributions are threefold. **(1)** We recast adaptive road representation as a region-wise visual action problem and introduce CAC to compose complementary feature operators according to scene evidence. **(2)** We introduce an outcome-driven verify-and-repair mechanism that evaluates whether an issued action actually resolves the target region before allocating additional computation. **(3)** We develop topology-rewarded policy learning that connects regional processing decisions to final road-network quality without requiring explicit action labels.

2 Related Work

2.1 Road-network extraction from remote sensing imagery

Road extraction has evolved from dense semantic segmentation toward explicit connectivity modeling and vectorized graph prediction. Early segmentation systems learned road appearance directly from aerial imagery (Mnih & Hinton, 2010), while DeepRoadMapper, D-LinkNet, orientation-aware refinement, and CoANet progressively introduced topology recovery, enlarged receptive fields, directional supervision, and connectivity attention (Mattyus et al., 2017; Zhou et al., 2018; Batra et al., 2019; Mei et al., 2021). Graph-growing approaches such as RoadTracer and VecRoad construct road networks sequentially (Bastani et al., 2018; Tan et al., 2020), whereas PolyMapper, Sat2Graph, Single-Shot Road Graph Extraction, and TopDiG represent roads more directly as geometric or topological structures (Li et al., 2019; He et al., 2020; Bahl et al., 2022; Yang et al., 2023). RNGDet and RNGDet++ further formulate adjacent-vertex prediction with transformers and multi-scale features (Xu et al., 2022; 2023). Recent methods continue to improve the efficiency–topology trade-off or convert dense predictions into vector graphs, including DeH4R, Seg2Vector, and MaGRoad (Gong & Ji, 2026; Wang et al., 2026; Guan et al., 2026). Our work is complementary to these graph decoders: we focus on how intermediate visual representations are adaptively processed before the final graph is produced.

2.2 Foundation models for road mapping

Large pretrained vision models provide transferable representations that can reduce task-specific architectural complexity. Vision Transformer established patch-token processing as a scalable visual backbone (Dosovitskiy et al., 2021), while SAM and SAM2 extended large-scale pretraining to promptable segmentation (Kirillov et al., 2023; Ravi et al., 2024). SAM-Road demonstrates that such embeddings can support efficient road-graph extraction (Hetang et al., 2024), and subsequent work

extends this direction to broader geographic coverage and more robust graph construction (Yin et al., 2025). VERA-Road uses a frozen visual backbone and a frozen road readout, but adds a separate vision-language controller that reasons over visual context and intermediate prediction states. This separates *representation execution* from *action selection*.

2.3 Learning decision policies from task outcomes

A difficulty in region-adaptive processing is the absence of direct action supervision: road datasets provide masks or graphs, not labels telling a model which feature transformation should be applied to a given region. We therefore treat action selection as a policy-learning problem. Group Relative Policy Optimization (GRPO) replaces a learned value baseline with group-normalized rewards and has been used to optimize language-model reasoning policies (Shao et al., 2024). We adapt this idea to road extraction by defining rewards from changes in graph-level metrics and by penalizing unnecessary actions. The resulting objective learns processing decisions from their downstream consequences rather than from hand-crafted action annotations.

3 Method

3.1 Road extraction as region-wise visual decision making

We consider paired observations (I, M), where $I \in \mathbb{R}^{H\times W\times C_I}$ is an RGB image and $M \in \mathbb{R}^{C_M\times H\times W}$ is its spatially aligned multispectral observation. The target is a road graph $G = (V, E)$ with vertices V in the image domain and edges $E \subseteq V \times V$. Training uses a reference graph $G^\star = (V^\star, E^\star)$.

Let $p \in \mathcal{P}$ index overlapping patches. A frozen vision encoder $\mathcal{E}$ maps an RGB patch I_p to a feature tensor

$$F_{0,p} = \mathcal{E}(I_p) \in \mathbb{R}^{C_f\times H_f\times W_f}. \tag{1}$$

A frozen road readout $\mathcal{D}$ decodes the patch features $\mathbf{F}_0 = \{F_{0,p}\}_{p\in\mathcal{P}}$ into a road-probability map, a keypoint map, and an initial graph,

$$(P_0^{\text{road}}, P_0^{\text{key}}, G_0) = \mathcal{D}(\mathbf{F}_0). \tag{2}$$

The aligned multispectral patches M_p provide auxiliary cross-band evidence.

The central design of VERA-Road is to avoid applying the same feature operation everywhere. From the initial prediction state $S_0 = (\mathbf{F}_0, P_0^{\text{road}}, P_0^{\text{key}}, G_0)$, we form candidate regions using prediction uncertainty and graph structure. Each region is represented by a mask $R_r \in [0, 1]^{H\times W}$ and a context package containing local, neighboring, and global imagery, the current prediction maps, graph visualization, and numerical measurements. A shared vision-language policy $\mathcal{Q}_\theta$ maps these observations to discrete region-level actions. CAC executes the first action, OGR verifies the resulting state and optionally issues a repair action, and TRPL trains the same policy from the final graph outcome. The overall framework is illustrated in Figure 1.

3.2 Contextual Action Composer

Complementary visual operators. CAC defines an action space $\mathcal{B}_A = \{D, G, C\}$ containing three operators with different processing behavior. The *detail operator* extracts Haar high-frequency responses from $F_{0,p}$ to emphasize fine road boundaries and local discontinuities. The *geometry operator* applies directional and dilated convolutions to strengthen elongated structures and spatial continuity. The *cross-band operator* combines normalized multispectral features with inter-band contrasts and maps them to the visual feature grid. Each operator produces a residual correction $\Delta F_{k,p}$ with the same shape as $F_{0,p}$.

Region observation and action prediction. For each region r, visual context $\mathcal{C}_r^0$ contains local, neighboring, and global views together with road/keypoint predictions and the current graph. Image measurements $\mathbf{e}_r$ summarize frequency, spatial, and cross-band evidence, while prediction measurements $\mathbf{d}_r^0$ summarize uncertainty and road-graph structure. The action-policy observation is

$$\mathcal{O}_A = \{(\mathcal{C}_r^0, \mathbf{e}_r, \mathbf{d}_r^0)\}_r. \tag{3}$$

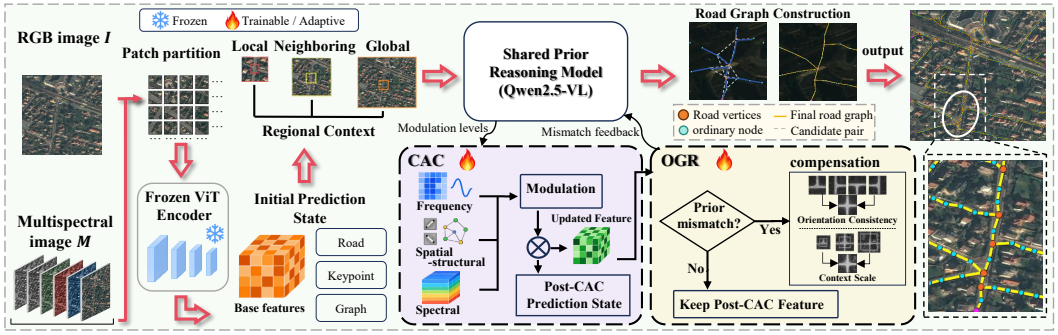


Figure 1: Overview of the proposed VERA-Road framework.

Given an action prompt p_A, the policy predicts a vector of discrete action levels,

$$\mathbf{a}_r = [\mathcal{Q}_\theta(\mathcal{O}_A; p_A)]_r \in \{0, \dots, L_A\}^{|\mathcal{B}_A|}. \tag{4}$$

We convert the levels into continuous coefficients $\alpha_{r,k} = a_{r,k}/L_A$. Multiple operators may be active simultaneously, which allows the controller to express mixed processing strategies rather than forcing a single hard choice.

Spatial action composition. Regions may overlap, so naively summing their action strengths can amplify corrections. We therefore construct a normalized spatial gate for operator k,

$$A_k(x) = \frac{\sum_r R_r(x)\alpha_{r,k}}{\max\{1, \sum_r R_r(x)\}}, \qquad x \in \Omega. \tag{5}$$

Cropping and resizing A_k to the patch feature grid gives $A_{k,p}$. The action-composed feature is then

$$F_{A,p} = F_{0,p} + \sum_{k \in \mathcal{B}_A} A_{k,p} \odot \Delta F_{k,p}, \tag{6}$$

where $\odot$ denotes element-wise multiplication with channel broadcasting. When all action coefficients are zero, the corresponding feature remains unchanged. Decoding $\mathbf{F}_A$ yields the post-action state S_A.

3.3 Outcome-Guided Repair

An issued action is not automatically assumed to be correct. OGR explicitly evaluates its effect and allocates additional computation only to regions that remain unresolved.

Outcome verification. For each region, we recompute prediction measurements from S_A and augment them with the action magnitude and the road-probability change relative to S_0. After min–max normalization with bounds fixed on the training calibration split, the valid measurements are averaged into an unresolved score

$$U_r = \frac{\sum_j v_{r,j}\, \bar{d}^A_{r,j}}{\sum_j v_{r,j}}, \qquad \sum_j v_{r,j} > 0, \tag{7}$$

where $\bar{d}^A_{r,j}$ is a normalized measurement and $v_{r,j} \in \{0, 1\}$ marks validity. Regions with $U_r < \tau_v$ retain the CAC output directly. Regions with $U_r \geq \tau_v$ are considered unresolved and are forwarded to the repair selector. This verification step converts the pipeline from one-shot feature processing into a closed-loop decision process.

Repair selection. For unresolved regions, we refresh the visual context using the post-action road maps and graph, obtaining $\mathcal{C}^A_r$. The updated observation $\mathcal{O}_R$ is queried with repair prompt p_R, producing discrete requirement levels

$$\mathbf{m}_r = [\mathcal{Q}_\theta(\mathcal{O}_R; p_R)]_r\,. \tag{8}$$

The entries are partitioned into a directional-alignment group $\mathcal{I}_{\mathrm{dir}}$ and a context-scale group $\mathcal{I}_{\mathrm{sc}}$. Their normalized responses are

$$s_r^{(q)} = \frac{1}{L_R|\mathcal{I}_q|} \sum_{j \in \mathcal{I}_q} m_{r,j}, \quad q \in \{\mathrm{dir}, \mathrm{sc}\}. \tag{9}$$

A fixed resolver chooses the higher-scoring repair only when its score, score margin, and support count satisfy calibration thresholds; otherwise no second-stage operation is executed.

Directional and contextual repair. The *Directional Alignment Repair* re-encodes rotated versions of a patch, inversely aligns their feature grids, and averages the aligned encodings. It is designed for regions whose elongated road evidence is unstable under the original orientation. The *Multi-Scale Context Repair* re-encodes patch-centered crops at multiple spatial scales and averages their aligned features, providing a wider context for regions that cannot be resolved from the original crop alone. Both repairs reuse the frozen encoder $\mathcal{E}$. Let $F_{q,p}$ denote the feature returned by repair q and $B_{q,p}$ the spatial mask of regions assigned to it. The repaired feature is

$$F_{R,p} = F_{A,p} + \sum_{q \in \{\mathrm{dir},\mathrm{sc}\}} B_{q,p} \odot (F_{q,p} - F_{A,p}). \tag{10}$$

Unselected locations remain unchanged. The final maps and road graph are obtained from $\mathcal{D}(\mathbf{F}_R)$.

3.4 Topology-Rewarded Policy Learning

Road annotations supervise the desired geometry and connectivity, but they do not state which regional processing action should be chosen. TRPL therefore learns the action policy from the consequences of its decisions.

Operator training and policy initialization. We first train the three CAC operator branches with map and edge supervision plus residual-diversity regularization, then freeze them. Candidate CAC assignments are executed and ranked using a weighted combination of TOPO-F1 and APLS. High-performing assignments provide action targets for supervised initialization of the policy. Repair targets are constructed from the measured benefit of the two repair operators, with positive repair supervision retained only when graph quality improves over the post-CAC state by a specified margin. The shared policy is initialized using supervised fine-tuning with LoRA (Hu et al., 2022).

Graph-level reward. After initialization, we optimize the policy with final extraction feedback. For each training image, a group $\mathcal{G}$ contains multiple recorded decision sequences. A sequence includes the CAC action and, when triggered, the OGR repair decision. Its reward is

$$R_i = \sum_j \lambda_j \frac{\Delta q_{ij}}{s_j} - \lambda_C C_i, \tag{11}$$

where $q_j \in \{\mathrm{TOPO\text{-}F1}, \mathrm{APLS}\}$, $\Delta q_{ij} = q_j(G_i, G^\star) - q_j(G_0, G^\star)$, and $s_j > 0$ is the calibration-set standard deviation of the corresponding metric gain. C_i penalizes the number and cost of unique active operations, including the second policy query when repair is triggered. The cost term discourages the controller from improving reward merely by activating every available operation.

Group-relative optimization. The group mean $\mu_{\mathcal{G}}$ and standard deviation $\sigma_{\mathcal{G}}$ define the relative advantage

$$\widehat{A}_i = \frac{R_i - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \epsilon}. \tag{12}$$

We assign the same sequence-level advantage to the recorded policy responses and optimize a clipped GRPO-style objective

$$\mathcal{L}_{\mathrm{TRPL}} = \mathbb{E}_{i,s}\left[\frac{1}{|y_{is}|}\sum_t \left(-\min\left(\rho_{ist}\widehat{A}_i, \mathrm{clip}(\rho_{ist}, 1-\epsilon_c, 1+\epsilon_c)\widehat{A}_i\right) + \beta\kappa_{ist}\right)\right], \tag{13}$$

where ρ_{ist} is the likelihood ratio between the current and previous policies, ϵ_c is the clipping range, and κ_{ist} regularizes deviation from the reference policy. Only the LoRA parameters of $\mathcal{Q}_\theta$ are updated in this stage.

Table 1: Comparison with representative road extraction methods on SpaceNet. Values are percentages.

Method	F1 ↑	Precision ↑	Recall ↑	APLS ↑
RNGDet (Xu et al., 2022)	82.62	91.54	75.63	67.49
MaGRoad (Guan et al., 2026)	84.37	87.64	80.31	72.39
DeH4R (Gong & Ji, 2026)	85.11	86.74	81.98	72.43
Seg2Vector (SAM) (Wang et al., 2026)	83.26	90.64	77.98	73.42
Seg2Vector (SAM2) (Wang et al., 2026)	84.21	91.38	78.67	74.56
VERA-Road (Ours)	**86.79**	**93.47**	**82.51**	**87.42**

4 EXPERIMENTS

4.1 DATASET, METRICS, AND IMPLEMENTATION

Dataset. We evaluate on the SpaceNet road extraction benchmark (Van Etten et al., 2018). The supplied experimental split contains 2,780 images of 1300×1300 pixels at 0.3 m/pixel, with 2,213 images for training and 567 for testing. The imagery spans Las Vegas, Paris, Shanghai, and Khartoum, providing substantial variation in urban layout and road morphology.

Metrics. We report pixel/road extraction F1, precision, and recall, together with Average Path Length Similarity (APLS), the routing-oriented metric introduced for the SpaceNet road challenge (Van Etten et al., 2018), which compares shortest-path structure between predicted and reference road networks. TOPO-F1 is additionally used inside the policy reward because it directly reflects graph-level structural quality.

Implementation. VERA-Road is implemented in PyTorch and trained on two NVIDIA GeForce RTX 5090 GPUs. AdamW is used with learning rates 1×10^{-4} for CAC, 1×10^{-4} for supervised policy initialization, and 5×10^{-6} for TRPL refinement. The three stages are trained for 50, 30, and 30 epochs, respectively. The base parameters of Qwen2.5-VL-7B-Instruct and the OGR repair operators are frozen; only the CAC parameters and the LoRA parameters of the policy model are updated.

4.2 COMPARISON WITH REPRESENTATIVE METHODS

Table 1 compares VERA-Road with representative road-graph extraction baselines on SpaceNet. We use the published method names RNGDet and MaGRoad. For Seg2Vector, we distinguish the SAM and SAM2 configurations supported by the released implementation (Wang et al., 2026; Kirillov et al., 2023; Ravi et al., 2024). VERA-Road achieves the highest value in all four reported metrics. Relative to the strongest baseline value in each column, it improves F1 by 1.68 percentage points, precision by 1.93 points, recall by 0.53 points, and APLS by 12.86 points. The particularly large APLS margin indicates that the gain is not limited to pixel-level recognition: the extracted graph better preserves route-level connectivity while simultaneously improving precision and recall.

Figure 2 shows qualitative examples from three scenes. Across the illustrated urban and sparse-road cases, the final prediction retains long connected segments and produces clean road geometry. These visualizations are consistent with the large APLS gain in Table 1: the improvement is most visible when fragmented local evidence must be converted into a coherent network.

4.3 ABLATION STUDY

We isolate the contribution of the three stages in Table 2. Adding CAC to the frozen base raises F1 from 80.73 to 83.58 and APLS from 82.56 to 84.57, showing that region-dependent action composition is useful even without second-stage repair. Adding OGR further improves all reported metrics, with APLS reaching 85.18. Finally, TRPL raises the full model to 86.79 F1, 93.47 precision, and 87.42 APLS. The stage-wise trend supports the intended closed-loop interpretation: action

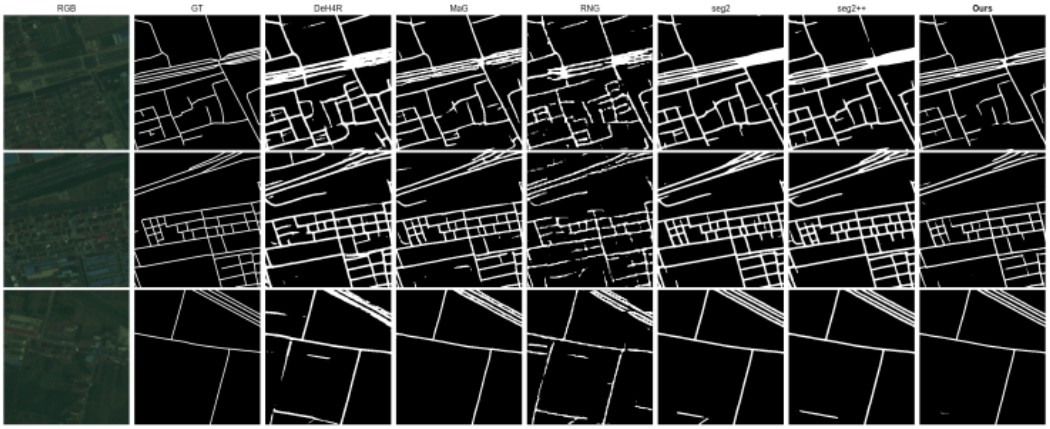


Figure 2: Qualitative comparison on SpaceNet. The examples illustrate how the final model preserves long road segments and junction structure while reducing spurious road responses.

Table 2: Stage-wise ablation of VERA-Road on SpaceNet.

Variant	CAC	OGR	TRPL	F1 ↑	Prec. ↑	APLS ↑
Base				80.73	90.42	82.56
+ CAC	✓			83.58	92.76	84.57
+ OGR	✓	✓		84.02	93.41	85.18
Full VERA-Road	✓	✓	✓	**86.79**	**93.47**	**87.42**

composition provides the initial representation gain, repair handles difficult residual regions, and topology-rewarded learning further improves how the policy allocates these operations.

4.4 When should repair be triggered?

The verification threshold τ_v controls how conservatively regions are routed to OGR. Figure 3 reveals a consistent trade-off between repair frequency and graph quality: permissive triggering sends many regions to the second stage, whereas a more selective setting reduces unnecessary repair and yields stronger connectivity. The trend indicates that OGR is most effective when it acts as a targeted fallback for unresolved regions rather than as another globally applied processing block.

4.5 Computation-performance trade-off

Figure 4 compares dense TFLOPs per image with extraction quality as the pipeline is progressively enabled. The base model requires 17.82 dense TFLOPs/image. Adding CAC increases the cost to 90.37 TFLOPs/image and provides the largest first-stage accuracy gain. Enabling OGR increases dense computation only slightly to 91.62 TFLOPs/image because repair is applied selectively. The full model reaches 132.08 TFLOPs/image after policy-related processing and achieves the strongest reported extraction quality. Table 1 reports the corresponding recall result. The parameter count is 8.39B, dominated by the frozen vision-language policy; only a small subset of parameters is updated through LoRA during TRPL.

5 Discussion and Limitations

The experiments suggest that the main benefit of VERA-Road comes from treating feature processing as a decision that can be revised. This perspective differs from simply stacking more feature branches: the controller chooses how strongly each operator participates, and OGR explicitly checks whether the resulting state warrants additional computation. The sensitivity study further shows that this verification step matters because unconditional repair can reduce graph quality.

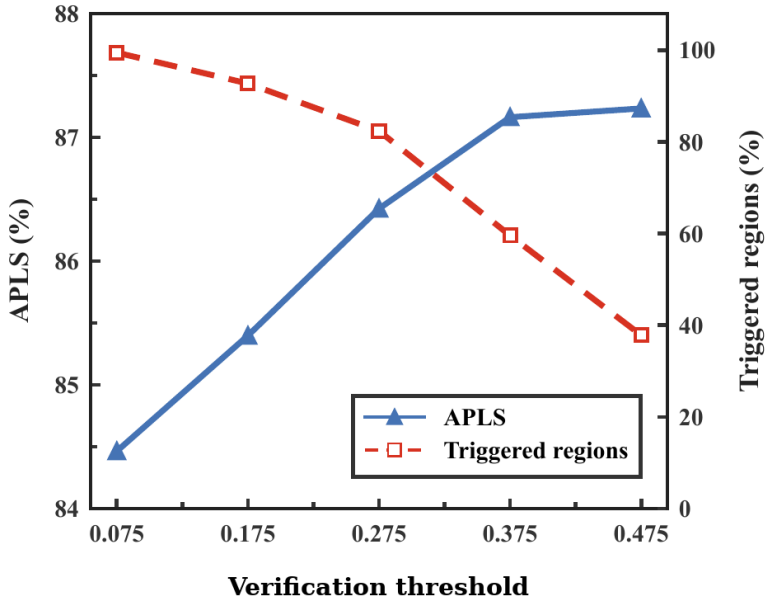


Figure 3: Sensitivity to the outcome-verification threshold, illustrating the trade-off between repair frequency and graph quality.

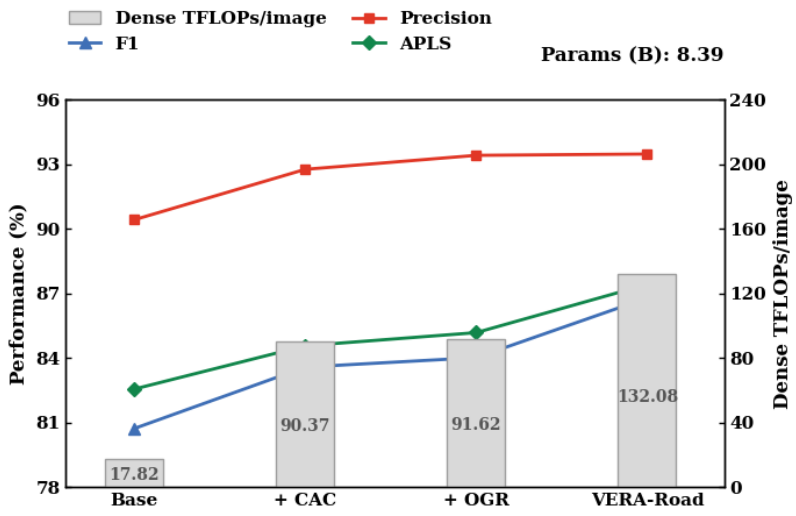


Figure 4: Computation-performance trade-off across progressively enabled stages of VERA-Road.

The current evidence also has clear limitations. First, the supplied experiments are concentrated on SpaceNet, so the present results do not by themselves establish cross-dataset robustness. Second, the vision-language controller contributes substantial model capacity even though most of its parameters remain frozen. Third, the action and repair spaces are manually defined; the policy learns when to use them but does not discover entirely new operators. Finally, the reward depends on graph metrics computed after extraction, which increases training-time evaluation cost. These limitations motivate future work on smaller controllers, broader benchmarks, and more flexible action spaces.

6 CONCLUSION

We presented VERA-Road, a road-network extraction framework built around region-wise visual decisions. The method replaces a fixed feature-processing path with an observe–act–verify–repair loop. CAC converts visual evidence into adaptive combinations of detail, geometry, and cross-band operators; OGR verifies the action outcome and selectively applies directional or multi-scale repair; and TRPL learns the shared policy from downstream topology feedback. On the supplied SpaceNet experiments, the complete model improves both pixel-level extraction and graph connectivity, with the strongest gain observed in APLS. The results support outcome-aware regional processing as a useful direction for adaptive road representation.

AI USE STATEMENT

Generative AI tools were used only for language polishing and improving the clarity and fluency of the manuscript. All scientific content, technical claims, and experimental results were produced and verified by the authors.

ETHICS STATEMENT

This work uses publicly available remote-sensing imagery and road-network annotations and does not involve human or animal subjects. The method is intended for automated mapping and geospatial analysis; deployment in safety-critical or high-stakes settings should include independent validation of map accuracy.

REPRODUCIBILITY STATEMENT

Section 3 specifies the region action, verification, repair, and policy-learning procedures, while Section 4 reports the dataset split, optimization settings, training schedule, ablations, and evaluation metrics. Full reproduction additionally requires the exact action/repair prompts and configuration files.

REFERENCES

Gaetan Bahl, Mehdi Bahri, and Florent Lafarge. Single-shot end-to-end road graph extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1402–1411, 2022. doi: 10.1109/CVPRW56347.2022.00146.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4720–4728, 2018. doi: 10.1109/CVPR.2018.00496.

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*, pp. 10385–10393, 2019. doi: 10.1109/CVPR.2019.01063.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dengxian Gong and Shunping Ji. Deh4r: A decoupled and hybrid method for road network graph extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–11, 2026. doi: 10.1109/TGRS.2026.3679934.

Wenfei Guan, Jilin Mei, Tong Shen, Xumin Wu, Shuo Wang, Chen Min, and Yu Hu. Beyond endpoints: Path-centric reasoning for vectorized off-road network extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13254–13263, 2026.

Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M. Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. Sat2graph: Road graph extraction through graph-tensor encoding. In *European Conference on Computer Vision*, pp. 51–67, 2020.

Congrui Hetang, Haoru Xue, Cindy X. Le, Tianwei Yue, Wenping Wang, and Yihui He. Segment anything model for road network graph extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2556–2566, 2024. doi: 10.1109/CVPRW63382.2024.00262.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Zuoyue Li, Jan Dirk Wegner, and Aurelien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1715–1724, 2019. doi: 10.1109/ICCV.2019.00180.

Ruyi Liu, Junhong Wu, Wenyi Lu, Qiguang Miao, Huan Zhang, Xiangzeng Liu, Zixiang Lu, and Long Li. A review of deep learning-based methods for road extraction from high-resolution remote sensing images. *Remote Sensing*, 16(12):2056, 2024. doi: 10.3390/rs16122056.

Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3438–3446, 2017. doi: 10.1109/ICCV.2017.372.

Jie Mei, Rou-Jing Li, Wang Gao, and Ming-Ming Cheng. Coanet: Connectivity attention network for road extraction from satellite imagery. *IEEE Transactions on Image Processing*, 30:8540–8552, 2021. doi: 10.1109/TIP.2021.3117076.

Volodymyr Mnih and Geoffrey E. Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pp. 210–223, 2010. doi: 10.1007/978-3-642-15567-3_16.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Radle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. Vecroad: Point-based iterative graph exploration for road graphs extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8910–8918, 2020. doi: 10.1109/CVPR42600.2020.00893.

Adam Van Etten. City-scale road extraction from satellite imagery. *arXiv preprint arXiv:1904.09901*, 2019.

Adam Van Etten, Daniel Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Shengqi Wang, Dongyang Zhang, Junmin Liu, Xiangyong Cao, and Chengli Tan. Seg2vector: Remote sensing road graph extraction via segmentation-to-vector transformation. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026. doi: 10.1109/TGRS.2026.3725197.

Zhenhua Xu, Yuxuan Liu, Lu Gan, Yuxiang Sun, Xinyu Wu, Ming Liu, and Lujia Wang. Rngdet: Road network graph detection by transformer in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12, 2022. doi: 10.1109/TGRS.2022.3186993.

Zhenhua Xu, Yuxuan Liu, Yuxiang Sun, Ming Liu, and Lujia Wang. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *IEEE Robotics and Automation Letters*, 8(5):2991–2998, 2023. doi: 10.1109/LRA.2023.3264723.

Bingnan Yang, Mi Zhang, Zhan Zhang, Zhili Zhang, and Xiangyun Hu. Topdig: Class-agnostic topological directional graph extraction from remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274, 2023. doi: 10.1109/CVPR52729.2023.00128.

Pan Yin, Kaiyu Li, Xiangyong Cao, Jing Yao, Lei Liu, Xueru Bai, Feng Zhou, and Deyu Meng. Towards satellite image road graph extraction: A global-scale dataset and a novel method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1527–1537, 2025. doi: 10.1109/CVPR52734.2025.00150.

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 Workshops*, pp. 182–186, 2018. doi: 10.1109/CVPRW.2018.00034.