RoadGIE: Towards A Global-Scale Aerial Benchmark for Generalizable Interactive Road Extraction

Supplementary Material

A. WorldRoadSeg-360K

A.1. Annotation

Each sample in the `WorldRoadSeg-360K` dataset consists of a high-resolution aerial image and its corresponding road mask. After the initial dataset construction, we invited multiple domain experts in remote sensing and computer vision to carefully curate and select high-quality image-mask pairs for inclusion in the final benchmark. The overall annotation pipeline comprises the following three stages:

- Stage 1: We collect satellite imagery from urban areas across the globe using Google Maps, with a spatial resolution ranging from 0.8 to 1.1 meters per pixel. For each selected city, we define a rectangular region of interest spanning 15–45 km in length. We then retrieve vector road maps from OpenStreetMap (OSM) [37] within these regions and apply morphological operations to thicken the roads, thereby simulating realistic road widths.
- Stage 2: The thickened road maps are used as prompts, together with the original aerial images, as inputs to a variety of state-of-the-art segmentation models, including SAM [18], HQ-SAM [16], RobustSAM [3], and our proposed `RoadGIE`. These models generate initial road masks, which are further refined through morphological post-processing to remove small isolated regions and enhance structural continuity.
- Stage 3: Finally, we perform manual quality control on the generated masks, selecting only those with high boundary accuracy and strong structural completeness. These high-quality samples form the final `WorldRoadSeg-360K` dataset.

A.2. Statistics

The `WorldRoadSeg-360K` dataset spans six continents, with the number of samples per continent summarized in Table 7. As shown, the image distribution is relatively balanced across continents, ensuring strong geographic diversity. To rigorously evaluate the generalization ability of road extraction models, we construct the test set by selecting cities that are entirely excluded from the training set, rather than sampling a fixed proportion of images from each city. This setting better reflects the real-world deployment scenario where models are applied to previously unseen regions.

Table 7. Statistics of WorldRoadSeg-360K across continents.

Continents	Countries	Cities	Images
Asia	3	39	94,096
North America	2	35	55,220
South America	13	44	27,480
Europe	2	30	130,623
Africa	16	66	47,982
Oceania	2	9	11,546
Total	**38**	**223**	**366,947**

A.3. Quality-based Data Partitioning

The classification of high- and low-quality subsets was performed using quantitative criteria based on model consistency and agreement with initial labels. At the early stage of this work, we reproduced multiple state-of-the-art road segmentation models and trained them independently on public remote sensing datasets such as DeepGlobe and SpaceNet. These models were then used to infer predictions on the RoadGIE dataset.

For each image, we computed (1) the average pairwise IoU across all model predictions, and (2) the average IoU between each model prediction and the initial annotation. If either value fell below a threshold (0.65 for model-to-model consistency or 0.60 for model-to-label agreement), the image was flagged as low quality. All such samples were subsequently reviewed by annotators to ensure accurate labeling. To validate the effectiveness of this partitioning, we conducted training experiments using different data combinations. In particular, we compared pretraining on the full dataset followed by fine-tuning on either all data or only the high-quality subset. Table 8 demonstrate that fine-tuning on high-quality data notably improves final performance, especially in terms of APLS.

Table 8. Training performance under different quality-based data partitioning.

Pretrain Subset	Finetune Subset	Dice	APLS
Low	High	83.0	61.7
High	High	83.1	61.7
Low + High	Low + High	82.7	61.3
Low + High	High	**83.5**	**62.0**

B. RoadGIE

B.1. Mitigating Model Degradation

In multi-round interactive road extraction, we observe a critical issue of state forgetting and foreground degradation,

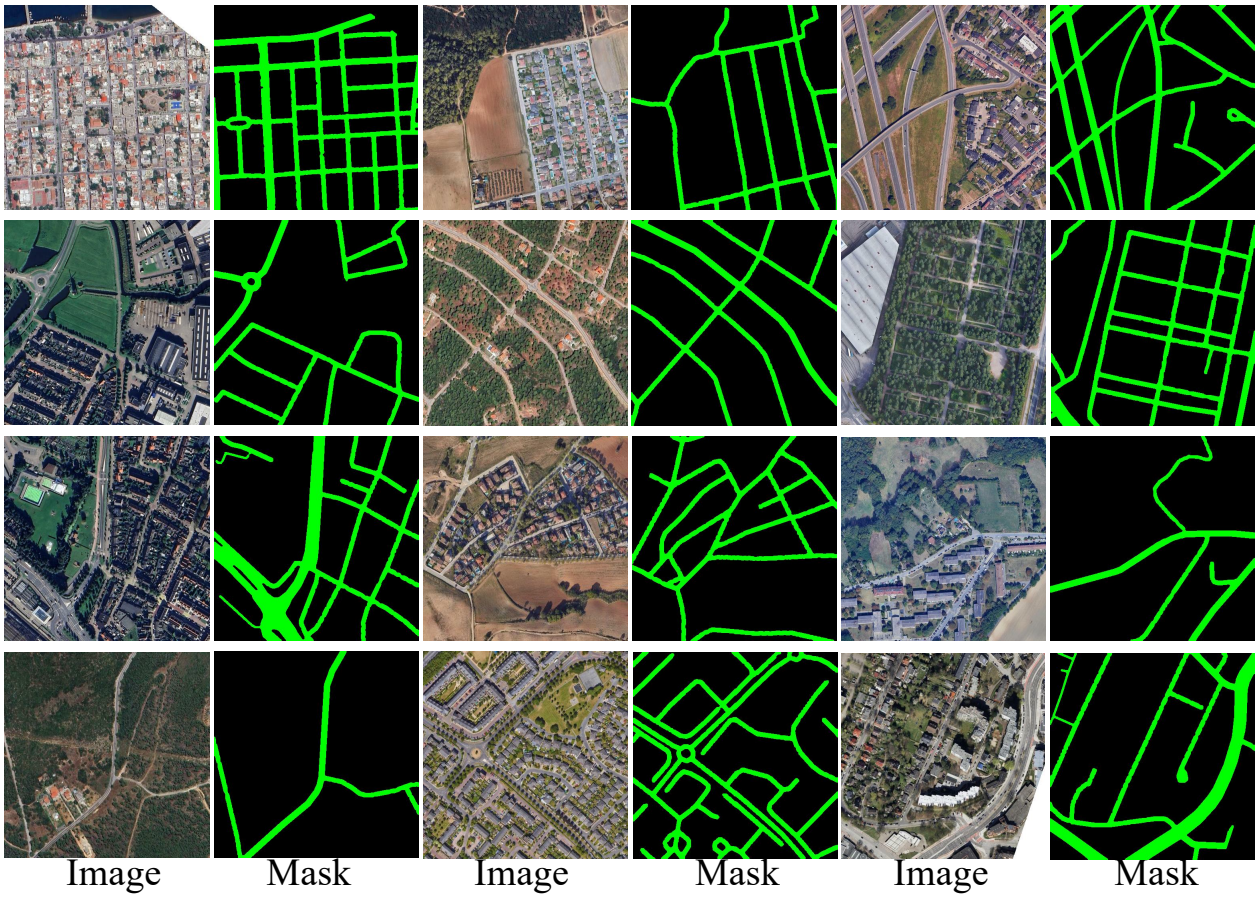

Figure 8. Visualization of representative images and their corresponding masks from the WorldRoadSeg-360K dataset. The displayed images include both urban and mountainous areas, and some road segments are occluded.

Table 9. Effectiveness of EG-Prompt across various interactive models.

Method	Dice	APLS
EISeg	68.3	50.1
+ EG-Prompt	**70.6** (+2.3)	**51.6** (+1.5)
ScribbleSeg-B3	75.9	55.7
+ EG-Prompt	**78.8** (+2.9)	**58.0** (+2.3)
SAM (ViT-h)	71.8	52.4
+ EG-Prompt	**75.6** (+3.8)	**55.3** (+2.9)
PRISM-2D	64.8	48.2
+ EG-Prompt	**66.9** (+2.1)	**49.6** (+1.4)
ScribblePrompt	79.2	57.7
+ EG-Prompt	**80.9** (+1.7)	**59.2** (+1.5)

where the model gradually loses the contextual semantics of previously segmented regions. As interaction progresses, it tends to overfit to the current prompt while ignoring earlier prompt contexts, causing previously extracted road areas to be mistakenly discarded. This semantic drift undermines structural consistency and overall interaction coherence.

To address this, we propose Prompt-excluded Skeleton Loss, an extension of the Skeleton-based Recall Loss [17]. By explicitly excluding prompt regions from gradient computation, the model is discouraged from over-relying on prompt points and instead learns to recover road structures based on global context and topological continuity.

While Focal Loss and Dice Loss are effective for pixel-level segmentation under class imbalance, they are not designed to model structural dependencies, as shown in Table 5. Applying prompt exclusion to these losses may impair their discriminative power and destabilize training due to their localized supervision. In contrast, Skeleton-based Recall Loss [17] operates at a structural level, evaluating the completeness of elongated objects like roads. The prompt-exclusion mechanism naturally complements this objective by focusing learning on the contextual geometry of the road rather than prompt-localized signals.

In summary, Prompt-excluded Skeleton Loss serves as a structure-aware constraint, purposefully integrated into Skeleton-based Recall Loss [17] to mitigate model degradation and preserve structural consistency in multi-step interactive segmentation.

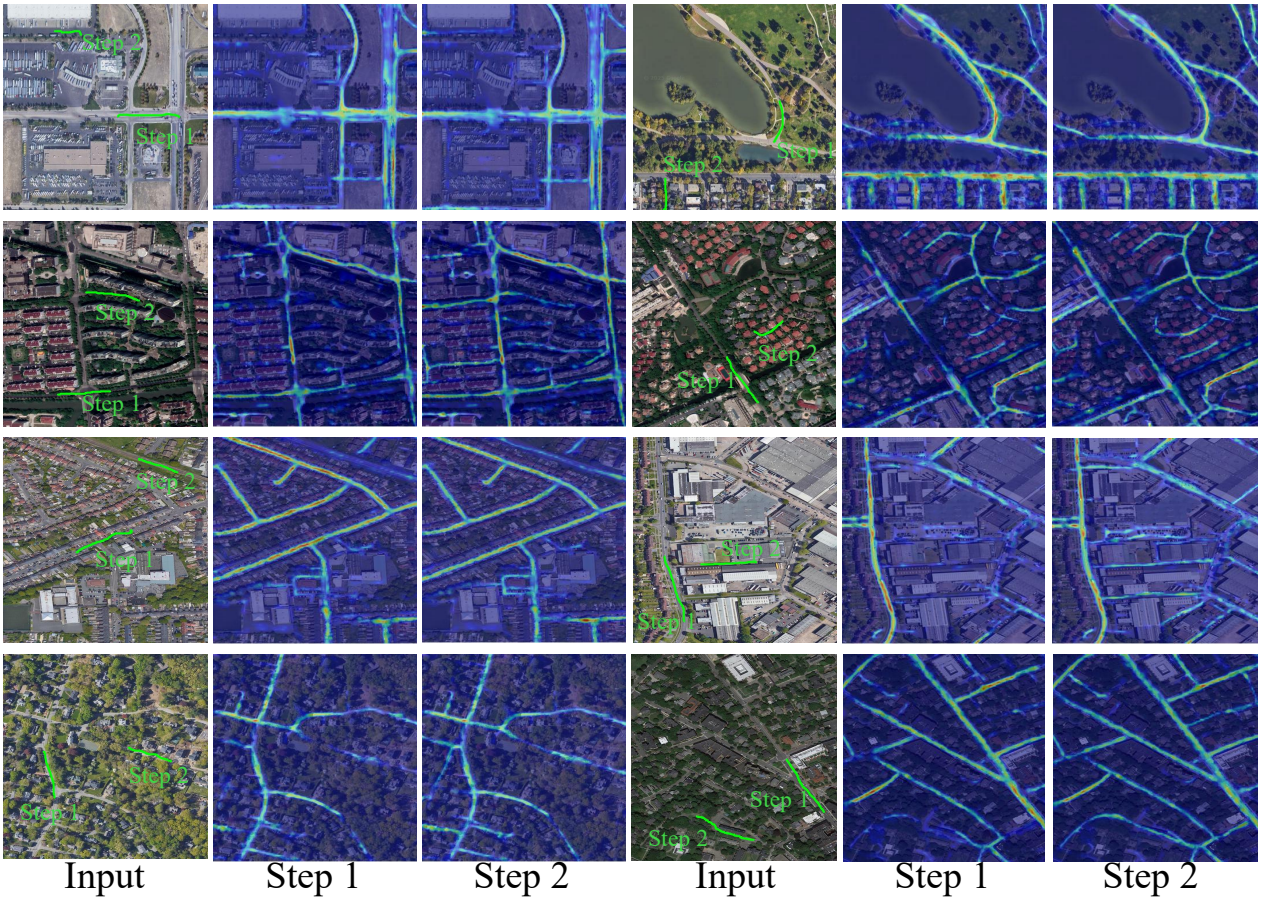

Figure 9. Visualization of the model's output heatmaps after the first and second rounds of user interaction. As shown in the figure, the activation values are highest in the interactive regions, and road areas similar to the interaction regions are also effectively activated.

Table 10. RoadGIE as a refinement module on various backbone architectures.

Arch	Encoder	Decoder	Fold1	Fold2	Fold3	Fold4	Fold5	Avg
CNN	SegNeXt_base	DaFormer	75.1	75.2	76.0	76.1	79.1	76.3 (↓0.7)
	EfficientNetV2-L	CoANet	75.0	75.3	75.8	75.8	78.9	**76.1 (↑1.4)**
	EfficientNetB7	CoANet	74.9	74.8	75.6	75.1	79.0	**75.9 (↑1.0)**
	ConvNeXt_Large	CoANet	75.0	75.5	75.9	75.8	80.9	**76.6 (↑0.8)**
Transformer	Swin_Small	BiconNet	75.1	75.2	75.7	75.6	78.6	76.0 (↓1.4)
	PvT_v2_b4	MSMDFFNet	75.2	75.4	76.1	76.0	79.5	**76.5 (↑1.5)**
	Dual_ViT_b	URoadNet	75.8	75.3	76.4	76.3	78.6	76.5 (↓0.4)
	SMT_base	URoadNet	75.4	76.0	76.3	76.0	79.8	**76.7 (↑2.8)**
	Uniformer_base	URoadNet	75.6	75.8	76.3	76.2	79.2	**76.6 (↑1.2)**
	WaveViT_base	MSMDFFNet	75.7	75.5	76.4	76.1	79.5	**76.6 (↑2.1)**
	MiT_b2	ConnNet	74.9	75.0	75.4	75.6	78.1	**75.8 (↑0.2)**
CNN+Transformer	iFormer_Large	DconnNet	75.3	75.4	76.2	76.0	79.6	**76.5 (↑1.7)**
	HorNet_Small_GF	CoANet	75.2	75.3	76.0	75.8	80.1	**76.5 (↑1.2)**
	CoaT_4L_Small	MSMDFFNet	75.5	75.7	76.3	76.0	79.4	**76.6 (↑3.8)**
	CoaT_4L_Lite_Med	MSMDFFNet	75.6	75.4	76.2	76.2	79.5	**76.6 (↑3.7)**
	CoaT_5L_Parallel	MSMDFFNet	75.6	75.5	76.6	75.9	79.7	**76.7 (↑2.2)**

B.2. Robustness of Expert-guided Prompt

We evaluated EG-Prompt on a range of lightweight and mainstream interactive segmentation models. The consistent performance gains confirm its generalizability and effectiveness, as shown in Table 9. Interestingly, we found that models trained with EG-Prompt can also serve as

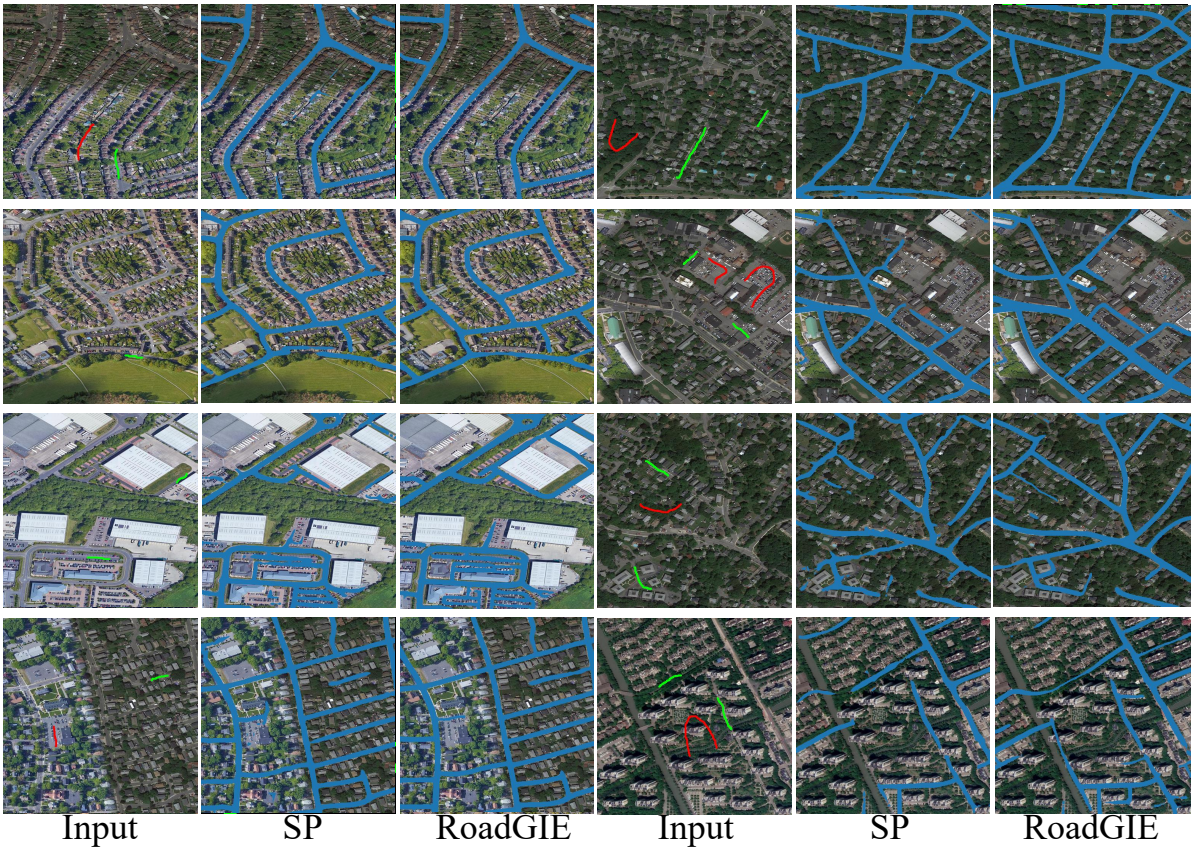

Figure 10. Visualization comparison of different methods. "SP" stands for ScribblePrompt.

strong refinement modules. By feeding predictions from other road extraction models as prompts into RoadGIE, we observed widespread improvements across both CNN and Transformer architectures, as shown in Table 10.

Table 11. Comparison with task-specific and lightweight road extraction models.

Encoder	Decoder	Params (M)	Time (ms)	Dice	APLS
ConvNeXt*	CoANet	13.9	112.8	77.5	56.8
SMT*	URoadNet	3.8	51.3	80.1	59.0
WaveViT*	MSMDFFNet	5.1	101.4	78.7	57.9
iFormer*	DconnNet	2.9	60.8	79.2	58.5
Ours	**Ours**	**3.7**	**39.5**	**80.9**	**59.8**

B.3. Efficient of Model Architecture

A more comprehensive comparison including lightweight and task-specific models is important for a fair and practical evaluation. To this end, we conducted a systematic study of leading road segmentation methods, analyzing various encoder-decoder combinations to identify key architectural components. This guided the design of our own model, which is lightweight, efficient, and suitable for real-time interactive applications. Table 11 presents a detailed comparison between our method, lightweight baselines, and state-of-the-art models tailored for road extraction.

We observed that strip convolutions are highly effective in this task. This may be attributed to the elongated and high-aspect-ratio nature of road structures. Similar principles have also been applied in remote sensing detection tasks such as Strip-RCNN. In addition, rotated convolutions used in methods like URoadNet and MSMDFFNet can be seen as functionally equivalent to strip convolutions. These observations directly inspired the design of our DAM module, which achieves competitive performance with fewer parameters. Furthermore, we leveraged the predictions of multiple task-specific models to construct our Expert-Guided Prompting strategy. This approach uses expert knowledge to guide prompt generation during training, focusing on uncertain regions and improving the model's efficiency and robustness. This is one of the key factors behind the effectiveness of our interactive method.

Table 12. Ablation on training with different backpropagation steps.

Backprop Steps Used	Dice	APLS
Step 5 only	77.6	56.7
Step 1 and 5	79.3	58.4
Step 1, 3, 5	80.2	58.6
All Steps (1–5)	**80.4**	**58.9**

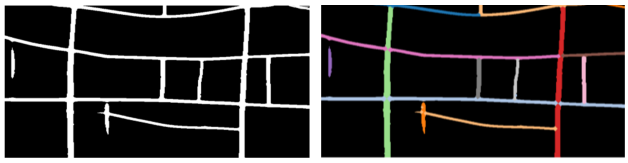

Figure 11. The visualization of topo-semantic coupled instantiation strategy. Different colors represent roads of different instances.

B.4. Impact of Different Backpropagation Steps

Our current implementation accumulates gradients across all interaction steps and performs a single backpropagation at the end. This allows the model to learn from the entire interaction trajectory, but it also scales the training time approximately linearly with the number of steps. To address this, we experimented with a simplified scheme where gradients are not accumulated in the first $n-1$ steps and only the final step contributes to backpropagation. This significantly reduces training time. During inference, this scheme performs comparably well when only 1 to 3 rounds of prompting are needed. However, in more complex cases where 4 or more steps are beneficial, we observed a noticeable performance drop. Table 12 shows the average Dice and APLS scores over five inference rounds, comparing different backpropagation schemes. Using all steps for gradient update yields the best overall performance.

B.5. Effect of Soft Weighting for Skeleton Loss

In our current implementation, we set the skeleton loss to zero within prompt regions to avoid overfitting and to encourage the model to learn global structure primarily from the unlabeled areas. We also conducted pilot experiments on a small-scale dataset by applying soft weights to the skeleton loss. Specifically, we assigned a small weight to prompt regions and a larger weight to non-prompt regions. The results show that this soft weighting scheme can slightly improve performance when the weights of Focal Loss and Soft Dice Loss are relatively low. However, when these two primary losses are given higher weights, the benefits of soft skeleton loss diminish and in some cases even slightly degrade performance.

We hypothesize that this may be due to the fact that Focal Loss and Soft Dice Loss already provide sufficient supervision within the prompt regions. Therefore, focusing the skeleton loss entirely on non-prompt areas leads to better structural recovery in regions where user guidance is absent. The experimental results are summarized in Table 13.

B.6. Effect of Topo-Semantic Coupled Instantiation

Table 14 demonstrates the effectiveness of our instantiation strategy across different methods, highlighting its critical role in improving road structural completeness. By treating roads of different levels as distinct categories and randomly assigning diverse road sets to the prompt regions during training, our approach effectively serves as a form of structure-aware data augmentation, enhancing the robustness and generalization of prompt-based inputs. Figure. 11 presents a visualization of the outcomes of our road instantiation strategy. Figure. 9 presents heatmaps of predicted road activations across different steps, where the user-provided prompts consistently activate a broader range of semantically similar road segments, further underscoring the practical value and efficiency of our strategy.

Table 13. Effect of soft weighting for skeleton loss in prompt and non-prompt regions.

Focal Loss	Soft Dice Loss	Prompt Weight	Non-Prompt Weight	Dice	APLS
0.25	0.25	0.2	0.8	77.4	56.8
0.25	0.25	0.3	0.7	77.5	57.0
0.5	0.5	0.2	0.8	77.9	57.2
1.0	1.0	0.2	0.8	77.6	57.0
1.0	1.0	0	1.0	**78.3**	**57.6**

Table 14. Effectiveness of topo-semantic coupled instantiation.

Method	Dice	APLS
EISeg	68.3	50.1
+ Instantiation	**69.5** (+1.2)	**50.8** (+0.7)
ScribbleSeg-B3	75.9	55.7
+ Instantiation	**77.4** (+1.5)	**56.9** (+1.2)
SAM (ViT-h)	71.8	52.4
+ Instantiation	**73.5** (+1.7)	**54.0** (+1.6)
PRISM-2D	64.8	48.2
+ Instantiation	**65.6** (+0.8)	**48.9** (+0.7)
ScribblePrompt	79.2	57.7
+ Instantiation	**79.9** (+0.7)	**58.5** (+0.8)

B.7. Implementation Details

Table 15 presents the hyper-parameter settings used during the training of our model. `Samples Per Epoch` is set to 1000, indicating that each training epoch uses a random subset of 1000 images. RandomAffine, RandomBrightnessContrast, RandomVariableGaussianBlur, RandomVariableGaussianNoise, RandomHorizontalFlip and RandomVerticalFlip are used for data augmentation.

B.8. Experimental Settings

All baseline models are evaluated using the same input format as `RoadGIE`, including positive clicks or scribbles, negative clicks or scribbles, and the logits from the previous prediction. We exclude bounding box prompts from all comparisons, as we observe that incorporating bounding boxes consistently leads to suboptimal performance across models.

Table 15. Training hyper-parameters.

Parameters	Value
Optimizer	Adam
Learning Rate	3×10^{-4}
Batch Size	8
Number of Epoch	1000
Learning Rate Schedule	Cosine Decay
Prompt Iter	5
Samples Per Epoch	1000
Image Size	512×512
Image Processing	Random crop to 512×512

C. Examples

C.1. WorldRoadSeg-360K

Fig. 8 presents representative remote sensing road images from the WorldRoadSeg-360K dataset, covering diverse terrain types. The images include roads of various levels and highlight typical cases where roads are occluded by vegetation.

C.2. RoadGIE

Figure. 9 illustrates the evolution of the model's activation maps after user interactions. As observed, the model responds to user input by accurately highlighting the intended road segments, while simultaneously generalizing to other structurally similar regions. In contrast, initially spurious areas with low activation are effectively suppressed. These results underscore RoadGIE's ability to infer road semantics and topology based on minimal user guidance.

Figure. 10 presents a visual comparison between the second-best method, ScribblePrompt [42], and our approach. As shown in the figure, our method is able to activate more previously disconnected road segments after user interaction, while preserving the optimal segmentation of regions that were already well delineated.