remote sensing

MDPI

Article

Geometry-Aware Supervision for Connectivity-Preserving Road Extraction via SAM2

Liming Xie, Zhiqian Jiang, Yu Zhang *, Jinlong Li and Kai Yang

School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 610031, China; xlm@swjtu.edu.cn (L.X.); jiangzhiqian@my.swjtu.edu.cn (Z.J.); jinlong_lee@swjtu.edu.cn (J.L.); yangkai@swjtu.edu.cn (K.Y.)

* Correspondence: zhangyu2@swjtu.edu.cn

Highlights

What are the main findings?

- Geometry supervision turns a parameter-efficient SAM2 adaptation into a connectivity-preserving extractor: aggregate CC-agreement rises from 0.4762 (D-LinkNet) to 0.6874, i.e. 44%, and is 10% above the same-backbone SAM2-Adapter (0.6263), while CC-err falls from 4.0172 to 1.0861. These margins are 8.1× and 2.4× the empirical uncertainty scale of the metric, whereas the IoU margin over the strongest baseline stays within its own scale, so the gain is topological rather than regional. It is also regime-dependent: joint training lifts Massachusetts CC-agreement by 0.2645 (2.5× that domain's scale), and the geometric branch still clears the scale with only 10% of DeepGlobe annotations.
- The gain carries a measurable cost. Aggregate APLS falls to 0.5812, below all baselines, because a better-connected mask can trade away path precision. A 2×2 ablation attributes this drop to the soft-clDice term (-0.0644) rather than to the geometric branch (-0.0041); geometry improves connectivity only in combination with soft-clDice. Region overlap and CC-agreement also rank models differently.

What are the implications of the main findings?

- The geometric prior is a deterministic function of the mask the task already requires, so it needs no extra annotation or generator, and its heads are removed after training (recovering 0.73 M parameters). It can therefore be added to an existing road-segmentation backbone as a training-time regulariser with no inference cost.
- Because IoU is nearly blind to a few-pixel break and the two metric families rank models differently, selecting or reporting on IoU alone can hide connectivity loss. Road-extraction comparisons should report both families against an empirical uncertainty scale and treat only differences exceeding that scale as improvements. The construction depends on mask geometry rather than road class, so it extends to rivers, canals, pipelines, and field boundaries.

Abstract

Road extraction from remote sensing imagery is a structured problem: a mask with high region overlap may still contain breaks that make the road network unusable, and the difficulty grows when annotated sources are small, heterogeneous, and unequally represented during joint training. We revisit this problem on a SAM2 encoder while keeping all added complexity in training rather than in the deployed network. First, we do not use SAM2 as a black box: its top-down pathway refines only the coarsest levels, leaving the quarter-resolution features our decoder needs semantically shallow, so a feature-enhancement neck

Check for updates

Academic Editor: Zhenwei Shi

Received: 15 September 2026
Revised: 10 November 2026
Accepted: 14 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

restores the missing top-down path and a directional attention module matches the elongated geometry of roads. Second, we decompose road structure into five complementary geometric views—skeleton, direction, width, neighbourhood affinity, and endpoints—used as a shared training-time prior; each is a deterministic function of the segmentation mask, so the prior needs no extra annotation and no learned generator, and its heads are removed after training, leaving the inference graph free of auxiliary computation; the feature-enhancement neck and directional attention module remain in the deployed model, as they serve the segmentation head. A soft-clDice term supervises centrelines, and a square-root sampler rebalances the 5.2:1 update-frequency skew across sources. On Massachusetts, DeepGlobe, and CHN6-CUG, the model reaches an aggregate IoU of 0.7403 and CC-agreement of 0.6874, improving CC-agreement by 44% over D-LinkNet and 10% over the strongest same-backbone adaptation. We also make evaluation part of the contribution: we measure an empirical uncertainty scale for topology metrics and hold every claim to exceed it. Held to this standard, the prior's CC-agreement gain clears the scale only in the scarce-annotation regime that limits operational mapping yet costs nothing to retain elsewhere. Finally, region overlap and CC-agreement rank models differently, so both metric families should be reported against a stated measurement scale.

Keywords: road network extraction; remote sensing image segmentation; geometric auxiliary supervision; connectivity preservation; SAM2; parameter-efficient fine-tuning; multi-dataset joint training

1. Introduction

Road networks are fundamental geographic features for navigation, urban planning, disaster response, and map updating. In these applications, the useful output is not simply a binary image but a connected representation of the network. A small gap in a narrow road can disconnect a route while changing IoU only marginally. Consequently, road extraction should be evaluated by both region accuracy and structural continuity. Figure 1 shows what this means on a single validation tile. A baseline prediction recovers almost the entire road area, so its IoU is high and the output looks correct, yet its centreline is severed at several narrow occluded points. Each gap is a few pixels across and therefore changes region overlap almost not at all, while turning one traversable network into several disconnected pieces. Region overlap cannot express this difference, so a model selected on IoU alone may not be the model that produces a usable network.



Figure 1. A prediction can be accurate as a region and unusable as a network. A single output from D-LinkNet [1] on a DeepGlobe validation tile is shown in two ways. (**a**) The input image. (**b**) The prediction against the reference, with green for recovered road, red for missed reference road, and blue for false positive. (**c**) The centreline of that same prediction, with each connected piece in a separate colour and the dashed box marking the window enlarged in (**d**). (**d**) One interruption enlarged.

Recent remote sensing models improve road extraction with long-range context and deformable transformers [2,3], directional or strip convolutions [4], connectivity branches and connectivity-preserving objectives [5–7], graph reasoning [8], boundary-aware fusion [9], and iterative or direction-guided refinement [10]. These strategies reduce fragmentation, but they do not remove the mismatch between pixel-wise training objectives and network-level usability. The mismatch is particularly important in multi-source training. Massachusetts, DeepGlobe, and CHN6-CUG differ in sensor, geography, resolution, annotation convention, and training-set size. Proportional sampling allows the largest source to dominate the optimisation, whereas a model trained on a small source alone may not observe enough examples to learn domain-general road structure. The controlled results in this study also show that a model can rank highly by IoU while producing poor connectivity, so an overlap-only comparison does not fully describe the quality of the extracted feature. Figure 1 shows the same phenomenon on a single tile: almost the entire road area is recovered, so by region overlap the output appears satisfactory, yet the network is severed at several narrow occluded points and its connectivity is far below what the IoU would suggest. The interruption in panel (d) is a gap a few pixels across, contributing almost nothing to region overlap while disconnecting the route; the centreline is drawn rather than the mask because a road is 10–20 pixels wide and a three-pixel gap is invisible at tile scale. The example uses D-LinkNet rather than the proposed model, so the effect belongs to the metrics and not to our method; the tile was selected to show the divergence rather than sampled at random, and the domain-averaged behaviour is in Section 5.1.

We address these issues with a design that concentrates its added complexity in training while keeping the deployed network simple. On the architecture side, we do not treat SAM2 as a fixed feature source: because its top-down pathway refines only the coarsest levels, we add a feature-enhancement neck that restores the top-down path for the quarter-resolution features our decoder needs, together with a directional attention module suited to the elongated geometry of roads. On the supervision side, the resulting shared feature predicts the road mask together with five geometric descriptions derived from the same mask—skeleton, direction, width, local affinity, and endpoints or junctions—that act as a decomposed structural prior. These targets require no additional annotation and no learned generator, and their prediction heads are disabled at inference. We combine this prior with soft-clDice and a square-root domain sampler. The former encourages centreline agreement, while the latter, evaluated as an isolated intervention, gives smaller sources a larger share of updates without discarding the statistical mass of larger sources.

The study focuses on when this intervention is useful rather than treating one aggregate score as sufficient evidence. We report metrics by domain, quantify the measurement scale of the topology metrics, compare controlled baseline families, and vary the amount of training annotation while holding the domain fixed. We also state the remaining failure cases explicitly: the proposed model reduces fragmentation in the reported experiments, but it still produces disconnected segments when occlusion removes both road appearance and reliable structural evidence.

The main contributions are as follows.

1. **A SAM2 encoder adapted for dense road prediction, not used as a black box.** We diagnose that the SAM2 top-down pathway refines only its two coarsest levels, so the quarter-resolution features our decoder consumes are semantically shallow. We restore the missing top-down path with a feature-enhancement neck and add a directional attention module whose 1×7 and 7×1 kernels are matched to the elongated geometry of roads. This targeted redesign, rather than a generic fine-tuning of the backbone, is what makes a foundation-model encoder effective for thin-structure extraction.

2. **A decomposed geometric prior that lives only in training.** We represent road structure through five complementary and mutually reinforcing views—skeleton, direction, width, neighbourhood affinity, and endpoints/junctions—that jointly shape one shared feature together with a soft-clDice term. Every view is a deterministic function of the mask that already supervises segmentation, so the prior needs no extra annotation and no learned label generator; the auxiliary heads are removed after training, reducing the parameter count by 0.73 M without adding any auxiliary computation to the deployed model; the feature-enhancement neck and directional attention module remain, as they serve the segmentation head directly. A domain-resolved ablation shows these five views are best treated as one integrated prior rather than as independently tunable modules, which is why they are supplied together rather than stacked selectively.
3. **An empirical measurement scale for topology metrics, supplemented by paired bootstrap significance testing.** We quantify how far connectivity metrics fluctuate between epochs of a converged run and use that scale as a diagnostic indicator; formal significance is established by paired bootstrap confidence intervals and permutation tests across three random seeds. The three-sigma scale is 0.106 for Massachusetts CC-agreement and 0.037 for DeepGlobe—larger than the gaps separating several configurations routinely reported as improvements. Held to this standard, a single-domain experiment that varies only annotation volume locates precisely where the prior pays off: its CC-agreement gain falls monotonically from 0.0479 at 10% of DeepGlobe to 0.0100 at full annotation, clearing the scale only in the scarce-annotation regime that limits operational mapping. Because the prior costs nothing at inference, it is retained unconditionally.
4. **Evidence that region overlap and CC-agreement rank models differently.** The highest-IoU baseline on DeepGlobe produces one of the most fragmented networks, so selecting a model on IoU alone can return a map that is accurate as a region yet unusable as a network. We therefore argue that region-overlap and topology metrics should be reported together and against a stated measurement scale, and we adopt this protocol throughout.

2. Related Work

2.1. Recent Road Extraction and Connectivity Modeling

Road extraction from remote sensing imagery has been driven by deep segmentation networks, and recent surveys document a steady shift from plain encoder–decoder models towards architectures that model long-range context and road structure explicitly [11]. A first line of work targets occlusion and elongated-feature context. CRRENet extracts multi-view contextual information with parallel dilated convolutions and multi-stage feature fusion for winding rural roads under heavy vegetation cover [12], while CRNet decouples global and local context and refines skip connections to suppress background noise, reporting gains in both accuracy and connectivity under occlusion [13]. A second line combines complementary backbones: HSN-Net integrates a Transformer with a CNN and adds a road-continuity perception module to reduce fragmented segments on DeepGlobe and CHN6-CUG [14]. More broadly, state-space (Mamba) models have begun to replace quadratic attention for efficient dense prediction of remote sensing scenes, as in UMFormer and UrbanSSF [15,16], indicating that global context can be captured at lower cost. These methods confirm that connectivity requires explicit modelling, but most add inference-time modules, post-processing, or task-specific architectural capacity. Our framework instead uses geometric predictions only as training supervision and evaluates their contribution under a common backbone and optimisation protocol.

2.2. Topology-Aware and Geometry-Guided Supervision

Region losses such as Dice and binary cross-entropy do not directly penalise changes in the centreline or component structure. The clDice loss addresses this gap by comparing soft skeletons of predictions and reference masks [17]. Skeleton Recall Loss provides a lower-cost alternative for preserving connectivity in thin tubular structures [18]. Such topology-aware objectives have been carried into road extraction: SAM2-RoadNet, for instance, attaches a soft-clDice term to a foundation-model encoder to enforce structural continuity and reduce fragmentation [19]. In parallel, auxiliary targets have been used to encode orientation, boundaries, corners, or local connectivity as additional supervision signals [3,5,7].

The present work differs in emphasis. We use five deterministic geometric representations derived from the binary road mask and ask how they should be combined rather than which single one wins. Under a unified schedule, no proper subset consistently outperforms the full set, and the differences between subsets largely fall within the measured uncertainty. We read this not as evidence that the tasks are redundant but as evidence that they express one shared structural capacity: each view constrains the same feature from a different geometric angle, so the informative object is the prior as a whole. We therefore supply all five as an integrated prior—complete because it costs nothing at inference to retain—while declining the stronger and unsupported claim that every individual task is indispensable.

2.3. Foundation Models and Multi-Source Training

Foundation models provide a strong initialization for dense prediction, but remote sensing imagery differs from the natural-image distribution used in their pretraining. SAM2 supplies a hierarchical image encoder with strong visual features [20]; recent remote sensing adaptations such as RS-SAM modify the foundation model to better capture multi-scale remote sensing structure [21], and RSAM-Seg inserts scale and feature adapters into the ViT blocks to inject task-specific priors, reporting competitive road extraction even under few-shot supervision [22]. This adapter route has been specialised to roads: SAM2MS transfers SAM2 with an adapter and a multi-scale subtraction decoder [23], while SAM2-RoadNet uses the SAM2 encoder purely as a feature extractor, adds receptive-field blocks and a weighted BiFPN, and supervises with soft-clDice [19]. These works establish that a SAM2 encoder with parameter-efficient adaptation is a strong basis for road extraction, but they improve connectivity mainly by adding architectural capacity—multi-scale fusion necks, subtraction decoders, or refinement blocks that persist at inference. Our design keeps the adaptation minimal and moves the structural inductive bias entirely into training-time geometric supervision, so the deployed network carries no auxiliary-head cost at inference. We compare parameter-efficient and full adaptation directly through SAM2-LoRA, SAM2-Adapter, and full fine-tuning controls.

Multi-source training adds a second difficulty because datasets differ in both distribution and size. Instead of introducing domain-specific branches or requiring domain labels at inference, we use square-root sampling to moderate the update frequency imbalance. This simple data-level intervention is evaluated separately from the geometric branch so that the two sources of improvement are not conflated.

3. Materials and Methods

3.1. Problem Formulation and Framework Overview

Let $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{M}_i, d_i)\}_{i=1}^{N}$ denote the union of the training sources, where $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ is an RGB remote sensing image, $\mathbf{M}_i \in \{0,1\}^{H\times W}$ is its road mask, and $d_i \in \{1,\ldots,D\}$ is

the domain label. The primary model learns $f_\theta(\mathbf{I}) = \hat{\mathbf{M}}$, while the training objective also uses geometric targets derived from $\mathbf{M}$.

The framework, which we refer to as GARE (Geometry-Aware Road Extraction), has two training-time interventions (Figure 2). First, a shared feature map feeds one segmentation head and five lightweight geometric heads. The auxiliary heads encourage the feature map to represent road structure in addition to appearance. Second, square-root sampling assigns each source a probability based on the square root of its training size, so that the smallest source retains a meaningful share of the updates, with each mini-batch drawn from a single domain selected according to p_d. At inference, only the segmentation head is evaluated. The deployed computation graph therefore contains no geometric heads or pseudo-label generation.

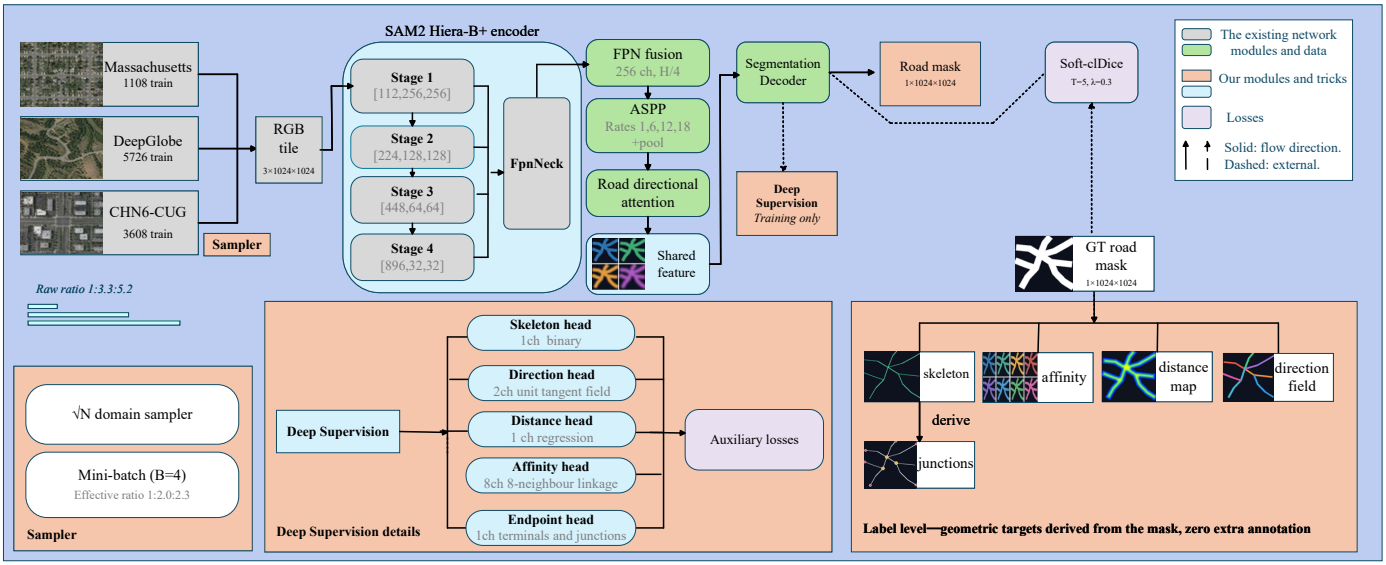


Figure 2. Overview of the proposed geometry-aware multi-source framework. The SAM2 encoder is drawn as its two published parts (the Hiera trunk and the pyramid neck); the neck produces one 256-channel shared feature at quarter resolution, which the segmentation head and five geometric heads read. The horizontal band separates components shared by training and inference from those that exist only during training. Among the five geometric targets, only the endpoint/junction heatmap depends on the skeleton; the remaining four (direction, distance, affinity, and skeleton itself) are computed independently from the road mask $\mathbf{M}$ via parallel operators, and the arrows should be read as parallel rather than sequential. The soft-clDice module receives both the predicted mask and the ground truth mask $\mathbf{M}$ as inputs, as specified in Equations (6)–(7).

3.2. Geometric Auxiliary Targets

The operators that generate the five geometric targets from the reference mask are shown in Figure 3. The five targets describe complementary views of the same road geometry. The skeleton is obtained by morphological thinning and represents the centreline and global connectivity. The direction field contains a unit tangent vector at road pixels and encodes local orientation. The distance transform gives each road pixel its distance to the nearest background pixel and therefore represents local width. The affinity map records whether each pixel and each of its eight neighbours belong to the road, providing a local redundancy of the mask boundary: because each affinity channel is the logical AND of the mask and a one-pixel shift, the eight channels reconstruct the same binary mask up to a one-pixel offset, so the signal is not an independent source of structural information but a regulariser that encourages the feature to encode local boundary consistency. The endpoint/junction heatmap marks terminals and branch points detected on the skeleton. It represents graph nodes that are especially sensitive to missed connections. Every operator

is deterministic and has no parameters, so the five geometric targets add no labelling beyond the mask that already supervises segmentation; the arrows in Figure 3 emphasise that the segmentation loss and all five geometric losses are supervised by the same tensor. The rendering choices in that figure affect only its legibility: the tensors passed to the losses are those produced by the operators, unmodified. Endpoint dilation uses a radius of two pixels to reduce the sparsity of that target.

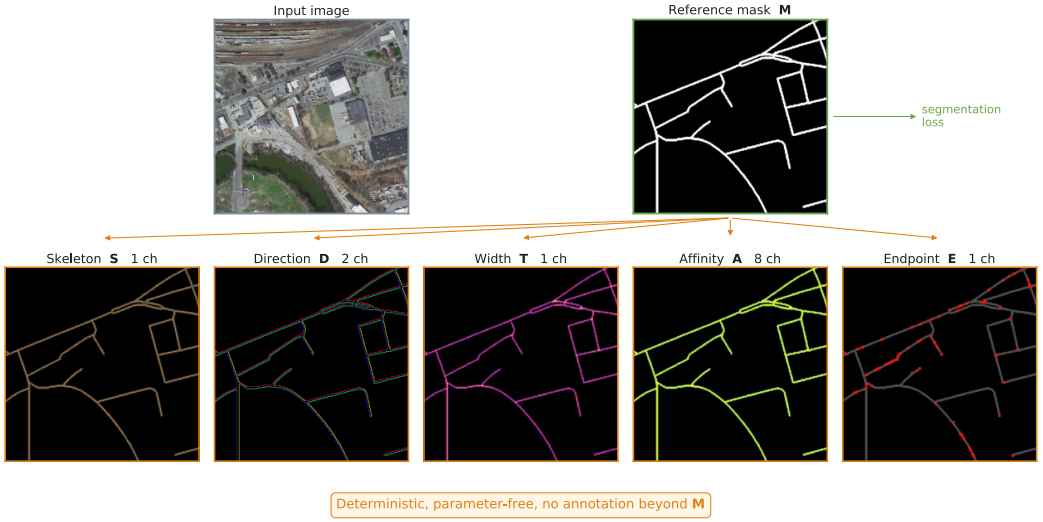


Figure 3. Generation of the five geometric targets from one reference mask, shown on a Massachusetts validation tile. **Top**: the input image and the reference mask **M**. **Bottom**: the five targets, labelled with their channel counts, in the order produced by thinning, rotation of the mask gradient, distance transform, eight-neighbour consistency, and branch-point detection. The skeleton is overlaid on a dimmed copy of the mask; the direction field uses hue to encode the tangent angle; the distance transform and affinity map use sequential colour maps.

3.3. Automatic Geometric Target Generation

All geometric targets are generated from the reference mask without additional annotation or a learned model. Skeletons use Zhang–Suen thinning [24]. Direction fields use Sobel gradients, a 90° rotation, and vector normalisation. Distance maps use the Euclidean distance transform. For the eight-neighbour affinity map, each channel is the logical AND of the mask and its one-pixel translation. A 3×3 crossing-number test on the skeleton detects endpoints and junctions, and those points are dilated by a radius of two pixels so that the target is not a handful of isolated pixels. Target generation runs in the data-loading workers.

Figure 3 shows the five operators applied to one validation tile. The mask at the top is the same tensor that supervises the segmentation head, so the six supervision signals of this framework are all functions of a single annotated mask.

3.4. SAM2 Encoder and Feature-Enhancement Neck

We use the SAM2 Hiera-base-plus image encoder and fine-tune it end to end. The encoder learning rate is one tenth of the decoder learning rate.

The encoder already contains a feature pyramid, and we state why a second one follows it. Figure 2 shows this second pyramid and the reason it is required. The Hiera trunk emits four levels at strides 4, 8, 16 and 32 with 112, 224, 448 and 896 channels, respectively. SAM2 projects each to 256 channels and discards the coarsest output, so three levels reach the downstream network. Its top-down pathway, however, is applied only at the two coarsest levels: the stride-32 level enters the stride-16 level through one projection and one nearest-neighbour upsampling, while the stride-4 and stride-8 outputs

are one-by-one projections of their trunk stages with no contribution from deeper layers. Since our decoder reads quarter resolution, those features would be shallow. We therefore apply a standard feature pyramid [25] across the three levels, which supplies the missing top-down path and reduces them to one map; the design serves the dense prediction task rather than the prompt-conditioned decoding SAM2 was built for.

The fused map is processed by atrous spatial pyramid pooling [26] and enhanced by a directional attention module. For a shared feature map $\mathbf{F}$, the directional module is

$$\mathbf{F}_{\text{dir}} = \mathbf{F} \odot \sigma(\text{Conv}_{1\times7}(\mathbf{F}) + \text{Conv}_{7\times1}(\mathbf{F})). \quad (1)$$

The resulting 256-channel feature map $\mathbf{F}'$ is used by all prediction heads.

3.5. Prediction Heads and Inference

The segmentation head uses an FPN decoder with two upsampling stages and emits the full-resolution road probability map. It additionally emits a deeply supervised intermediate prediction at half resolution, which is used only in the training objective. Each geometric head is a two-layer convolutional block followed by bilinear upsampling. The skeleton, distance, and endpoint/junction heads emit one channel; the direction head emits two channels; and the affinity head emits eight channels.

The geometric heads are enclosed in a training-mode branch. They are not executed during inference, and their targets are not generated in the deployment pipeline. The inference graph therefore retains the encoder, the feature-enhancement neck, and the directional attention module alongside the segmentation head, but none of the five geometric heads. The output is the segmentation probability map thresholded at 0.5.

Two consequences of this arrangement are worth separating. All six heads read the same tensor $\mathbf{F}'$, so no head has private capacity and any structure a geometric head requires must be present in the representation that the segmentation head also consumes; this is the sense in which the auxiliary tasks act as a regulariser on a shared feature rather than as independent predictors. Separately, the geometric heads exist only while training, so the deployed computation graph is that of a single-head segmentation network. The geometric quantities are supervision signals and never inputs: at inference the network receives an image and emits a mask, and applying the trained model to newly acquired imagery requires no skeleton, no direction field, and no annotation of any kind. What the construction removes is therefore the marginal cost of the structural prior, not the cost of the segmentation mask itself; the method remains fully supervised.

A branch that is discarded before deployment invites the question of whether its heads are learning at all, since a head that never converges would leave the shared feature unchanged and cost nothing to remove. Figure 4 addresses this directly from the training record. Each curve is normalised by its own value at the first recorded epoch, because the geometric terms use different objectives—binary cross-entropy, cosine, and smooth L_1—and span two orders of magnitude in absolute value, so they cannot be read on a shared axis. All five losses reach a plateau, so each head receives a usable signal from the shared feature. The reductions are not compared across tasks: the direction curve moves least, but its target is degenerate over most of the pixels it supervises (Section 6.5), so that movement does not measure what the head learned. Removing the heads reduces the parameter count by 0.73 M without adding auxiliary computation to the deployed graph.

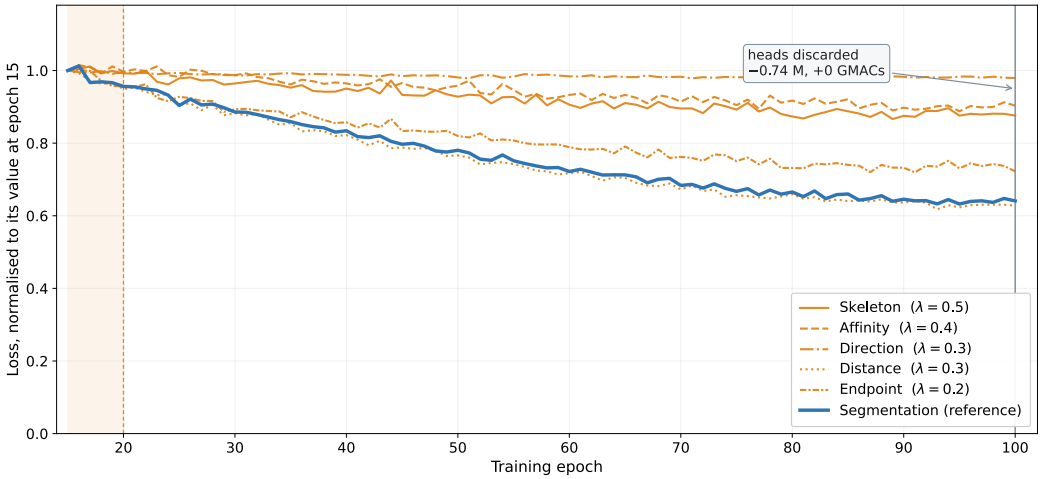


Figure 4. Auxiliary supervision is introduced under a warm-up, converges, and is then removed. The segmentation loss is drawn in blue for reference. The shaded interval is the recorded part of the warm-up; the dashed line at its edge is epoch 20, after which the geometric weights act at full strength. The solid line at the right marks the end of training, where the five heads are removed.

3.6. Square-Root Domain Sampling

Following square-root sampling [27], for a domain d with N_d training images, the sampling probability is

$$p_d = \frac{\sqrt{N_d}}{\sum_{d'=1}^{D} \sqrt{N_{d'}}}. \tag{2}$$

The square root reduces the difference in update frequency between domains while preserving the larger statistical contribution of larger sources. In this study, the effective ratio between the largest and smallest sources changes from about 5.2:1 under proportional sampling to 2.3:1. At each step a domain is drawn with probability p_d, and the mini-batch is filled from that domain only, so no mini-batch mixes images from different sources.

3.7. Training Objective

The total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{cl}} \mathcal{L}_{\text{clDice}} + \sum_{k \in \{\text{sk,dir,dist,aff,ep}\}} \lambda_k(e) \mathcal{L}_k. \tag{3}$$

The segmentation loss combines three region terms with a deeply supervised auxiliary prediction:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}} + \lambda_{\text{focal}} \mathcal{L}_{\text{Focal}} + \lambda_{\text{tv}} \mathcal{L}_{\text{Tversky}} + \lambda_{\text{ds}} \mathcal{L}_{\text{Dice}}^{\text{aux}}, \tag{4}$$

where $\mathcal{L}_{\text{Dice}}$, $\mathcal{L}_{\text{Focal}}$, and $\mathcal{L}_{\text{Tversky}}$ are the Dice [28], focal [29], and Tversky [30] losses, and $\mathcal{L}_{\text{Dice}}^{\text{aux}}$ is applied to the intermediate decoder prediction at half resolution. We use $\lambda_{\text{dice}} = 1.0$, $\lambda_{\text{focal}} = 0.5$, $\lambda_{\text{tv}} = 0.5$, and $\lambda_{\text{ds}} = 0.3$. Auxiliary weights use a linear warm-up:

$$\lambda_k(e) = \lambda_k^{\max} \min\left(1, \frac{e}{E_{\text{warmup}}}\right). \tag{5}$$

3.7.1. Soft-clDice Loss

Let $\mathcal{S}(\cdot)$ be the differentiable soft-skeleton operator [17] with τ iterations. Topology precision and sensitivity are

$$T_{\text{prec}} = \frac{\sum \mathcal{S}(\hat{\mathbf{M}}) \odot \mathbf{M}}{\sum \mathcal{S}(\hat{\mathbf{M}})}, \qquad T_{\text{sens}} = \frac{\sum \mathcal{S}(\mathbf{M}) \odot \hat{\mathbf{M}}}{\sum \mathcal{S}(\mathbf{M})}. \tag{6}$$

The loss is [17]

$$\mathcal{L}_{\text{clDice}} = 1 - \frac{2T_{\text{prec}}T_{\text{sens}}}{T_{\text{prec}} + T_{\text{sens}}}. \tag{7}$$

We set $\lambda_{\text{cl}} = 0.3$ and $\tau = 5$.

3.7.2. Geometric Auxiliary Losses

Skeleton and affinity use binary cross-entropy. Direction is trained by cosine similarity on road pixels $\Omega = \{(i,j) \mid M_{i,j} = 1\}$:

$$\mathcal{L}_{\text{dir}} = 1 - \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \frac{\mathbf{D}_{i,j} \cdot \hat{\mathbf{D}}_{i,j}}{\|\mathbf{D}_{i,j}\| \|\hat{\mathbf{D}}_{i,j}\| + \epsilon}. \tag{8}$$

Averaging over all of Ω has a consequence we return to in Section 6.5: $\mathbf{D}$ is a zero vector wherever the mask gradient vanishes, which is the interior of a road, and those positions contribute a constant with no gradient. The main experiments therefore use this Sobel form. A skeleton-derived variant that restricts supervision to a ± 5 px valid-pixel band around the centreline avoids the degenerate interior entirely; its aggregate CC-agreement (0.669) is slightly below that of the Sobel form (0.6874), but its Massachusetts endpoint F1 is higher (0.6428 vs. 0.6354) , as shown in Table 1 below (these are direction-target diagnostic values from the auxiliary endpoint head, distinct from the post-hoc skeleton-derived End. F1 so it is a viable alternative when path-endpoint accuracy is prioritised. A quantitative diagnosis of the degeneracy and the skeleton-derived variant is given in Section 6.5.

Distance uses smooth-L_1 regression on road pixels, and the endpoint/junction heatmap uses binary cross-entropy with positive reweighting. The maximum auxiliary weights are $\lambda_{\text{sk}} = 0.5$, $\lambda_{\text{dir}} = 0.3$, $\lambda_{\text{dist}} = 0.3$, $\lambda_{\text{aff}} = 0.4$, and $\lambda_{\text{ep}} = 0.2$.

Table 1. Direction-target diagnostic comparison: endpoint F1 under three direction supervision variants. These values are computed directly from the auxiliary endpoint heatmap prediction during training, which targets the endpoint head explicitly; they differ from the post-hoc skeleton-derived End. F1, because that metric is extracted from the final segmentation mask's skeleton after inference, where direction supervision has already shaped the segmentation. Bold marks the best per domain.

Direction Supervision	Massachusetts	DeepGlobe	CHN6-CUG
Sobel gradient (main experiments)	0.6354	0.6155	0.5710
Skeleton-derived tangent	**0.6428**	0.6087	**0.5794**
None (direction head removed)	0.6079	**0.6186**	0.5682

3.8. Training Procedure

Algorithm 1 summarises the complete procedure. The geometric targets are regenerated from the reference mask inside the data-loading workers at every iteration, and the geometric heads are discarded before inference.

Algorithm 1 Geometry-aware multi-source training

Require: training datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_D\}$ and number of epochs E
Ensure: trained parameters θ

```
Initialise the SAM2 Hiera-base-plus encoder from pretrained weights
Set p_d ∝ √N_d
for epoch e = 1 to E do
    for each mini-batch sampled from one domain do
        Extract shared feature F′ through the encoder and neck
        Predict M̂ and the five geometric outputs
        Derive geometric targets from M
        Compute L_total
        Backpropagate with gradient clipping at maximum norm 1.0
        Update parameters with AdamW and differential learning rates
    end for
    Update learning rates by cosine annealing
end for
Disable the geometric heads and retain the segmentation head for inference
return θ
```

3.9. Implementation Details

We use AdamW with cosine learning-rate annealing. The decoder learning rate is 2×10^{-4} and the encoder learning rate is 2×10^{-5}. Training uses 1024×1024 inputs, batch size 4, eight data-loading workers, FP32 precision, and one GPU per experiment. Augmentation includes horizontal and vertical flips, random rotation, and photometric jitter. The auxiliary warm-up lasts 20 epochs, and the main runs use seed 42, while the three-seed repetition of the significance analysis uses seeds 42, 123, and 2024. The final model is trained for 100 epochs.

4. Experimental Setup

4.1. Datasets

We evaluate three road extraction datasets with different sensors, regions, resolutions, and training-set sizes. Massachusetts Roads [31] contains 1108 training images, a 14-image validation set, and an official 49-image test set; we use the 49-image test set as our evaluation set because the 14-image validation set is too small for reliable bootstrap significance testing (Section 5.1). DeepGlobe Roads [32] contains 6226 labelled satellite images; we hold out 500 images from the training split for validation, leaving 5726 training images. CHN6-CUG [33] contains 4511 tiles from six Chinese cities, with 3608 training images and 903 validation images.

All images are processed at 1024×1024. Massachusetts and DeepGlobe images are resized, and CHN6-CUG images are upsampled from 512×512. The resulting training-set sizes differ by about 5.2:1.

4.2. Baselines and Model Variants

The controlled comparison includes D-LinkNet, DeepLabV3+, and SegFormer as representative road-specific and general segmentation baselines [1,26,34]. We also evaluate SFA-Net (a recent lightweight remote sensing segmentation model [35]) and D2LS (a recent dynamic-dictionary model for remote sensing image segmentation [36]). Their implementations are retrained in the same multi-source pipeline.

To verify that the gains are not merely an artifact of the parameter-efficient tuning approach, we also include comparisons with SAM2-LoRA [37], SAM2-Adapter [38], and full fine-tuning.

We additionally evaluate soft-clDice on our backbone, neck, and sampler with all geometric heads disabled. This control isolates the effect of the auxiliary geometric branch from the topology-aware objective. Both models share the same 100-epoch schedule.

For component analysis, the Naive mixing variant uses proportional sampling and segmentation loss only, the Balanced sampling variant adds square-root sampling, and GARE further adds the geometric branch and soft-clDice. Naive mixing also uses a shorter schedule and a shorter warm-up, so the first transition is an incremental diagnostic rather than a strict one-variable causal comparison. Single-domain specialists are trained with the same backbone and optimiser. A separate geometric ablation compares subsets of the five targets.

4.3. Evaluation Metrics and Protocols

We report IoU, clDice, connectivity, and connected-component error (CC-err). IoU and clDice describe region and centreline overlap; connectivity and CC-err describe whether the predicted network has the same number of pieces as the reference. Because this metric compares only connected-component counts and cannot detect false bridges or incorrect path merges, we refer to it hereafter as CC-agreement (CC-agr.) rather than general connectivity; a path-level metric (APLS) is reported alongside it in the revised experiments (Section 5.5). CC-agreement is defined as

$$\text{CC-agr.} = 1 - \min\left(1, \frac{|CC_{\text{pred}} - CC_{\text{gt}}|}{\max(1, CC_{\text{gt}})}\right), \tag{9}$$

where CC denotes the number of eight-connected components, and CC-err is $|CC_{\text{pred}} - CC_{\text{gt}}|$.

To capture path-level structure beyond component counts, we report two additional metrics. The first is skeleton endpoint coverage. Let S_{gt} and S_{pred} be the skeletons of the reference and predicted masks (unit-width, eight-connected via Zhang–Suen thinning), and E_{gt}, E_{pred} the sets of degree-1 pixels (terminals) of each. With $N_\delta(v)$ denoting the δ-neighbourhood of a pixel ($\delta = 3\,\text{px}$),

$$\text{EP}_{\text{rec}} = \frac{|\{e \in E_{\text{gt}} : N_\delta(e) \cap S_{\text{pred}} \neq \varnothing\}|}{\max(1, |E_{\text{gt}}|)}, \quad \text{EP}_{\text{prec}} = \frac{|\{e \in E_{\text{pred}} : N_\delta(e) \cap S_{\text{gt}} \neq \varnothing\}|}{\max(1, |E_{\text{pred}}|)}, \tag{10}$$

and $\text{EP}_{\text{f1}} = 2\,\text{EP}_{\text{rec}}\,\text{EP}_{\text{prec}}/(\text{EP}_{\text{rec}} + \text{EP}_{\text{prec}})$. Low EP_{rec} indicates missed or broken connections, while low EP_{prec} flags false bridges or spurious extensions. The second is the Average Path Length Similarity (APLS) of Van Etten et al. [39]: both skeletons are converted to graphs, the shortest paths between randomly sampled node pairs are compared, and the normalised length ratios are averaged to produce a single score in $[0, 1]$; node pairs with no connecting path in either graph are counted as a length ratio of zero and included in the average, so APLS penalises disconnected pairs rather than ignoring them. APLS penalises both missing connections and detours, so it complements CC-agreement (which rewards only correct component counts) and endpoint coverage (which tests only terminal reachability).

Joint models are evaluated separately on each domain and on the sample-weighted union. Single-domain models provide specialist references. The annotation- efficiency experiment uses 10%, 25%, and 100% of DeepGlobe training images, with nested subsets, the same 500-image validation set, and approximately 57,000 optimiser steps at each level. The 50% level is not reported because one of its control runs was not completed.

To quantify metric stability, we detrend the plateau segment of each converged validation history and compute the epoch-to-epoch standard deviation. The reported 3σ values are used as a diagnostic indicator of epoch-to-epoch variation within a single run. They do

not replace proper significance testing: in the revised experiments we supplement them with paired bootstrap confidence intervals and permutation tests across multiple random seeds (Section 5.5). A difference inside the 3σ scale is noted as a likely tie, but formal significance is established by the bootstrap procedure.

Concretely, for each pair of configurations A, B evaluated on the same fixed validation set, let $d_i = m_A(\mathbf{x}_i) - m_B(\mathbf{x}_i)$ be the per-image metric difference at image i. We resample the paired differences with replacement $B = 10{,}000$ times to obtain a 95% confidence interval of the mean difference, and run a permutation test ($B = 5000$) to obtain a p-value. A difference is reported as significant (marked † in the tables) only when the confidence interval excludes zero and $p < 0.05$; otherwise the two configurations are described as not statistically distinguishable at the image-paired level. We pair by image rather than by run because the same validation image appears in both configurations: pairing eliminates the large inter-image variance of topology metrics, which would otherwise dwarf the treatment effect. Checkpoints are selected by validation IoU under an identical rule in every run, so only the random seed (42, 123, 2024) varies across the three repetitions. We report the mean ± standard deviation across seeds alongside the paired result to separate seed-level variability from image-level variability. The paired bootstrap is computed on the fixed validation set at the image level; the three-seed repetition is reported separately as seed-level variability and is not pooled into the bootstrap resampling.

All of the final topological results use the complete validation sets: 49 Massachusetts, 500 DeepGlobe, and 903 CHN6-CUG images.

5. Results

5.1. Measurement Scale of Topology Metrics

CC-agreement and CC-err are more sensitive to narrow gaps than region-overlap metrics. The following table reports the empirical 3σ floor used for interpretation. The floor is estimated separately by domain because the validation-set sizes differ.

CC-agreement fluctuates about eight times more than IoU in aggregate and is especially noisy on Massachusetts, whose validation set is the smallest. The thresholds are broadly consistent with an $n^{-1/2}$ scaling with validation set size, but the measured domain-specific values are used instead of extrapolated ones because several comparisons lie close to a threshold. Accordingly, small directional differences are not interpreted as findings. Checkpoints throughout this study are selected by validation IoU, which is stated with each table because the selection rule and the measurement scale have to be read together. The floor measures epoch-to-epoch variation within runs. For APLS and End. F1, which are computed on the skeleton and are not tracked in the epoch history, the floor is estimated from run-to-run variation across three independent training seeds (Section 5.5); the aggregate APLS floor of 0.0484 is the largest among the skeleton-derived metrics (CC-agreement 0.0260, End. F1 0.0061), reflecting the higher seed sensitivity of path-length similarity, while the overall largest aggregate floor is that of CC-err (0.1956).

Table 2 lists these floors for all six metrics, both in aggregate and for each domain. Two features govern how the entries should be read. First, the floors differ by more than an order of magnitude across metrics: IoU is the most stable (0.0033 in aggregate), CC-agreement is about eight times looser (0.0260), and CC-err is the loosest (0.1956), so the same absolute gap is negligible under one metric and decisive under another. Second, the floors scale with validation-set size, which is why Massachusetts, with only 49 evaluation images, shows the largest values for most metrics. Every improvement reported in this paper is therefore compared against the corresponding entry of this table rather than against a single global threshold.

Table 2. Empirical 3σ measurement floors estimated from linearly detrended plateau histories. These values are used to distinguish measurable effects from epoch-to-epoch variation; they serve as a diagnostic consistency check.

Metric	Aggregate	Massachusetts	DeepGlobe	CHN6-CUG
IoU	0.0033	0.0087	0.0043	0.0048
clDice	0.0067	0.0062	0.0040	0.0086
CC-agreement	0.0260	0.1060	0.0371	0.0262
CC-err	0.1956	2.5154	0.3708	0.0818
APLS	0.0484	0.0087	0.0503	0.0542
End. F1	0.0061	0.0079	0.0008	0.0106

5.2. Multi-Source Comparison

Table 3 reports the controlled comparison. Every model is retrained by us on the same three-source union, at the same input resolution, and evaluated on the same 1452 validation images, so the rows differ in architecture and supervision rather than in data or protocol. The baselines are grouped into road-specific and general segmentation networks, recent remote sensing models, SAM2 adaptations that share our backbone, and a topology-aware objective applied to our own backbone and neck.

Table 3. Multi-source road extraction results. Aggregate IoU is sample-weighted over the 1452 validation images, and the three IoU columns report the domain-level values. Domain-resolved APLS and End. F1 are given in Table 5. Models are evaluated with the same validation protocol; nominal best values are bolded, while interpretations use the measurement floors in Table 2. Arrows indicate the direction of improvement: an upward arrow (↑) means higher values are better, while a downward arrow (↓) means lower values are better.

Method	Mass. IoU	DeepGlobe IoU	CHN6 IoU	Overall IoU	clDice	CC-Agr.	CC-Err ↓	APLS ↑	End. F1 ↑
Road-specific and general segmentation networks									
D-LinkNet [1]	0.5918	0.6502	0.7585	0.7156	0.8466	0.4762	4.0172	0.6088	0.7270
DeepLabV3+ [26]	0.5843	0.6553	0.7550	0.7149	0.8459	0.4765	4.5847	0.6085	0.7275
SegFormer [34]	0.5862	0.6838	0.7479	0.7203	0.8519	0.5436	3.6116	0.6193	0.7364
Recent remote sensing segmentation models									
SFA-Net [35]	0.5855	0.6739	0.7421	0.7133	0.8459	0.5445	3.2225	0.6156	0.7275
D2LS [36]	0.5977	0.7116	0.7695	0.7438	0.8595	0.3752	16.6667	0.6007	0.7368
SAM2 adaptation baselines									
SAM2-LoRA [37]	0.6026	0.6569	0.7638	0.7215	0.8554	0.5635	2.6866	0.6215	0.7381
SAM2-Adapter [38]	0.6078	0.6685	0.7714	0.7304	0.8612	0.6263	1.8671	0.6366	0.7515
SAM2-FullFT	0.5962	0.6182	0.7246	0.6836	0.8133	0.5296	2.8354	0.5995	0.7087
Topology-aware objective control									
soft-clDice [17]	0.6317	0.6955	0.7749	0.7427	0.8798	0.6673	1.2390	0.5835	0.7460
Ours	0.6249	0.6914	0.7737	0.7403	0.8780	0.6874	1.0861	0.5812	0.7441

Against D-LinkNet, aggregate CC-agreement rises from 0.4762 to 0.6874 and CC-err falls from 4.0172 to 1.0861. Against SAM2-Adapter, which shares our encoder, CC-agreement rises from 0.6263 to 0.6874 and CC-err falls from 1.8671 to 1.0861. Both CC-agreement margins, 0.2112 and 0.0611, are 8.1 and 2.4 times the aggregate floor of 0.0260, so they are resolvable by a wide margin. Aggregate IoU moves far less: our 0.7403 is 0.0035 below D2LS, at the IoU floor of 0.0033; the paired bootstrap confidence interval includes zero ($p = 0.686$), so the two are statistically indistinguishable. The framework therefore separates from the baselines on topology and not on region overlap, which is the behaviour the geometric supervision and the centreline objective are designed to produce.

The comparison against SAM2-Adapter is the most informative one, because it holds the foundation-model capacity fixed and varies only what we add to it. A margin of 0.0611 in CC-agreement and 0.78 in CC-err under that constraint is attributable to the neck, the sampling, and the geometric branch together; Section 5.5 separates the branch from the rest.

D2LS is the clearest case for reporting both metric families. It attains the highest DeepGlobe IoU in the table, 0.7116 against our 0.6914, yet its CC-agreement on that domain is 0.2606 and its CC-err is 12.92, against 1.7480 for our model on the same domain; on CHN6-CUG its IoU of 0.7695 comes with a CC-err of 17.90, against 0.5659 for our model on the same domain. A reader ranking by IoU alone would select D2LS on DeepGlobe and receive a mask that recovers the road surface while cutting the network into an order of magnitude more pieces than the reference. The floors cannot explain a gap of this size: the DeepGlobe CC-err floor is 0.3708, and the gap between 12.92 and our 1.7480 on the same domain is more than thirty times it. Our model still leaves breaks under severe occlusion, but by a measure that reflects component-count agreement the two outputs are not close. Figure 5 shows that this is a general property of the comparison rather than one model's quirk: the horizontal and vertical positions are close to independent across all methods and all three domains.

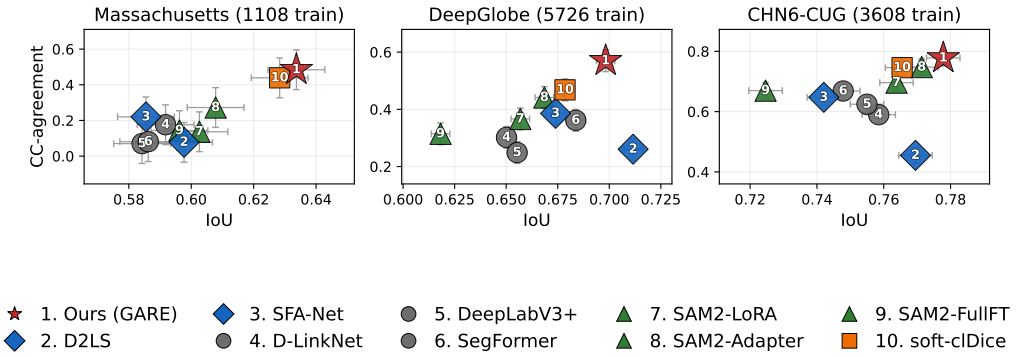


Figure 5. IoU against CC-agreement for the controlled multi-source comparison, taken from Table 3. Error bars are the domain-specific measurement floors. Models that are close on the horizontal axis can be far apart on the vertical one, which is the practical reason for reporting both families.

The paired bootstrap significance tests of Table 4 confirm that the topology gains over the region-overlap baselines are significant at the image-paired level, while the aggregate IoU margin over D2LS (-0.003) is not: its 95% CI includes zero. Against the soft-clDice control no metric is significant, as expected, since the control and our model share the clDice term and differ only in the geometric heads.

APLS, however, moves the other way: our aggregate APLS of 0.5812 is below every baseline (D-LinkNet 0.6088, SAM2-Adapter 0.6366), and the bootstrap tests mark most of those differences significant. The cause is the soft-clDice objective rather than the geometric branch: in the 2×2 ablation of Section 5.5, adding the clDice term drops aggregate APLS by 0.0644 (0.6479 to 0.5835), whereas adding the geometric supervision drops it by only 0.0041 (0.6479 to 0.6438). clDice optimises the topology of the skeleton, and a mask that is better connected can trade away the precision of the path it draws. The trade is not an artefact of our pipeline: CC-agreement and CC-err reward component-count agreement and give full marks to a false bridge, a wrong-shape component, or half the road, so they measure component-count agreement rather than path-level traversability, while APLS measures how accurately the path is drawn, and a method can improve one at the cost of the other.

Table 4. Paired image-level bootstrap significance tests for the controlled comparison. Δ is the aggregate difference (ours minus the listed method); † marks a 95% bootstrap CI that excludes zero. A downward arrow (↓) means lower values are better and the bold entries denote values that are both statistically significant and represent an improvement over the comparison.

Comparison	IoU	clDice	CC-Agr.	CC-Err ↓	APLS	End. F1
			Δ (Ours − Other) ± 95% CI			
Ours vs. D-LinkNet	**+0.025 ± 0.007** †	**+0.031 ± 0.008** †	**+0.211 ± 0.023** †	**−2.931 ± 0.332** †	**−0.028 ± 0.009** †	**+0.017 ± 0.008** †
Ours vs. DeepLabV3+	**+0.025 ± 0.007** †	**+0.032 ± 0.007** †	**+0.211 ± 0.023** †	**−3.499 ± 0.430** †	**−0.027 ± 0.009** †	**+0.017 ± 0.008** †
Ours vs. SegFormer	**+0.020 ± 0.007** †	**+0.026 ± 0.008** †	**+0.144 ± 0.022** †	**−2.525 ± 0.417** †	**−0.038 ± 0.010** †	+0.008 ± 0.008
Ours vs. SFA-Net	**+0.027 ± 0.007** †	**+0.032 ± 0.007** †	**+0.143 ± 0.023** †	**−2.136 ± 0.445** †	**−0.034 ± 0.009** †	**+0.017 ± 0.007** †
Ours vs. D2LS	−0.003 ± 0.006	**+0.018 ± 0.007** †	**+0.312 ± 0.023** †	**−15.581 ± 1.429** †	**−0.020 ± 0.007** †	**+0.007 ± 0.006** †
Ours vs. SAM2-LoRA	**+0.019 ± 0.006** †	**+0.023 ± 0.007** †	**+0.124 ± 0.020** †	**−1.601 ± 0.269** †	**−0.040 ± 0.009** †	+0.006 ± 0.007
Ours vs. SAM2-Adapter	+0.010 ± 0.006	**+0.017 ± 0.007** †	**+0.061 ± 0.020** †	**−0.781 ± 0.167** †	**−0.055 ± 0.009** †	−0.007 ± 0.007
Ours vs. SAM2-FullFT	**+0.057 ± 0.008** †	**+0.065 ± 0.008** †	**+0.158 ± 0.022** †	**−1.749 ± 0.270** †	**−0.018 ± 0.009** †	**+0.035 ± 0.008** †
Ours vs. soft-clDice	−0.002 ± 0.005	−0.002 ± 0.005	+0.020 ± 0.018	−0.153 ± 0.083	−0.002 ± 0.005	−0.002 ± 0.005
Ours vs. geo-only	+0.005 ± 0.005	+0.013 ± 0.006	**+0.036 ± 0.019** †	**−0.357 ± 0.102** †	**−0.063 ± 0.009** †	**−0.010 ± 0.007** †
Ours vs. no-geo	−0.001 ± 0.005	+0.009 ± 0.006	+0.023 ± 0.020	**−0.367 ± 0.121** †	**−0.067 ± 0.009** †	**−0.013 ± 0.007** †

The domain-resolved APLS of Table 5 localises the trade: our model is closest to the source-rich Massachusetts domain, where the clDice-driven trade is smallest, and furthest on CHN6-CUG, where the source is smallest and the trade is largest. Table 6 makes the complementarity of the two metric families explicit on eight synthetic cases built from a single validation tile, with D1 as the identity case that scores one on every metric. In four cases (D2, D3, D5, D6) the component count is broken while the skeleton geometry survives: CC-agreement falls to zero and CC-err rises to one, yet clDice and APLS both remain above 0.88. In the remaining three (D4, D7, D8) the count is preserved—CC-agreement 1.0, CC-err 0—while the geometry degrades progressively, down to clDice 0.05 and APLS 0.00 in D8. Neither family therefore subsumes the other, which is why both are reported throughout this paper.

Table 5. Domain-resolved APLS and End. F1 for the controlled comparison. Bold marks the domain best.

Method	APLS	End.F1	APLS	End.F1	APLS	End.F1
	Mass.		DeepGlobe		CHN6	
D-LinkNet	0.8003	0.9041	0.7060	0.8345	0.5445	0.6578
DeepLabV3+	0.7850	0.9058	0.7030	0.8363	0.5466	0.6576
SegFormer	0.7903	**0.9091**	**0.7363**	0.8659	0.5452	0.6553
SFA-Net	0.7901	0.8991	0.7284	0.8546	0.5437	0.6478
D2LS	0.7833	0.9007	0.7297	**0.8810**	0.5194	0.6481
SAM2-LoRA	0.8172	0.9062	0.7134	0.8465	0.5600	0.6690
SAM2-Adapter	0.8191	0.9071	0.7317	0.8543	**0.5740**	**0.6862**
SAM2-FullFT	0.8005	0.8985	0.6662	0.8047	0.5516	0.6453
soft-clDice	**0.8335**	0.9069	0.6971	0.8757	0.5070	0.6654
Ours	0.8282	0.8993	0.6998	0.8713	0.5021	0.6653

Table 6. Synthetic diagnostic cases showing where structural metrics give full marks while geometric metrics do not. D1 is the identity; the remaining cases fall into two groups: those that break the component count but keep the geometry (D2, D3, D5, D6), and those that keep the count but break the geometry (D4, D7, D8).

Case	CC-Agr.	CC-Err	clDice	APLS
D1 identity	1.0	0	1.0	1.0
D2 main-line break	0	1	0.9759	0.9318
D3 false bridge	0	1	0.8923	0.8889
D4 false bridge/loop	1.0	0	0.7421	0.7179
D5 over-extension	0	1	0.9677	0.967
D6 wrong geometry	0	1	0.9634	0.9091
D7 half predicted	1.0	0	0.6672	0.6667
D8 wrong shape	1.0	0	0.0519	0.0

5.3. Domain-Resolved Effects of Joint Training

Table 7 isolates the effect of training on the union rather than on one source, by comparing the joint model with three specialists that share its backbone, resolution, and optimiser. Ordering the rows by training-set size makes the pattern visible directly.

Table 7. Joint multi-source training versus single-domain specialists. Both models use the same backbone, resolution, and optimiser. A difference is interpretable only when it exceeds the domain-specific floor in Table 2.

Domain	Train	Single-Domain		Joint		Δ
		IoU	CC-Agr.	IoU	CC-Agr.	
Massachusetts	1108	0.6069	0.2199	0.6337	0.4844	+0.2645
CHN6-CUG	3608	0.7752	0.7284	0.7778	0.7780	+0.0496
DeepGlobe	5726	0.7014	0.5475	0.6983	0.5698	+0.0223

The connectivity gain scales inversely with the size of the source: 0.2645 on Massachusetts, which is 2.5 times that domain's floor; 0.0496 on CHN6-CUG, which is 1.9 times its floor; and 0.0223 on DeepGlobe, which is inside it. Aggregate IoU moves little in every case, so the union supplies structural information that a small source cannot expose on its own, without changing how much road area is recovered. Because the three datasets differ in acquisition difficulty as well as in size, the ordering identifies where the benefit appears but does not attribute it to volume alone.

A separate comparison shows that a DeepGlobe specialist evaluated on Massachusetts reaches connectivity 0.3822, compared with 0.2199 for the Massachusetts specialist, while its IoU is lower. This result is consistent with the view that road structure is more transferable across domains than surface appearance.

5.4. Ablation of Sampling and Geometric Supervision

Table 8 separates the two training-time interventions by adding them one at a time to a joint model that has neither. Because the naive-mixing run was trained on a shorter schedule than the other two, the first transition varies schedule as well as sampling and is read as a diagnostic rather than as an isolated causal estimate.

Table 8. Incremental ablation. Naive mixing uses proportional sampling and segmentation loss only; Balanced sampling adds square-root sampling; GARE adds the geometric branch and soft-clDice. A downward arrow (↓) means lower values are better and bold entries denote the best result in each column.

Variant	Overall IoU	Mass. IoU	Mass. CC-Agr.	clDice	CC-Agr.	CC-Err ↓
Naive mixing	0.7430	0.6026	0.3075	**0.8596**	0.6716	1.33
Balanced sampling	0.7429	0.6190	0.4699	0.8564	0.6769	1.27
GARE	**0.7456**	**0.6337**	**0.4844**	0.8588	**0.6964**	**1.08**

The two interventions act on different scales. Square-root sampling raises Massachusetts connectivity by 0.1624, above that domain's floor of 0.1060, which is the single largest step in the chain and comes at no cost beyond changing the sampling probabilities. The geometric branch adds a further 0.0145 there, inside the floor, together with the best aggregate IoU, connectivity, and CC-err in the table, none of which separates from the middle row by more than its floor. Sampling is therefore the only intervention that resolves in this comparison, and the branch is retained on the evidence of the annotation-volume experiment below.

The GARE row is the original 300-epoch main model, retained for continuity within this incremental chain; it is not on the unified 100-epoch schedule of Table 9, so its absolute values are not directly comparable to those in Table 3.

Table 9. Geometric supervision under a unified 100-epoch schedule. Only the auxiliary heads and their loss weights vary. The last four rows form a two-by-two design over the two task pairs on top of skeleton and soft-clDice: neither pair, affinity and endpoint only, direction and distance only, and both pairs. The indented rows are therefore alternatives to each other, not a cumulative sequence. A downward arrow (↓) means lower values are better.

Geometric Supervision	Aggregate				CC-Agr. by Domain		
	IoU	clDice	CC-Agr.	CC-Err ↓	Mass.	DG	CHN6
None (segmentation only)	0.7439	0.8584	0.6729	1.30	0.3585	0.5340	0.7669
Skeleton head only	**0.7448**	0.8592	0.6714	1.38	0.3582	0.5239	0.7701
soft-clDice only	0.7401	0.8567	0.6744	1.25	0.3990	0.5214	0.7740
Skeleton + clDice	0.7416	0.8578	0.6880	1.14	0.4464	0.5521	0.7763
with affinity + endpoint	0.7416	0.8597	**0.6926**	1.15	**0.5009**	**0.5542**	0.7796
with direction + distance	0.7446	**0.8615**	0.6918	**1.13**	0.4646	0.5515	**0.7818**
All five	0.7428	0.8604	0.6825	1.15	0.4564	0.5394	0.7740
Spread, rows 2–7	0.0047	0.0048	0.0212	0.25	0.1427	0.0328	0.0117
3σ floor	0.0033	0.0067	0.0260	0.1956	0.1060	0.0371	0.0262

5.5. Integrated Geometric Prior

The unified 100-epoch ablation in Table 9 tests whether the five geometric tasks are individually necessary. Because the last four rows add the two task pairs on top of the same skeleton and soft-clDice baseline, the indented rows are alternatives to each other rather than a cumulative sequence; the spread row measures the range over rows 2 to 7, that is, over every configuration that adds any auxiliary supervision to the first row, and is compared with the floor beneath it.

The first row is the essential control: it disables all geometric supervision while keeping the encoder, neck, sampling, and schedule fixed, so the table can answer whether the branch helps at all rather than only which combination is best.

Across the six supervised configurations, the spread is 0.0047 in IoU and 0.0212 in aggregate connectivity, both at or below the corresponding floors, and the complete branch differs from the control by −0.0011 IoU, +0.0020 clDice, +0.0096 connectivity, and −0.15 CC-err. On the aggregate, no individual target is shown to be inferior. By domain the picture is different: the Massachusetts connectivity spread reaches 0.1427 against a floor of 0.1060, and two configurations exceed the control there by 0.1424 and 0.1061, whereas the DeepGlobe and CHN6-CUG spreads of 0.0328 and 0.0117 stay inside their floors. The branch is thus resolvable where the source is small and its internal composition is not, which is the reason we treat the five targets as one prior rather than ranking them.

We note that 'Ours' in Table 3 (IoU 0.7403, CC-agreement 0.6874) and "All five" in Table 9 (IoU 0.7428, CC-agreement 0.6825) share the same five-target configuration and the same unified 100-epoch training schedule but come from different training runs with different random seeds. The differences (0.0025 IoU, 0.0049 CC-agreement) are within the multi-seed confidence intervals reported in Section 5.5 and illustrate the run-to-run variability of topology metrics that motivates our bootstrap significance testing.

Because the aggregate effect of the branch sits inside the measurement floor while the clDice term does not, we separate the two with a clean 2 × 2 factorial ablation in Table 10: geometric supervision on/off crossed with the clDice term on/off, everything else fixed. The interaction is small in IoU (+0.0039) and clDice (+0.0016) but large and positive in connectivity (+0.0328) and in CC-err (−0.1426), so the point estimates are consistent with the two components reinforcing each other on topology rather than acting independently, and the clDice term is the dominant contributor; we note that the interaction is a point estimate and we do not claim a statistically significant synergy. Geometric supervision alone slightly lowers CC-agreement (0.6640 to 0.6513); only when combined with the soft-clDice term does the configuration reach 0.6874. The improvement over soft-clDice alone (0.6673 to 0.6874, +0.020) is not statistically significant under the paired bootstrap (95% CI includes zero). We therefore state the contribution of the geometric branch as an improvement in the combined configuration rather than as an independently significant gain.

Table 10. Clean 2 × 2 factorial ablation of the geometric branch (on/off) crossed with the soft-clDice term (on/off), with the encoder, neck, sampling, and schedule fixed. The interaction row is 4 − 2 − 3 + 1; a positive interaction in connectivity and a negative one in CC-err mean the two components reinforce each other on topology. The APLS column shows the cost of that reinforcement: the clDice term accounts for almost all of the APLS loss (0.6479 to 0.5835), while the geometric branch accounts for almost none (0.6479 to 0.6438). Arrows indicate the direction of improvement: an upward arrow (↑) means higher values are better, while a downward arrow (↓) means lower values are better.

Geometric Supervision	IoU	clDice	CC-Agr.	CC-Err ↓	APLS ↑
① None	0.7412	0.8687	0.6640	1.4532	0.6479
② Geo only	0.7349	0.8653	0.6513	1.4428	0.6438
③ clDice only	0.7427	0.8798	0.6673	1.2390	0.5835
④ Geo + clDice	0.7403	0.8780	0.6874	1.0861	0.5812
Interaction = ④ − ② − ③ + ①	+0.0039	+0.0016	+0.0328	−0.1426	+0.0018

Table 11 decomposes the connectivity comparison by domain. Our model leads every domain in connectivity, but the margin over SAM2-Adapter concentrates in Massachusetts (0.5227 vs. 0.2725) and DeepGlobe (0.5679 vs. 0.4415), where the source is small; on CHN6-CUG the margin is thin (0.7626 vs. 0.7478). CC-err follows the same pattern. This mirrors the joint-versus-single result: the benefit of structural supervision is largest where the source cannot supply the structure on its own.

Table 11. Domain-resolved connectivity and CC-err for the controlled comparison. Our model leads every domain in connectivity and trails none in CC-err.

Method	CC-Agr.	CC-Err	CC-Agr.	CC-Err	CC-Agr.	CC-Err
	Mass.		DeepGlobe		CHN6	
D-LinkNet	0.1761	15.2857	0.3016	5.8880	0.5892	2.3699
DeepLabV3+	0.0702	30.5918	0.2489	7.6020	0.6245	1.5028
SegFormer	0.0810	30.5510	0.3617	5.6580	0.6695	1.0166
SFA-Net	0.2179	23.8776	0.3888	4.8740	0.6484	1.1872
D2LS	0.0774	32.1224	0.2606	12.9220	0.4549	17.9014
SAM2-LoRA	0.1366	13.0204	0.3657	4.8440	0.6962	0.9313
SAM2-Adapter	0.2725	8.2857	0.4415	3.3860	0.7478	0.6777
SAM2-FullFT	0.1434	13.2857	0.3144	5.1600	0.6696	0.9812
soft-clDice	0.4211	4.8776	0.5394	2.0020	0.7515	0.6190
Ours	0.5227	3.9184	0.5679	1.7480	0.7626	0.5659

5.6. Effect of the Geometric Branch Against Annotation Volume

The fixed-domain experiment on DeepGlobe varies only the amount of training annotation. The 10%, 25%, and 100% subsets are nested, and the same 500-image hold-out is excluded before subsampling and used for every evaluation. The epoch budgets are scaled inversely with the training fraction (400, 160, and 40 epochs), so each level receives approximately 57,000 optimiser steps.

Table 12 reports the results. This is the cleanest test of the branch in the paper, because domain, validation set, and step budget are all held fixed and the pairs differ in one variable. The gain is positive at every level and declines monotonically as annotation grows, and it clears the DeepGlobe floor only at 10%. Two things follow. The branch is worth its training cost when annotation is scarce, which is the regime that limits operational mapping. And the decline is a statement about our resolution, not a demonstration that the effect reaches zero: at full annotation the remaining 0.0100 is simply smaller than what this validation set can distinguish. Figure 6 draws both curves against the floor.

Table 12. Effect of the geometric branch at three annotation levels on DeepGlobe. Each row is a matched pair of runs differing only in whether the branch is active.

Annotations	Images	CC-Agr.		Δ	Δ clDice
		w/o Geo	With Geo		
10%	573	0.4600	0.5079	+0.0479	−0.0022
25%	1432	0.4985	0.5183	+0.0198	+0.0048
100%	5726	0.5437	0.5537	+0.0100	+0.0042

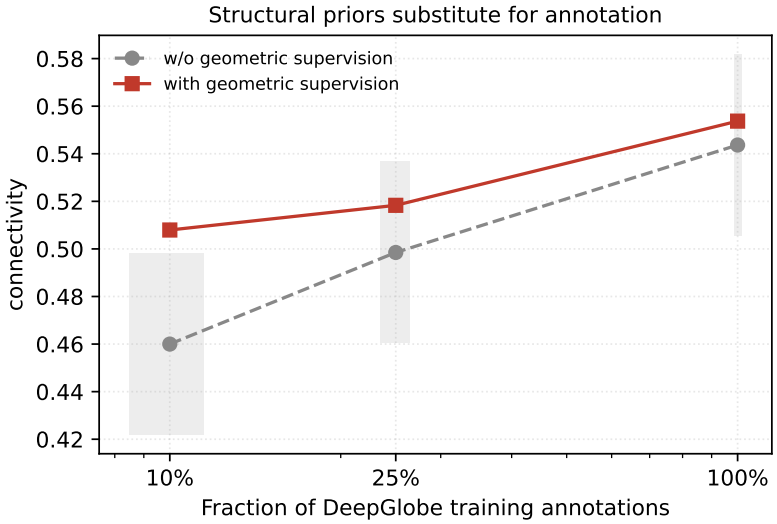


Figure 6. DeepGlobe connectivity at three annotation levels with and without geometric supervision, with the domain measurement floor of 0.0371 shown for reference. The gap between the two curves narrows as annotation grows, from 0.0479 at 10% to 0.0100 at full annotation.

5.7. Computational Cost

Table 13 separates the cost of the geometric branch from the cost of the backbone it attaches to. The two parameter columns differ only by the auxiliary heads, which are absent from the deployed graph, so the gap between them is the entire deployment cost of the branch.

Table 13. Computational cost at 1024 × 1024 and batch size 1. The auxiliary heads are disabled during inference.

Method	Train (M)	Infer. (M)	GMACs	Mem. (GB)
SFA-Net	10.70	10.70	28.8	0.37
DeepLabV3+	26.68	26.68	146.9	0.60
SegFormer	24.72	24.72	147.6	0.83
D-LinkNet	31.10	31.10	106.5	0.51
D2LS	94.02	94.02	373.8	4.88
Ours	74.65	73.92	893.3	1.49

The proposed model has higher arithmetic cost than the lightweight baselines, which is mainly due to the SAM2 encoder and wide feature-enhancement neck. The auxiliary heads account for 0.73 M parameters, about 1% of the training model, and do not add inference operations. Peak inference memory is 1.49 GB, lower than the 4.88 GB reported for D2LS under the same input protocol.

5.8. Qualitative Analysis

Figure 7 shows representative failure cases. The baselines often recover the road surface but break the route at canopy, shadow, or building occlusions. The proposed model retains more connections in these examples, but some breaks remain. This is expected: when an occlusion removes both local appearance and reliable neighbouring evidence, a mask-derived prior can encourage continuity but cannot guarantee recovery of an unobserved road.



Figure 7. Road-extraction examples on DeepGlobe. Columns show the input, the reference mask, the proposed prediction, and the baseline predictions. Red marks missed reference road and blue marks false positives. Fragmentation is reduced but not eliminated.

The same failure mode is more pronounced on Massachusetts, whose sparse rural network makes each interruption costly. Counted over that validation set, the proposed model predicts an average of 10.0 connected components per tile against 21.3 for D-LinkNet and 36.7 for DeepLabV3+. The comparison is consistent with the aggregate CC-err column of Table 3.

Figure 8 applies the same visualisation within the geometric branch, comparing the complete five-task configuration, skeleton plus soft-clDice, and non-geometric anchor under the unified 100-epoch schedule. The aggregate difference between these configurations does not clear the measurement floor, so the examples in Figure 8 illustrate the mechanism the branch is intended to supply rather than evidence that the configurations are separable; the quantitative comparison is Table 9.

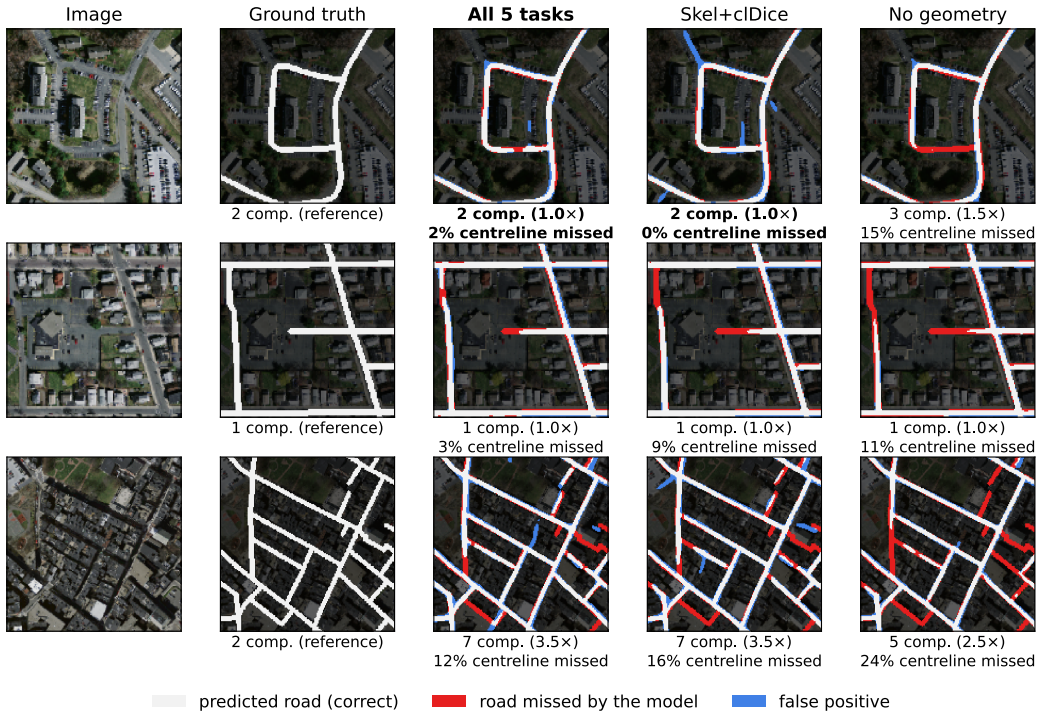


Figure 8. Qualitative comparison of the full geometric branch, skeleton plus soft-clDice, and the non-geometric anchor under the unified 100-epoch schedule. The three columns share the same schedule, so the comparison is not confounded by training length.

6. Discussion

6.1. When Geometry-Aware Supervision Helps

Two experiments with different designs agree on where the benefit lies. In the multi-source setting, joint training raises Massachusetts connectivity by 0.2645, well above that domain's floor of 0.1060, while the largest source shows no measurable gain. In the single-domain setting, where sampling and domain are held fixed, the geometric branch clears the floor only at 10% annotation. Both point to the same mechanism: structural supervision substitutes for regularities that a small sample does not expose reliably, and its marginal effect shrinks below our resolution once the sample is large. The distinction matters for how the result is used. We can say that the effect is detectable when annotation is scarce; we cannot say that it vanishes when annotation is plentiful, only that any remaining effect is smaller than 0.0371 on this domain.

6.2. Connectivity Requires More than IoU

The controlled comparison demonstrates why road extraction should not be ranked by IoU alone. D2LS has the best DeepGlobe IoU but much worse connectivity and CC-err than the proposed model. Similar differences occur among the other baselines, although their IoU values are close. A road mask can therefore be regionally accurate and still fail as a network.

Our method addresses this mismatch in three ways. Skeleton and affinity targets provide explicit signals for centreline continuity and local linkage. The direction and distance targets encourage feature representations that follow elongated, width-varying structures rather than isolated texture patches. Soft-clDice adds a direct centreline objective, and square-root sampling keeps the difficult small source visible during optimisation. These mechanisms mitigate fragmentation in the reported metrics and examples, but they do not guarantee a gap-free network. Severe canopy or building occlusion remains a failure case, and the geometric targets are derived from visible reference masks rather than from

hidden roads. The appropriate claim is therefore connectivity improvement under the tested conditions, not complete topology preservation.

6.3. The Auxiliary Branch as an Integrated Prior

The geometric ablation gives a clear answer to a question that is rarely asked directly: the individual targets are not separable at this measurement precision. Best aggregate IoU, best clDice, and best connectivity fall into three different subsets, and the spread across all six supervised configurations is 0.0047 in IoU and 0.0212 in connectivity, at or below the corresponding floors. The informative comparison is instead against the first row, which disables geometric supervision entirely, and it resolves on one domain only: on Massachusetts two configurations exceed that domain's floor of 0.1060, by 0.1424 and 0.1061, whereas no configuration exceeds the aggregate floor of 0.0260. The branch therefore acts as a single structural bias whose components are largely interchangeable at this precision, and its effect is detectable where the domain is small. For a practitioner this is more useful than a ranking of five targets, because it says the branch can be shortened without much loss and that the domain, not the target set, determines whether it is worth running at all.

We retain all five targets because together they describe road geometry completely and cost nothing at inference, while acknowledging that the ablation licenses a shorter branch: a practitioner constrained by training memory can drop targets and expect a penalty inside the floors reported here.

6.4. Scope Beyond Roads

The construction depends on the mask, not on the class. Any feature whose mask has a medial axis, a local orientation, a local width, and graph-like connectivity admits the same five targets from the same deterministic operators, which cover rivers, canals, pipelines, field boundaries, and coastlines. We have not tested any of them. What our results support is narrower and worth stating precisely: the targets are computable for such features, the training-only arrangement transfers unchanged, and the argument for measuring connectivity separately from region overlap applies wherever the feature is a network. Whether the effect sizes carry over is an open question, and the answer may differ by feature: rivers branch but rarely reconnect, whereas road networks contain cycles, so the endpoint and affinity targets may behave differently there.

6.5. Limitations

This study has four main limitations.

First, a mask-derived prior cannot recover evidence that the image does not contain. Where canopy or buildings remove both the local appearance of the road and the neighbouring context that would let it be inferred, the geometric targets can encourage continuity but cannot supply the missing observation. The residual breaks in Figure 7 are of this kind, and they place a ceiling on what any purely mask-supervised approach can achieve under heavy occlusion; closing that gap requires additional evidence, such as multi-temporal imagery or an explicit graph completion step, rather than a better target.

Second, all five targets are functions of the reference mask, so their quality is bounded by the quality of that mask. Coarsely traced annotation yields a skeleton that wanders off the true centreline and a distance transform that misstates width, and the network is then supervised towards those errors. The distance-transform target is the most exposed in this respect, because it is least reliable exactly where road width changes abruptly, at junctions and at transitions between road classes. The approach therefore suits datasets whose masks are geometrically faithful and is not a way to extract structure from loose labels.

Third, the framework buys connectivity at a computational price. At 1024×1024 it requires 893.3 GMACs against 106.5 for D-LinkNet (Table 13), which is set by the SAM2

encoder and the wide neck rather than by the auxiliary branch. Removing the geometric heads costs nothing at deployment, but the backbone still places the model outside the envelope of on-board or near-real-time processing, and large-area mapping with it is a batch operation.

Fourth, the direction target is degenerate over most of the road pixels it supervises. Because **D** is built from the mask gradient, it vanishes in the road interior, so only pixels near a boundary carry orientation information and the head is trained on a target that is a constant vector over much of the road. A diagnostic on the training set confirms this: on average approximately 54% of road pixels receive a zero-gradient (constant) direction target, with per-image fractions ranging from 23% to 76% depending on road width. A skeleton-derived variant that assigns a tangent to every centreline pixel and restricts supervision to a ± 5 px valid-pixel band avoids this degeneracy entirely: the direction field is well-defined along the entire centreline rather than only at the boundary. In practice, the skeleton-derived variant yields a slightly lower aggregate CC-agreement (0.669 vs. 0.6874) but a higher Massachusetts endpoint F1 (0.6428 vs. 0.6354; see Table 1), suggesting that it improves path-endpoint reachability at a small cost in component-count agreement. Removing the direction head altogether drops Massachusetts endpoint F1 to 0.6079, confirming that direction supervision—even in its degenerate Sobel form—contributes meaningfully. The degeneracy therefore constrains the interpretability of the direction head rather than the reported gains; we retain the Sobel form in the main experiments because it requires no skeleton computation at training time and produces comparable aggregate connectivity, but the skeleton-derived form is a viable alternative when endpoint accuracy is prioritised.

7. Conclusions

We presented GARE (Geometry-Aware Road Extraction), a framework for road network extraction from remote sensing imagery that concentrates its added complexity in training while keeping the deployed network a single-head segmentation model. On the architecture side, we do not treat SAM2 as a fixed feature source: because its top-down pathway refines only the coarsest levels, we restore the missing top-down path with a feature-enhancement neck and add a directional attention module matched to the elongated geometry of roads, so that a foundation-model encoder becomes effective for dense thin-structure prediction. On the supervision side, five geometric descriptions of a road—centreline, tangent direction, local width, neighbourhood affinity, and endpoints—are computed deterministically from the segmentation mask that is already required for the task, supervised alongside it through a shared representation, and then removed. The branch adds no annotation, and its removal recovers 0.73 M parameters and removes all auxiliary computation from the deployed graph; the neck and directional attention module remain, as they serve the segmentation head.

The framework separates from the baselines on topology rather than on region overlap. Aggregate connectivity rises from 0.4762 for D-LinkNet to 0.6874, and the connected-component error falls from 4.0172 to 1.0861; against SAM2-Adapter, which shares our encoder, connectivity rises from 0.6263 to 0.6874. Those margins are 8.1 and 2.4 times the measurement scale we establish for the metric, whereas the aggregate IoU margin over the strongest baseline is at its scale. Joint training over the union raises Massachusetts connectivity by 0.2645, two and a half times that domain's scale, and the geometric branch clears the scale at 10% of DeepGlobe annotation. The benefit concentrates where annotation is scarce, which is the regime that limits operational mapping, and since the branch costs nothing to deploy there is no reason to withhold it elsewhere.

The construction depends on the geometry of the mask rather than on the road class, so it applies directly to rivers, canals, pipelines, and field boundaries. The endpoint and

affinity heads already predict the primitives a graph needs, and retaining them at inference—rather than discarding them as in the current design—would let the model emit network objects rather than binary masks, which is the representation that routing, accessibility analysis, and change detection actually consume.

Author Contributions: Conceptualization, Z.J. and L.X.; methodology, Z.J.; software, J.L.; validation, Y.Z. and K.Y.; formal analysis, Z.J.; investigation, L.X.; resources, L.X.; data curation, Z.J.; writing—original draft preparation, Z.J.; writing—review and editing, Z.J. and L.X.; visualization, L.X. and Y.Z.; supervision, Y.Z.; project administration, K.Y. and Y.Z.; funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2023YFB2603700.

Data Availability Statement: The three datasets used in this study are publicly available: Massachusetts Roads (https://www.cs.toronto.edu/~vmnih/data/) (accessed on 17 June 2026), DeepGlobe Road Extraction (https://competitions.codalab.org/competitions/18467) (accessed on 17 June 2026), and CHN6-CUG (https://grzy.cug.edu.cn/zhuqiqi) (accessed on 17 June 2026). All geometric auxiliary targets are derived deterministically from the released masks and require no additional annotation. Code and configuration files supporting the reported results are available from the corresponding author on reasonable request.

Acknowledgments: The authors thank the creators of the Massachusetts Roads, DeepGlobe, and CHN6-CUG datasets for enabling reproducible research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASPP	Atrous spatial pyramid pooling
CC	Connected component
clDice	Centreline Dice
FPN	Feature pyramid network
GMAC	Giga multiply-accumulate operations
IoU	Intersection over union
LoRA	Low-rank adaptation
SAM	Segment anything model

References

1. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–1924. https://doi.org/10.1109/CVPRW.2018.00034.
2. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. https://doi.org/10.1016/j.jag.2022.102987.
3. Tao, J.; Chen, Z.; Sun, Z.; Guo, H.; Leng, B.; Yu, Z.; Wang, Y.; He, Z.; Lei, X.; Yang, J. Seg-Road: A segmentation network for road extraction based on transformer and CNN with connectivity structures. *Remote Sens.* **2023**, *15*, 1602. https://doi.org/10.3390/rs15061602.
4. Zhao, L.; Ye, L.; Zhang, M.; Jiang, H.; Yang, Z.; Yang, M. DPSDA-Net: Dual-path convolutional neural network with strip dilated attention module for road extraction from high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 3741. https://doi.org/10.3390/rs15153741.
5. Mei, J.; Li, R.-J.; Gao, W.; Cheng, M.-M. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. https://doi.org/10.1109/TIP.2021.3117076.

6. Abdollahi, A.; Pradhan, B.; Alamri, A. SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5617815. https://doi.org/10.1109/TGRS.2022.3143855.
7. Zhang, L.; Yuan, S.; Dong, R.; Zheng, J.; Gan, B.; Fang, D.; Liu, Y.; Fu, H. SWCARE: Switchable learning and connectivity-aware refinement method for multi-city and diverse-scenario road mapping using remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103665. https://doi.org/10.1016/j.jag.2024.103665.
8. Liu, G.; Shan, Z.; Meng, Y.; Akbar, T.A.; Ye, S. RDPGNet: A road extraction network with dual-view information perception based on GCN. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102009. https://doi.org/10.1016/j.jksuci.2024.102009.
9. Li, W.; Lan, T.; Fan, S.; Jiang, Y. RBFNet: A region-aware and boundary-enhanced fusion network for road extraction from high-resolution remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 16608–16624. https://doi.org/10.1109/JSTARS.2024.3433552.
10. Zhou, G.; He, C.; Wang, H.; Xie, Q.; Chen, Q.; Hong, L.; Chen, J. RIRNet: A direction-guided post-processing network for road information reasoning. *Remote Sens.* **2024**, *16*, 2666. https://doi.org/10.3390/rs16142666.
11. Wang, X.; Jin, X.; Dai, Z.; Wu, Y.; Chehri, A. Deep learning-based methods for road extraction from remote sensing images: A vision, survey, and future directions. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 55–78. https://doi.org/10.1109/MGRS.2024.3491014.
12. Li, L.; Yi, J.; Lei, P.; Lou, H.; Li, X.; Lin, H. Rural road extraction from remote sensing images based on multi-view contextual information and multi-stage features. *Appl. Soft Comput.* **2025**, 113975. https://doi.org/10.1016/j.asoc.2025.113975.
13. Yang, Z.; Yao, H.; Li, Q.; Ni, W.; Wu, J.; Wang, Q. Semantic–spatial feature refinement network for road extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–10. https://doi.org/10.1109/TGRS.2026.3665299.
14. Huang, B.; Lu, Y.; Yang, R.; Tao, Y.; Wang, S.; Shi, Y. HSN-Net: A hybrid segmentation neural network for high-resolution road extraction. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. https://doi.org/10.1109/LGRS.2025.3558511.
15. Li, L.; Yi, J.; Fan, H.; Lin, H. A lightweight semantic segmentation network based on self-attention mechanism and state space model for efficient urban scene segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3562185.
16. Wang, Z.; Yi, J.; Chen, A.; Chen, L.; Lin, H.; Xu, K. Accurate semantic segmentation of very high-resolution remote sensing images considering feature state sequences: From benchmark datasets to urban applications. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 824–840. https://doi.org/10.1016/j.isprsjprs.2025.01.017.
17. Shit, S.; Paetzold, J.C.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J.P.W.; Bauer, U.; Menze, B.H. clDice—A novel topology-preserving loss function for tubular structure segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 16555–16564. https://doi.org/10.1109/CVPR46437.2021.01629.
18. Kirchhoff, Y.; Rokuss, M.R.; Roy, S.; Kovacs, B.; Ulrich, C.; Wald, T.; Zenk, M.; Vollmuth, P.; Kleesiek, J.; Isensee, F.; et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In *Computer Vision—ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 15135, pp. 218–234. https://doi.org/10.1007/978-3-031-72980-5_13.
19. Feng, R.; Guo, Z.; Du, X.; Wu, T. SAM2-RoadNet: Topology-aware multi-scale road extraction from high-resolution remote sensing images. *Remote Sens.* **2026**, *18*, 913. https://doi.org/10.3390/rs18060913.
20. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment anything in images and videos. *arXiv* **2024**, arXiv:2408.00714.
21. Zhang, E.; Liu, J.; Cao, A.; Sun, Z.; Zhang, H.; Wang, H.; Sun, L.; Song, M. RS-SAM: Integrating multi-scale information for enhanced remote sensing image segmentation. In Proceedings of the Asian Conference on Computer Vision (ACCV), Hanoi, Vietnam, 8–12 December 2024; pp. 280–296. https://doi.org/10.1007/978-981-96-0966-6_17.
22. Zhang, J.; Li, Y.; Yang, X.; Jiang, R.; Zhang, L. RSAM-Seg: A SAM-based model with prior knowledge integration for remote sensing image semantic segmentation. *Remote Sens.* **2025**, *17*, 590. https://doi.org/10.3390/rs17040590.
23. Zhang, P.; Li, J.; Wang, C.; Niu, Y. SAM2MS: An efficient framework for HRSI road extraction powered by SAM2. *Remote Sens.* **2025**, *17*, 3181. https://doi.org/10.3390/rs17183181.
24. Zhang, T.Y.; Suen, C.Y. A fast parallel algorithm for thinning digital patterns. *Commun. ACM* **1984**, *27*, 236–239. https://doi.org/10.1145/357994.358023.
25. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. https://doi.org/10.1109/CVPR.2017.106.
26. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.

27. Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; van der Maaten, L. Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 181–196.
28. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 4th International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. https://doi.org/10.1109/3DV.2016.79.
29. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324.
30. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In Proceedings of the International Workshop on Machine Learning in Medical Imaging (MLMI), Quebec City, QC, Canada, 10 September 2017; pp. 379–386. https://doi.org/10.1007/978-3-319-67389-9_44.
31. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010; pp. 210–223. https://doi.org/10.1007/978-3-642-15567-3_16.
32. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A challenge to parse the Earth through satellite images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–17209. https://doi.org/10.1109/CVPRW.2018.00031.
33. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. https://doi.org/10.1016/j.isprsjprs.2021.03.016.
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34, pp. 12077–12090.
35. Hwang, G.; Jeong, J.; Lee, S.J. SFA-Net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sens.* **2024**, *16*, 3278. https://doi.org/10.3390/rs16173278.
36. Zou, X.; Li, Y.; Zhang, S.; Li, K.; Wang, S.; Tao, P.; Xing, J.; Lang, C. Dynamic dictionary learning for remote sensing image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–26 October 2025; pp. 22457–22466. https://doi.org/10.1109/ICCV51701.2025.02085.
37. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022. https://doi.org/10.48550/arXiv.2106.09685.
38. Chen, T.; Lu, A.; Zhu, L.; Ding, C.; Yu, C.; Ji, D.; Li, Z.; Sun, L.; Mao, P.; Zang, Y. SAM2-Adapter: Evaluating and adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv* **2024**, arXiv:2408.04579. https://doi.org/10.48550/arXiv.2408.04579.
39. Van Etten, A.; Lindenbaum, D.; Bacastow, T.M. SpaceNet: A remote sensing dataset and challenge series. *arXiv* **2019**, arXiv:1807.01232.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.