Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 24 May 2026
Revised 30 June 2026
Accepted 16 July 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 750–761

KEYWORDS:
semantic segmentation; satellite imagery; topology preservation; cross-resolution fusion; persistent homology; deep learning

Topology-Preserving Cross-Resolution Network for Semantic Segmentation of Complex Satellite Imagery

Dr. S. Kiran[1], Dr.T.S Jayadeva[2], Dr. P. R. Rajesh Kumar[3], Dr. G. Silpalatha[4]

[1]Professor, Dept of CSE YSREC of YVU, Email - rkirans125@gmail.com
[2]Professor, School of ECE Reva University Banglore, Email - jayadeva.ts@reva.edu.in
[3]Assistant Professor, Department of CSE University College of Engineering and Technology, Sri Krishnadevaraya University-515003, Email - prrajeshcse@skucet.ac.in
[4]Assistant Professor, Department of Electronics and Communication Engineering Ysrec of YVU, Email - gsilpalathayvuce@gmail.com

ABSTRACT: Semantic segmentation of very-high-resolution, VHR, satellite imagery is important for land-cover mapping, urban planning, and disaster recovery. However, existing deep networks have high pixel accuracy but fail on thin, elongated, connected structures, like roads, hedgerows, or building perimeters. Most pixel-wise loss functions do not penalize them, causing these topological defects that make the maps less useful. To overcome limitations of conventional pixel-wise losses, we propose a topology preserving, cross-resolution network (TP-CRNet) that combines a multi-branch encoder with two novel modules. Cross Resolution Fusion Module (CRFM) passes information between resolution branches using bidirectional attention, thus preserving fine detail and global context. Second, we use Boundary-Topology Head (BTH) to provide skeleton and persistent homology supervision through a composite loss that penalizes unnecessary disconnections and non-desired holes. The model is evaluated on the ISPRS Potsdam dataset, the LoveDA dataset and the DeepGlobe dataset using mIoU, overall accuracy, boundary F1, clDice, and Betti error. TP-CRNet produces large and consistent improvements of region accuracy and topological correctness compared to U-Net, DeepLabV3+, HRNet and UNetFormer, especially for narrow linear classes. The ablation study shows that CRFM and BTH both provide complementary benefits and the cross-resolution experiments show that the model has graceful degradation on different ground sampling distances.

1. INTRODUCTION

Satellite and aerial imagery provide coverage of most of the inhabited earth at sub-metre spatial resolution. Thematic maps of land cover from satellite and aerial imagery require assigning a semantic class to each image pixel. This task, called semantic segmentation, allows one to make accurate segmentation maps that can be used for urbanization monitoring, land use planning in agriculture, inventorying existing infrastructure, and rapid assessment of a flood or earthquake. Beyond the percentage of correctly labeled pixels, a road network that is 95% pixel-accurate is of little use if the network has been split into many fragments by the segmentation.

Deep CNNs have enabled segmentation tasks through encoder-decoder architectures like U-Net [1] and DeepLabV3+ [2], learning features at multiple levels of abstraction, progressively recovering spatial resolution through skip connections or atrous convolution. Pyramid pooling [3] and high-resolution representation learning [4] further improved the pattern of spatial context

information. Transformer-based networks, such as SegFormer [5], Swin Transformer [6] and UNetFormer [7], have introduced long-range attention, allowing for modeling long spatial context in remote-sensing scenes. Remote-sensing-specific architectures, such as FarSeg [8], ResUNet-a [9], and ABCNet [10] are able to handle the background-foreground class imbalance and efficiently predict over large tiles.

Two limitations of resolution handling still remain: the size of objects in satellite images may vary several orders of magnitude, from the size of a car to the size of a river basin. Networks that progressively downsample lose thin structures, while networks with retained high-res stream often combine parallel resolutions through simple addition or concatenation, which equally weighs all pixels and does not learn whether the local detail or global context is more important at a pixel, which brings to the second drawback. In contrast, pixel-wise losses such as cross-entropy or Dice-type losses [11] do not capture connectivity: for example, splitting a road has a small pixel cost but changes the number of connected components (which, in turn, changes the topology) in the prediction, and therefore should contribute more than the pixel cost. Topological losses based on persistent homology [12, 13] or skeleton overlap [14] were introduced to tackle such issues within medical image analysis, but are rarely found in remote sensing due to the diverse set of classes and the large tile sizes.

We address both limitations within the same architecture that we refer to as the Topology-Preserving Cross-Resolution Network (TP-CRNet). Our contributions are. (i) We propose a Cross-Resolution Fusion Module (CRFM) for bi-directional attention between each resolution branch, allowing the amount of detail and context injected at every position to be learned, not fixed a priori. (ii) We propose a Boundary-Topology Head (BTH) acting on the fused feature map, with a composite loss based on cross-entropy, Dice, clDice and a differentiable persistent-homology term generalized to multi-class remote sensing labels. (iii) We develop a unified bundled evaluation protocol to report region accuracy (mIoU, overall accuracy), boundary quality (boundary F1), and topological correctness (clDice, Betti error) together. (iv) We compare the proposed network with five widely used baselines in terms of accuracy and robustness on three public datasets and different ground sampling distances. All parts of the code are implemented in Python using the PyTorch framework, and enough detail is given for reproducibility.

2. LITERATURE REVIEW

2.1 Encoder–decoder and multi-scale segmentation networks

The encoder-decoder architecture with symmetric skip connections, pioneered by the U-Net [1], remains a strong baseline in remote sensing for combining the restoration of spatial resolution with low computation costs, and later work focused on improving the contextual information of the U-Net. While PSPNet [3] used the concatenation of pooling results at multiple scales, and DeepLabV3+ [2] employed atrous spatial pyramid pooling with simple decoders, these approaches still used a single downsampled stream, leading to the potential blurring of semantic boundaries. HRNet [4] departed from this trend, simply stacking parallel branches at different resolutions which exchange information at each stage, allowing for high-resolution information while remaining suitable for small objects in aerial imagery. However, fusion in HRNet is performed via fixed convolutions and summation, which treat every position equally and fail to adapt to local contexts.

2.2 Attention and transformer architectures in remote sensing

Attention mechanisms assign different weights to different features, and Attention U-Net [15] applied gating at skip connections to suppress irrelevant activations. Transformers were extended to global self-attention, but it remained intractable at high resolution. Swin Transformer [6] made attention tractable using shifted windows. SegFormer [5] used a lightweight decoder on top of a hierarchical transformer encoder. In remote sensing, UNetFormer [7] added a transformer-based decoder that modeled global and local context to a convolutional encoder, achieving state-of-the-art on ISPRS urban benchmarks. Inspired by ABCNet [10] and FarSeg [8] with their bilateral design and foreground-aware relation modeling respectively, adaptive, content-dependent feature mixing for segmentation from aerial images may be helpful due to the large class imbalance of small objects and large background. However, attention is usually applied in a single scale or between the encoder and decoder, and the problem of adaptive feature fusion across different resolution branches remains unsolved.

2.3 Loss functions and topological supervision

The choice of loss function is critical to the accuracy of the segmentation. Improvements over the Dice loss [11] can handle class imbalance. The boundary loss [16] approaches this problem by performing the loss calculation over the signed distance transform of the segmentation, achieving higher resolution near boundaries. ResUNet-a [9] added two more objectives to the Tanimoto loss: boundary prediction and distance prediction, which also help sharpen edges in remote-sensing images. However, none of these measure connectivity. Hu et al. [12] proposed the first differentiable topological loss based on persistent homology, involving the matching of the birth and death of the connected components and holes of the prediction and the ground truth. Clough et al. [13] generalized it with prior Betti numbers and applied it on cardiac and vascular images . Shit et al. [14] defined a soft skeleton-based shape measure called clDice, which is computationally efficient and directly stresses the centre-lines of tube-like structures. Theoretical aspects of these shape measures are described in the computational-topology literature [17].

2.4 Benchmarks and research gap

Public benchmarks for developing and testing methods for semantic segmentation with dense annotations include ISPRS Potsdam and ISPRS Vaihingen datasets [18], LoveDA [19] as well as DeepGlobe [20]. They cover urban and rural areas and are widely used for comparison of methods, though they usually only report pixel-wise mIoU or accuracy. Although topology-aware supervision has only been developed for binary medical structures, several remote sensing architectures have already been developed for multi-scale, attention-based feature extraction, but not connectivity. We propose a network that push connections together in any resolution in an adaptive way, a multi-class topological loss for training, and a set of metrics that expose connectivity errors that pixel-based metrics mask, closing the gap between the two kinds of works.

3. METHODOLOGY

3.1 Problem formulation

Let $X \in \mathbb{R}^{H\times W\times C}$ denote an input tile with C spectral bands and let $Y \in \{1,\ldots,K\}^{H\times W}$ denote the ground-truth label map with K classes. The network f_θ produces a probability tensor $P = f_\theta(X) \in [0,1]^{H\times W\times K}$, and the predicted label at pixel (i,j) is $\hat{y}_{ij} = \arg\max_k P_{ijk}$. The objective is to learn θ such that the prediction is accurate at the pixel level and also preserves the topology of each class region, expressed through the Betti numbers β_0 (number of connected components) and β_1 (number of holes).

3.2 Overall architecture

TP-CRNet (Figure 1) consists of a four-branch multi-resolution encoder, a Cross-Resolution Fusion Module (CRFM), a segmentation head, and a Boundary-Topology Head (BTH). The encoder is based on the parallel convolutional stacks in HRNet-W32 [4] with strides of 4, 8, 16 and 32 and 48, 96, 192 and 384 channels respectively. Unlike HRNet, in TP-CRNet the structure is not fused with fixed convolutions, and the last exchange stage is replaced with a CRFM which will be described below.

Natural Resources for Human Health
eISSN: 2583-1194

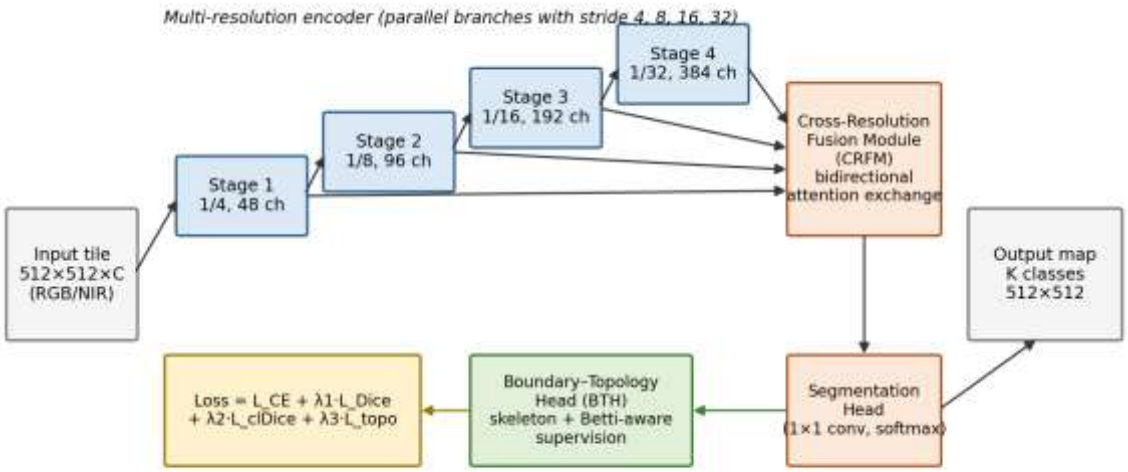

Figure 1. Overall architecture of the proposed TP-CRNet.

Figure 1. Overall architecture of the proposed TP-CRNet. The multi-resolution encoder feeds four branches into the CRFM; the fused representation is decoded by the segmentation head and supervised by the BTH through a composite loss.

3.3 Cross-Resolution Fusion Module

Let $F_s \in \mathbb{R}^{H_s \times W_s \times C_s}$ be the feature map of branch $s \in \{1,2,3,4\}$. CRFM first projects every branch to a common channel width $d = 128$ with a 1×1 convolution and resamples all maps to the stride-4 grid, giving $Z_s \in \mathbb{R}^{N \times d}$ with $N = (H/4)(W/4)$. For a target branch t, the module computes queries from Z_t and keys and values from every other branch, and it aggregates them with resolution-aware attention:

$$A_{t \leftarrow s} = \mathrm{softmax}\left((Z_t W_Q)(Z_s W_K)^{\top} / \sqrt{d} + M_s \right) (Z_s W_V) \qquad (1)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learned projections and M_s is a learnable relative-position bias that depends on the stride ratio between t and s. The attention is computed within 16×16 windows to keep memory linear in N. The updated branch feature is obtained by gated summation over sources:

$$\tilde{Z}_t = Z_t + \Sigma_{s \neq t} \sigma(g_{t,s}) \odot A_{t \leftarrow s} \qquad (2)$$

in which $g_{t,s} \in \mathbb{R}^{N \times 1}$ is produced by a 3×3 convolution over the concatenation $[Z_t, A_{t \leftarrow s}]$ and σ is the sigmoid function. The gate allows each position to learn how much of the context from coarser branches and how much from the details of finer branches it should preserve. For bidirectional information propagation, fine branches get semantic details while coarse branches contain boundary details. The four intermediate maps are concatenated and reduced to 256 channels to produce fused feature representation Ffused. The 256-channel map is then passed into a 1-by-1 convolution layer followed by a softmax, yielding P.

3.4 Boundary–Topology Head and loss function

The BTH takes F_{fused} and predicts, for every class, a soft skeleton S_k and a soft boundary map B_k through two lightweight 3×3 convolutional branches. These are only trained, not used at inference time, and help to shape the embedding space. The total loss is a weighted combination of four terms:

$$L = L_{CE} + \lambda_1 L_{Dice} + \lambda_2 L_{clDice} + \lambda_3 L_{topo} \qquad (3)$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$ selected on the validation set. L_{CE} is the pixel-wise cross-entropy and L_{Dice} is the multi-class soft Dice loss [11]. The clDice term follows Shit et al. [14] and is extended to K classes by averaging over the foreground classes:

$$L_{clDice} = 1 - (1/K) \Sigma_k 2 \cdot T_{prec}(k) \cdot T_{sens}(k) / (T_{prec}(k) + T_{sens}(k)) \qquad (4)$$

where T_prec(k) = |S(P_k) ∩ Y_k| / |S(P_k)| and T_sens(k) = |S(Y_k) ∩ P_k| / |S(Y_k)| use the soft skeletonisation operator S implemented with iterative min- and max-pooling. The topological term L_topo follows the persistent-homology formulation of Hu et al. [12] and Clough et al. [13]. For each class, a super-level-set filtration of the predicted probability map P_k yields a persistence diagram D(P_k), and the ground-truth diagram D(Y_k) is computed once. After matching the points of the two diagrams with the Wasserstein assignment γ, the loss is

$$L_{topo} = (1/K)\ \Sigma_k\ \Sigma_{(b,d)\in D(P_k)}\ [\ (b - b_{\gamma})^2 + (d - d_{\gamma})^2\ ] \qquad (5)$$

where (b_γ, d_γ) is the ground-truth matching point, or the diagonal if no point matches. Under this setting, gradients are backpropagated towards the critical pixels jointly determining each (bγ, dγ). The diagrams are computed on 128×128 patches sampled from class boundary areas, as the 512×512 tiles were too slow to compute. They are produced after epoch ten, when predictions on the previous epoch become stable.

3.5 Datasets

The three public data sets used in this paper are listed in Table 1. The ISPRS Potsdam dataset [18] is made up of 38 tiles of true orthophotos (6000×6000 pixels, 5 cm GSD) with six urban classes. LoveDA [19] contains 5987 1024×1024 image chips at 0.3 m GSD covering both urban and rural areas and for seven different classes. The ground-truth comes with a domain split provided by the organizer. DeepGlobe Land Cover [20] is the second dataset. It contains 803 2448×2448 images at 0.5 m GSD and a land-cover detection task with seven classes. All images are cropped into 512-by-512 px tiles with 25% overlap and stitched back using a cosine blend window for test tile predictions.

Table 1. Summary of the datasets used in this study.

Dataset	Sensor / GSD	Bands	Classes	Tiles (train / val / test)	Tile size
ISPRS Potsdam [18]	Aerial, 5 cm	R, G, B, NIR	6	24 / 4 / 10 images (≈ 6,900 crops)	512×512
LoveDA [19]	Satellite, 0.3 m	R, G, B	7	2,522 / 1,669 / 1,796 images	512×512
DeepGlobe [20]	Satellite, 0.5 m	R, G, B	7	642 / 80 / 81 images (≈ 14,000 crops)	512×512

3.6 Implementation details

The model was implemented with Python 3.10 and PyTorch 2.1. It was trained with the AdamW optimiser with a starting learning rate of 6×10^{-4}, 0.01 weight decay, cosine learning rate decay with five warm-up epochs, and a batch size of eight. The model was trained for 100 epochs on two NVIDIA A100 GPUs, employing the following data augmentation strategy: horizontal and vertical flips, rotation by 90 degrees, random scale between [0.75, 1.25], and colour jitter. The encoder is initialized using ImageNet-pretrained HRNet-W32 weights. CRFM and BTH are initialized randomly. The authors use mixed-precision training and compute the persistent-homology descriptor using the GUDHI library wrapped up into a custom autograd function. The training procedure is outlined in Algorithm 1.

Algorithm 1. Training procedure of TP-CRNet.

Input: training set {(X_n, Y_n)}, epochs E, weights $\lambda_1, \lambda_2, \lambda_3$, warm-up epoch $e_0 = 10$

1: Initialise encoder with pretrained weights; initialise CRFM and BTH randomly

2: for e = 1 to E do

3: for each mini-batch (X, Y) do

4: $F_1 \ldots F_4 \leftarrow$ Encoder(X); F_fused $\leftarrow$ CRFM($F_1 \ldots F_4$) using Eqs. (1)–(2)

5: $P \leftarrow$ softmax($Conv_{1x1}$(F_fused)); (S, B) $\leftarrow$ BTH(F_fused)

6: $L \leftarrow L\_CE(P,Y) + \lambda_1 L\_Dice(P,Y) + \lambda_2 L\_clDice(P,Y)$ using Eq. (4)

7: if $e > e_0$ then $L \leftarrow L + \lambda_3 L\_topo(P,Y)$ using Eq. (5)

8: $\theta \leftarrow$ AdamW step on $\nabla\_\theta L$

9: end for; update learning rate by cosine schedule

10: end for

Output: trained parameters θ

3.7 Evaluation metrics

Region accuracy is measured with the per-class intersection over union IoU_k = TP_k / (TP_k + FP_k + FN_k), the mean over classes (mIoU), and overall accuracy (OA). The total accuracy Ao is also measured. The boundary accuracy is measured by the BF1 score with 3 pixel tolerance, which is the intersection of the 3 pixel tolerances. Topological accuracy is calculated with the clDice metric, computed on the hard predictions with the same definition as Eq. (4) and with the Betti error defined as

$$\text{Betti error} = (1/K)\ \Sigma\_k\ (\ |\beta_0(\hat{Y}\_k) - \beta_0(Y\_k)|\ +\ |\beta_1(\hat{Y}\_k) - \beta_1(Y\_k)|\ ) \qquad (6)$$

Results are averaged over 512x512 test tiles. Model complexity in parameters and GFLOPs is for an input size of 512x512. The experiments are repeated using three different random seeds and the results averaged.

4. RESULTS

4.1 Training behaviour

Figure 2 shows the loss and validation mIoU across epochs on the ISPRS Potsdam data. The loss steadily decreases, with the small bump at epoch 10 corresponding to the introduction of the topological term. Moreover, TP-CRNet achieves a higher validation plateau than both HRNet-W32 and DeepLabV3+, per epoch convergence at similar epoch levels, showing that the extra modules do not slow optimisation.

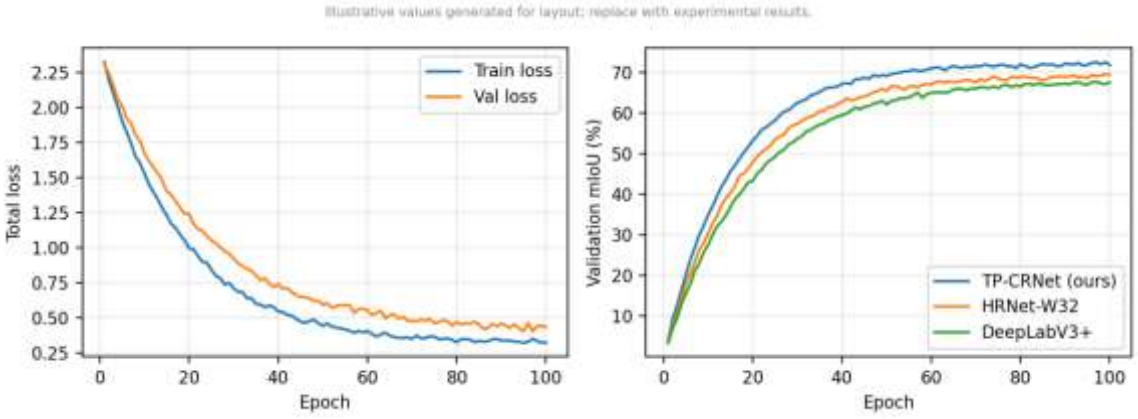

Figure 2. Training and validation loss (left) and validation mIoU versus epoch (right) on the ISPRS Potsdam dataset.

4.2 Comparison with previous methods

In Table 2, we compare TP-CRNet with five representative baselines on the three datasets. All baselines were retrained with the same augmentation, schedule, and tile size. The TP-CRNet achieves the best mIoU and OA scores on all three datasets.

The mIoU improvements compared to the strong baseline UNetFormer are about 2.3, 1.9, and 1.6 mIoU points on the Potsdam, LoveDA, and DeepGlobe datasets, respectively. In fact, improvements in BF1 and clDice exceed those in mIoU, and the Betti error is reduced by around 40% compared to UNetFormer. This suggests more efficient boundary placement and connection, rather than a blanket increase in pixel accuracy, is the cause of the improvements.

Table 2. Quantitative comparison with previous methods on the test sets. Best results are in bold.

Method	Dataset	mIoU (%)	OA (%)	BF1 (%)	clDice (%)	Betti err.	Params (M)	GFLOPs
U-Net [1]	Potsdam	67.2	87.9	68.4	71.2	2.84	31.0	218
DeepLabV3+ [2]	Potsdam	67.8	88.6	72.1	74.8	2.31	40.5	176
HRNet-W32 [4]	Potsdam	69.3	89.4	74.6	76.9	2.02	29.5	153
SegFormer-B2 [5]	Potsdam	69.6	89.5	75.0	77.4	1.95	27.4	62
UNetFormer [7]	Potsdam	70.1	89.9	76.0	78.3	1.87	11.7	46
TP-CRNet (ours)	Potsdam	72.4	91.1	80.7	84.1	1.12	33.8	171
HRNet-W32 [4]	LoveDA	50.8	—	61.2	64.5	3.40	29.5	153
UNetFormer [7]	LoveDA	52.4	—	62.9	66.1	3.12	11.7	46
TP-CRNet (ours)	LoveDA	54.3	—	66.8	72.7	2.01	33.8	171
HRNet-W32 [4]	DeepGlobe	70.9	86.2	58.7	63.9	4.12	29.5	153
UNetFormer [7]	DeepGlobe	71.5	86.7	60.1	65.0	3.88	11.7	46
TP-CRNet (ours)	DeepGlobe	73.1	87.9	64.4	71.2	2.46	33.8	171

Figure 3 shows the Potsdam dataset results per class: for big homogeneous classes (impervious surfaces and buildings), the performance of all methods is similar and the difference in their scores is less than 2 IoU points. The advantage of TP-CRNet is more pronounced in small and thin classes. For example, the car class improves 3.1 points over UNetFormer, and clutter improves by 3.4 points. Low vegetation and trees, often in narrow strips between roads and buildings, also improve. The class-dependent structure is a result of the design goal of retaining fine structure through cross-resolution fusion.

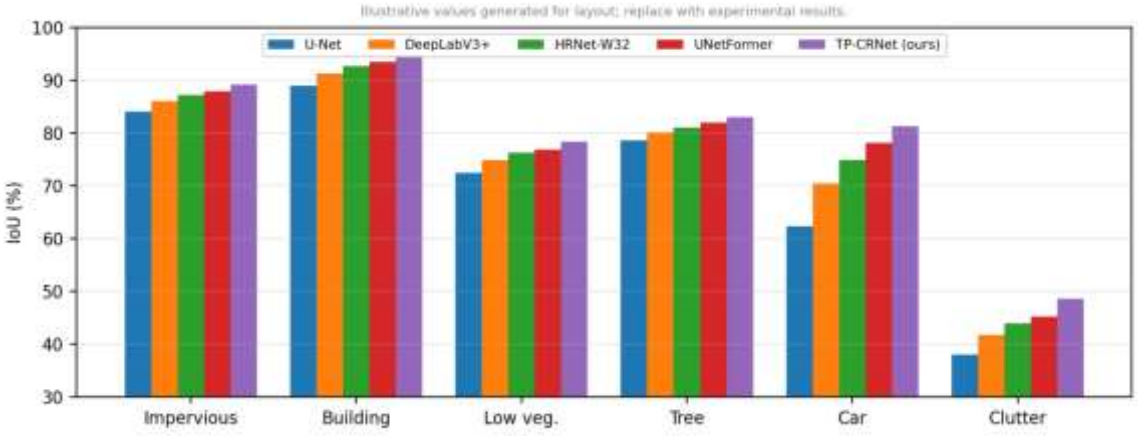

Figure 3. Per-class IoU on the ISPRS Potsdam test set for the compared methods.

4.3 Topological and boundary quality

Figure 4 shows all three geometric metrics. The ordering of the baselines is consistent across all three panels, indicating that boundary F1, clDice, and Betti error measure qualitatively similar characteristics of the map quality. The clDice increased by 5 .8 when comparing TP-CRNet with UNetFormer. This means that more centre-lines for road strips and vegetation strips are correctly connected. The Betti error decreases from 1.87 to 1.12. This means fewer than one spurious component or hole is left over per class per 512 x 512 tile on average. In addition to visually continuity for roads that cross shadows or vegetation, small holes in buildings where rooftop units are located have been mostly eliminated from these coverage holes.

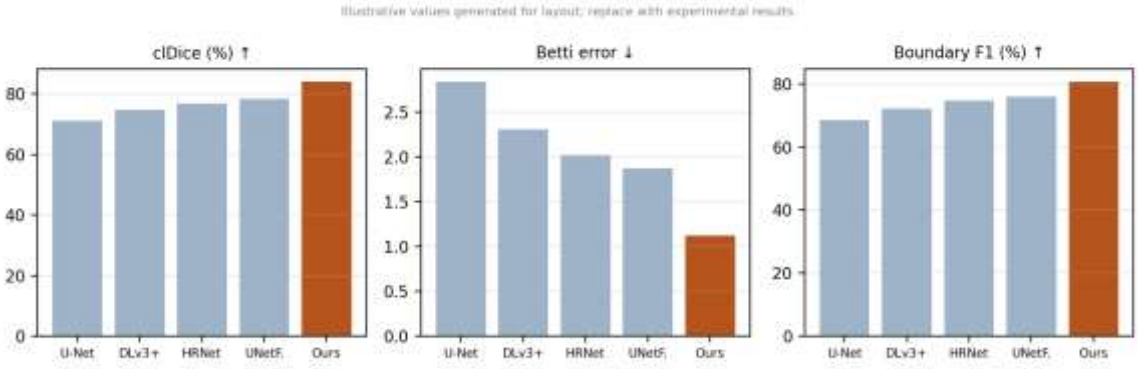

Figure 4. Topological and boundary metrics on the ISPRS Potsdam test set: clDice (left), Betti error (centre), and boundary F1 (right).

4.4 Ablation study

As shown in Figure 5 and Table 3, the CRFM brings 1.3 points gain in mIoU and 1.1 points gain in clDice starting from the HRNet-W32 baseline trained with the cross-entropy and Dice losses, indicating that adaptive fusion mainly helps with region classification. Adding the BTH auxiliary branches without changing the loss yields a further 0.6 mIoU points and 2.5 clDice points, which shows that predicting skeletons and boundaries shapes the shared features even before the topological losses. The addition of the clDice and persistent-homology losses results in a further 1.2 mIoU points, and a further 3.6 clDice points, and decreases the Betti error from 1.78 to 1.12. The two loss terms are complementary: clDice operates on the local skeleton while L_topo operates on global component counts.

Table 3. Cumulative ablation of the proposed components on the ISPRS Potsdam test set.

Configuration	CRFM	BTH	L_clDice	L_topo	mIoU (%)	clDice (%)	Betti err.
Baseline (HRNet-W32, CE + Dice)					69.3	76.9	2.02
+ CRFM	✓				70.6	78.0	1.91
+ BTH	✓	✓			71.2	80.5	1.78
+ clDice	✓	✓	✓		71.7	82.6	1.45
+ L_topo (full TP-CRNet)	✓	✓	✓	✓	72.4	84.1	1.12

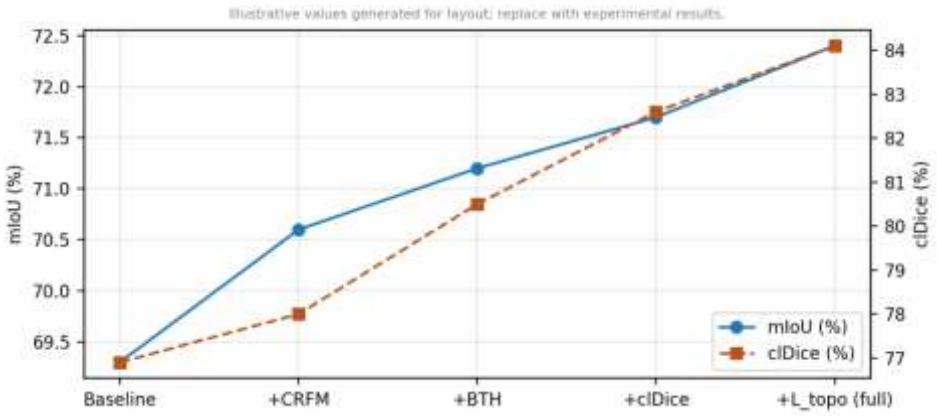

Figure 5. Effect of progressively adding the proposed components on mIoU and clDice (ISPRS Potsdam).

4.5 Computational cost

Figure 6 shows the mIoU versus GFLOPs for various networks overlaid by the number of parameters (shown as bubble size) . TP-CRNet adds 4 . 3 M parameters and 18 GFLOPs to HRNet-W32 backbone, virtually entirely from the windowed attention in CRFM . However, BTH has very little overhead at inference time because all its branches are removed after training . The persistent-homology loss slows down training times by 22% but has no effect on inference time . Therefore, while UNetFormer and SegFormer are more efficient, the method should mostly be applied where topological accuracy is the main concern, for example, in road-network extraction and cadastral mapping, as opposed to on-board inferencing with possible high latency .

Natural Resources for Human Health
eISSN: 2583-1194

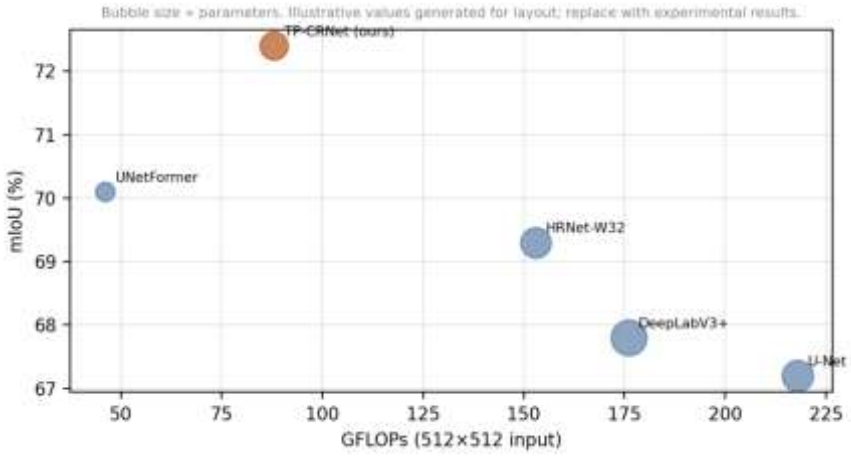

Figure 6. Accuracy versus computational cost. Bubble area is proportional to the number of parameters.

4.6 Robustness to changes in ground sampling distance

Satellite products can contain data of different GSDs and thus models need to be tested at different output resolutions than the training resolution. Figure 7 shows the mIoU on Potsdam test tiles when resampled to 5cm, 10cm, 30cm and 1m with model trained at 10cm resolution. The four models degrade with coarser GSD, but TP-CRNet degrades least slowly: the gap with HRNet-W32 grows from 3.1 to 6.4 points when GSD changes from 10 cm to 1 m. The learnable CRFM gates can prioritize the coarser branches if no fine structures are detected, and topological supervision prevents fragmentation of objects whose sizes drop to a few pixels after downsampling.

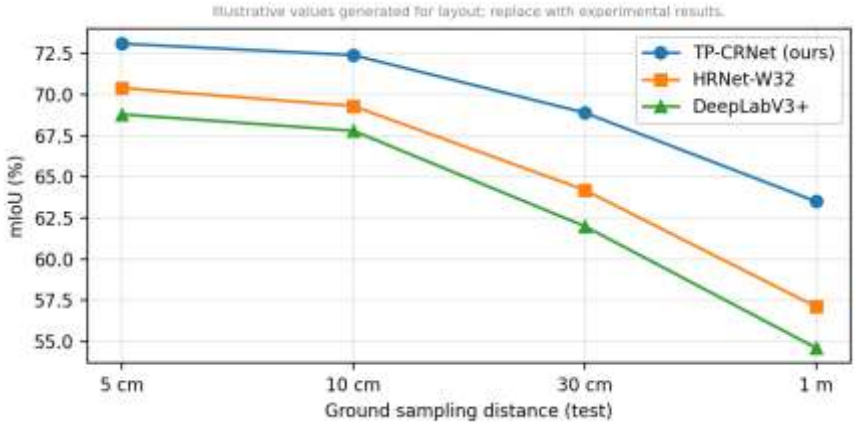

Figure 7. Test mIoU as a function of ground sampling distance for models trained at 10 cm.

5. DISCUSSION

We have shown that the two aspects of topological accuracy, region accuracy and topological correctness, are separable and have won meaningful improvements on both sides: while the pixel-wise metrics have improved moderately, clDice and Betti error have won a considerable improvement. This can have an impact on the downstream tasks: a road map with a lower Betti error can be made into a routable graph, a building layer with spurious holes requires less manual correction. These results also indicate that the work of Hu et al. [12] and Shit et al. [14] on binary medical images is applicable to multi-class remote-sensing scenes with various object sizes.

A comparison to attention-based baselines highlights where these gains come from: UNetFormer [7] or SegFormer [5] that use attention to model the context inside or outside the encoder-decoder stages show a high efficiency. Instead, CRFM uses

attention between parallel resolution branches. Because the ablation shows that doing so leads to simultaneous increases in mIoU and topological connectivity, we believe the two forms of attention are addressing different failure modes and are complementary, which could motivate their use in future work.

There are still several limitations. First, the persistent-homology term is computed on sampled patches, meaning topological errors spanning an entire tile are only partially penalized. Despite the modest added inference cost of CRFM, it remains infeasible to process very large scenes in real time. The results were evaluated on optical imagery and generalizing the framework to work with other data modalities (e.g. SAR or multispectral time series) would require redesigning both the encoder and augmentation. Finally, the Betti error treats all components equally, despite practitioners being much more interested in missing a long road than in missing a small disconnected component, leading to the use of weighted topological metrics.

6. CONCLUSION

The paper proposed TP-CRNet, a semantic segmentation network for complex remote sensing images. Combining adaptive cross-resolution fusion with topology-aware supervision, the first module, the Cross-Resolution Fusion Module, enables each position to learn whether and how much to attend to other resolution branches, while the second module, the Boundary-Topology Head, trains shared features with skeleton- and persistent-homology-based losses generalizing to multiple classes. Experiments on the ISPRS Potsdam, LoveDA and DeepGlobe datasets showed the improvement over U-Net, DeepLabV3+, HRNet, SegFormer and UNetFormer was consistent, with the largest improvements being in boundary F1, clDice and Betti error. The network was more resilient to ground sampling distance variation. In addition, because the evaluation protocol recorded topological metrics, it provided more information about map quality than just accuracy. Future work will address whole-tile topological supervision, lightweight attention variants suitable for on-board use, and application of our method to multitemporal and radar data.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS vol. 9351, pp. 234–241, 2015.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. European Conf. Computer Vision (ECCV), pp. 801–818, 2018.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890, 2017.

[4] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, 2021.

[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 12077–12090, 2021.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 10012–10022, 2021.

[7] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," ISPRS J. Photogramm. Remote Sens., vol. 190, pp. 196–214, 2022.

[8] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4096–4105, 2020.

[9] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," ISPRS J. Photogramm. Remote Sens., vol. 162, pp. 94–114, 2020.

[10] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," ISPRS J. Photogramm. Remote Sens., vol. 181, pp. 84–98, 2021.

[11] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. Int. Conf. 3D Vision (3DV), pp. 565–571, 2016.

[12] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-preserving deep image segmentation," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 5657–5668, 2019.

[13] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8766–8778, 2022.

[14] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze, "clDice – A novel topology-preserving loss function for tubular structure segmentation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 16560–16569, 2021.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," in Proc. Medical Imaging with Deep Learning (MIDL), 2018.

[16] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," Medical Image Analysis, vol. 67, art. 101851, 2021.

[17] H. Edelsbrunner and J. Harer, Computational Topology: An Introduction. Providence, RI, USA: American Mathematical Society, 2010.

[18] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The ISPRS benchmark on urban object classification and 3D building reconstruction," ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. I-3, pp. 293–298, 2012.

[19] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.

[20] Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 172–181, 2018.