BARE: Boundary-Aware with Resolution Enhancement for Tree Crown Delineation

Attavit Wilaiwongsakul[1,[0009−0003−3084−9419]] ✉, Bin Liang[1,[0000−0002−6605−2167]], Wenfeng Jia[2,[0000−0002−3996−5438]], Bryan Zheng[1,[0000−0003−1223−9230]], and Fang Chen[1,[0000−0003−4971−8729]]

[1] University of Technology Sydney, Sydney, Australia
`attavit.wilaiwongsakul@student.uts.edu.au`, `bin.liang@uts.edu.au`, `boyuan.zheng@uts.edu.au`, `fang.chen@uts.edu.au`
[2] Charles Sturt University, Australia
`wjia@csu.edu.au`

Abstract. Tree Crown Delineation (TCD) requires precise boundary segmentation, yet reduced-resolution architectures face inherent limitations due to decoder outputs at lower spatial resolutions. We propose BARE (Boundary-Aware with Resolution Enhancement), a simple architecture-preserving training strategy that combines external full-resolution loss supervision with class weighting to address dataset imbalance. BARE upsamples decoder output solely during training, maintaining inference efficiency while improving boundary precision. Through comprehensive evaluation on the OpenAerialMap Tree Crown Delineation (OAM-TCD) dataset, BARE demonstrates consistent improvements over baseline configurations across SegFormer, PSPNet, and SETR architectures, achieving Boundary Intersection-over-Union (B-IoU) improvements across diverse architectural paradigms. To rigorously evaluate boundary quality, we introduce B-IoU to TCD research. Our systematic evaluation validates that external full-resolution supervision consistently outperforms standard training approaches while maintaining computational efficiency, demonstrating that methodical training optimization offers a practical approach for enhancing boundary precision in reduced-resolution segmentation architectures. Code: `https://github.com/attavit14203638/bare`

Keywords: Tree Crown Delineation · Boundary IoU · Vision Transformers

1 Introduction

Precise tree crown boundaries are essential for accurate forest monitoring, yet automated methods struggle to achieve the boundary precision required for critical applications such as biomass estimation, biodiversity assessment, and climate change mitigation planning [6,14]. While pixel-level accuracy metrics may suggest high performance, boundary imprecision—manifesting as edge blurring, crown merging, and fragmented delineations—undermines the reliability of

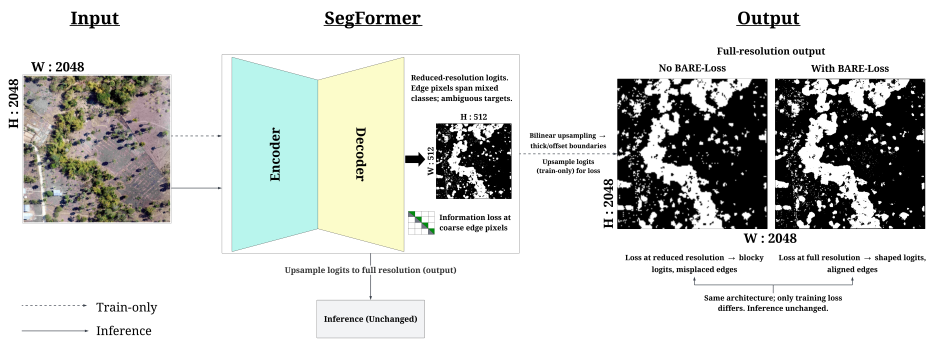


Fig. 1: Reduced-resolution decoder challenge.

downstream analyses in dense forest canopies where individual tree discrimination is paramount. Traditional computer vision approaches face particular challenges with high-resolution aerial imagery, where fine spatial details directly impact the quality of forest inventory and ecosystem monitoring applications.

The evolution to deep learning has revolutionized TCD, with Vision Transformer architectures demonstrating significant advantages through global context modeling [6,7,8]. However, several segmentation architectures including SegFormer [17], PSPNet [21], and SETR [22] share a fundamental bottleneck: their decoders produce predictions at reduced spatial resolution requiring upsampling, potentially compromising boundary quality. This architectural limitation manifests as a critical resolution mismatch where high-resolution input images (2048×2048 pixels) are processed through encoders that maintain spatial detail, but decoders output predictions at significantly reduced resolutions—SegFormer at 1/4, PSPNet at 1/8, and SETR at 1/16 of the original size. The subsequent bilinear upsampling to restore full resolution introduces boundary artifacts that severely impact crown separation accuracy in dense forest canopies where precise delineation between adjacent trees is paramount (Figure 1). To address current evaluation limitations in boundary assessment, we introduce boundary IoU (B-IoU) [2] to TCD evaluation—to our knowledge, the first application of this boundary-focused metric in TCD research. Unlike standard IoU metrics that treat all pixels equally, B-IoU specifically evaluates segmentation quality within narrow boundary regions, revealing that full-resolution loss supervision during training produces markedly cleaner crown boundaries with reduced edge artifacts and improved crown separation compared to standard reduced-resolution training approaches (Figure 2).

We propose **BARE** (Boundary-Aware with Resolution Enhancement): an architecture-preserving training strategy that combines external full-resolution loss supervision with class weighting to address data set imbalance. Rather than modifying architectural structures, our approach upsamples decoder output solely during training for loss computation, maintaining inference efficiency while improving boundary quality. In the OAM-TCD dataset, BARE achieves

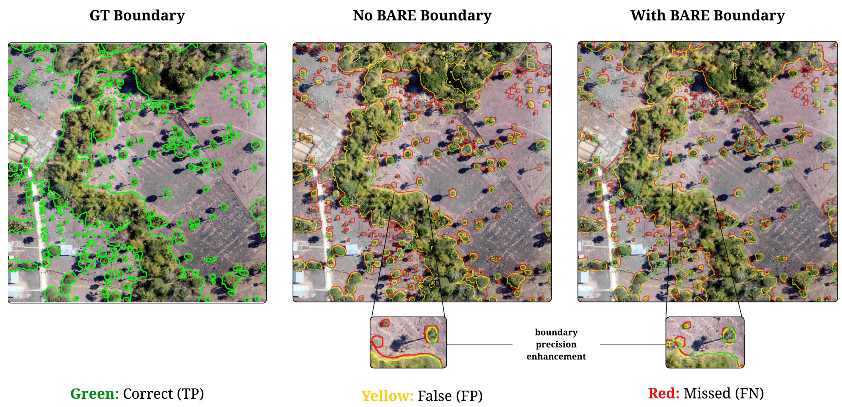


Fig. 2: B-IoU boundary precision evaluation.

superior boundary accuracy, demonstrating that methodical optimization of training supervision can be more effective than architectural modifications for boundary-sensitive applications. Our contributions are threefold: First, we propose BARE as an architecture-preserving approach that improves boundary quality without structural changes, offering a practical solution for deployment scenarios requiring computational efficiency. Second, we introduce B-IoU to TCD research, establishing the first boundary-focused assessment approach in this domain. Third, we demonstrate that external full-resolution supervision consistently benefits multiple reduced-resolution architectures (SegFormer, PSPNet, and SETR), while revealing that optimal training configurations exhibit architecture-dependent characteristics—informing practitioners when to apply full BARE versus its individual components.

2 Related Work

This section contextualizes our high-resolution TCD research by reviewing the evolution of automated TCD methods toward Transformer-based architectures, challenges in preserving fine-grained spatial details, and existing approaches for high-resolution semantic segmentation and boundary refinement.

Deep learning for TCD has evolved from traditional Convolutional Neural Networks (CNNs) like U-Net [6,12] to advanced self-attention models such as SegFormer [1,7,8,14]. Despite impressive benchmark performances exceeding ninety-five percent F1 scores [8,16], precise boundary delineation remains challenging due to dense canopies [4,14,23], small tree instances [6], shadow effects [13], and overlapping crowns [20]. Advanced models often introduce significant computational costs [3,14], necessitating trade-offs between accuracy and efficiency [9].

Semantic segmentation with Transformers in TCD features distinct architectural approaches. SETR [22] represents pure Vision Transformer design without hierarchical feature extraction, while SegFormer [17] employs hierarchical Transformer encoders with lightweight Multi-Layer Perceptron (MLP) decoders for multi-scale feature fusion. Both output logits at reduced resolution requiring upsampling, potentially losing fine-grained details crucial for TCD. This architectural diversity motivates evaluating training strategies that enhance boundary quality without architectural modifications [5].

High-resolution semantic segmentation in TCD addresses fine spatial detail recovery through learned upsampling, multi-scale fusion, and attention mechanisms [10,11,19], though increasing model complexity. Alternative approaches include interpolation-based upsampling and boundary refinement modules [15,18,20], albeit at increased computational cost. Our approach differs from dedicated boundary refinement modules, which often add computational overhead at inference time. Instead, BARE focuses on a training-only strategy to enhance boundary learning within the existing architecture, thus preserving computational efficiency. Classical CNN architectures like PSPNet [21] face similar challenges through pyramid pooling at reduced spatial resolution. Our work investigates external upsampling strategies requiring careful optimization for meaningful boundary improvements.

3 Methodology

We present BARE, an architecture-preserving training strategy that addresses the resolution bottleneck in segmentation models for TCD applications. BARE combines external full-resolution loss supervision with class weighting to improve boundary precision without architectural modifications.

3.1 BARE Framework Overview

The BARE framework targets segmentation architectures that produce decoder outputs at reduced spatial resolution, including SegFormer ($\frac{H}{4} \times \frac{W}{4}$), PSPNet ($\frac{H}{8} \times \frac{W}{8}$), and SETR ($\frac{H}{16} \times \frac{W}{16}$). Figure 3 illustrates our approach: rather than modifying architectural structures, BARE applies two complementary training strategies. Panel (a) shows standard training with loss computed at reduced resolution requiring ground truth downsampling, while panel (b) demonstrates BARE training with external upsampling for full-resolution loss supervision. The approach is broadly applicable across different resolution reduction factors while maintaining inference efficiency through training-only modifications. BARE combines (1) external full-resolution loss supervision that upsamples reduced-resolution logits solely during training, and (2) class weighting to address dataset imbalance. While external full-resolution supervision benefits all tested architectures, our results reveal that combining both components exhibits architecture-dependent effectiveness, with pure-transformer architectures like SETR showing sensitivity to the combined approach.

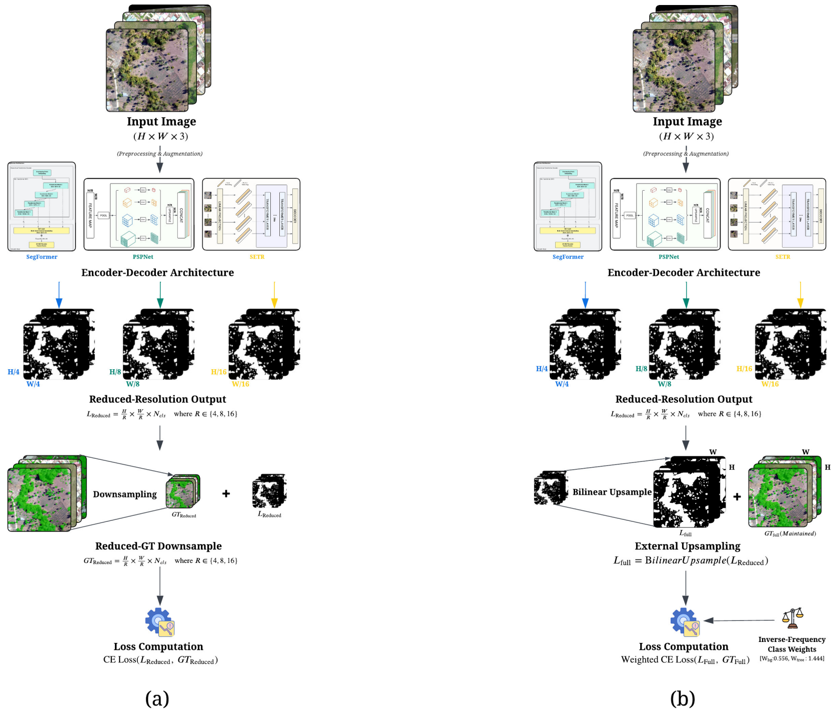


Fig. 3: BARE framework: (a) Standard vs (b) BARE training.

3.2 Reduced-Resolution Processing Across Architectures

Contemporary segmentation architectures commonly produce decoder outputs at reduced spatial resolution due to computational efficiency considerations. For an input image $I \in \mathbb{R}^{H \times W \times 3}$, these architectures generate logits $L_{reduced} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times N_{cls}}$ where R represents the resolution reduction factor and N_{cls} is the number of classes. The reduction factor varies by architecture: SegFormer uses $R = 4$, PSPNet uses $R = 8$, and SETR uses $R = 16$. Standard training applies loss computation directly on $L_{reduced}$ or its upsampled version, potentially compromising boundary quality due to the resolution mismatch with ground truth masks at full resolution $H \times W$.

3.3 External Full-Resolution Supervision Strategy

The core component of BARE is external full-resolution supervision, which bridges the resolution gap between reduced-resolution decoder outputs and full-resolution ground truth masks during training. As depicted in Figure 3(b), this strategy externally upsamples the reduced-resolution logits for loss computation while maintaining the original architecture unchanged. For an input image with spatial size $H \times W$ and a model producing reduced-resolution logits

$L_{reduced} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times N_{cls}}$ (R: resolution reduction factor; N_{cls}: number of classes), we bilinearly upsample to obtain full-resolution logits:

$$L_{full} = \text{BilinearUpsample}(L_{reduced}, \text{size} = (H, W)) \tag{1}$$

Here, $L_{full} \in \mathbb{R}^{H \times W \times N_{cls}}$ matches the spatial dimensions of the ground truth labels. This upsampling is applied externally during training only, preserving inference efficiency at test time. Predicted probabilities are computed via a per-pixel softmax over classes:

$$P_{i,c} = \frac{\exp(L_{full,i,c})}{\sum_{j=1}^{N_{cls}} \exp(L_{full,i,j})} \tag{2}$$

where i indexes pixels and c and j index classes.

3.4 Class Weighting Strategy

TCD datasets exhibit significant class imbalance, with tree crown pixels typically representing a minority class. To address this, BARE uses inverse-frequency class weights with unit-mean normalization across classes. For a dataset with N_{ds} total labeled pixels and n_c pixels belonging to class c, define the class frequency $f_c = \frac{n_c}{N_{ds}}$ and the class weight is:

$$w_c = \frac{\frac{1}{f_c}}{\frac{1}{N_{cls}} \sum_{j=1}^{N_{cls}} \frac{1}{f_j}} \tag{3}$$

Here, N_{cls} denotes the number of classes (2 for binary TCD), j indexes the classes in the summation, and f_j is the class frequency of class j (with $j = 1, \ldots, N_{cls}$). This normalization keeps weights inversely proportional to class frequency while ensuring their arithmetic mean equals one. For the OAM-TCD dataset with approximately 27.8% tree crown pixels ($f_{\text{tree}} \approx 0.278$, $f_{\text{background}} \approx 0.722$), this yields $w_{\text{background}} = 0.556$ and $w_{\text{tree}} = 1.444$.

3.5 Complete Loss Function Formulation

BARE employs Weighted Cross-Entropy (WCE) loss computed on full-resolution predictions. The complete loss formulation is:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{N_{cls}} w_c \cdot y_{i,c} \cdot \log(P_{i,c} + \epsilon) \tag{4}$$

where $N = H \times W$ represents the total number of pixels in the input image, N_{cls} denotes the number of classes (2 for binary TCD), w_c is the class weight for class c as computed in Equation 3, $y_{i,c} \in \{0, 1\}$ represents the binary ground truth label for pixel i and class c, and $P_{i,c}$ is the predicted probability for pixel i and class c obtained after softmax normalization. For computational stability, we add a small epsilon ($\epsilon = 1 \times 10^{-7}$) to the logarithm: $\log(P_{i,c} + \epsilon)$. Intuitively, ϵ

acts as a numerical safety net: if a predicted probability is exactly zero, log(0) is undefined; adding ϵ makes it a tiny positive value so $\log(P_{i,c} + \epsilon)$ remains finite. Its effect on the loss is negligible for typical probabilities and only matters in such extreme near-zero cases, but it prevents instabilities and training crashes. This formulation ensures that boundary pixels receive appropriate supervision at full spatial resolution while addressing class imbalance through weighted loss computation.

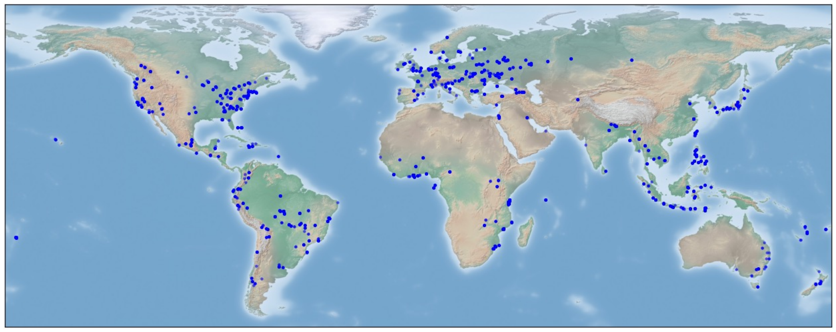
4 Experiments

This section presents our comprehensive experimental evaluation of the BARE framework across multiple reduced-resolution architectures. We begin with detailed dataset characteristics and experimental setup, followed by implementation details and baseline configurations. We then present our results with integrated analysis, demonstrating BARE's effectiveness in improving boundary precision for TCD applications across different architectural paradigms.

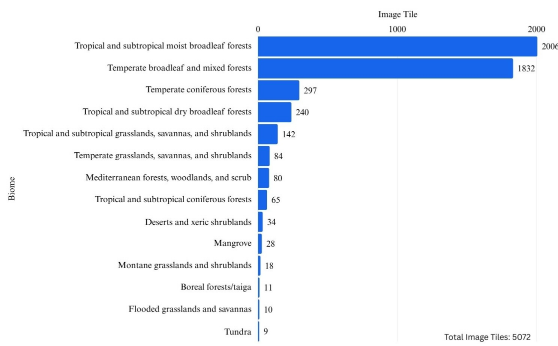
4.1 Dataset and Experimental Setup

All experiments are conducted on the OpenAerialMap Tree Crown Delineation (OAM-TCD) dataset [14], sourced from Hugging Face Hub as `restor/tcd`. This dataset provides global coverage across diverse forest environments spanning multiple continents and biomes, making it ideal for evaluating BARE's generalizability across different ecological contexts and architectural paradigms. As shown in Figure 4, the dataset encompasses temperate forests, tropical regions, and agricultural landscapes across North America, Europe, Africa, Asia, and Australia. The dataset features high-resolution aerial imagery (10 cm ground sampling distance) with 2048×2048 pixel images split into 4169/416/439 samples for training/validation/testing respectively. The binary segmentation task exhibits significant class imbalance with tree crown pixels representing 27.8% of training pixels, motivating our inverse-frequency class weighting strategy with weights of [0.556, 1.444] for background and tree crown classes respectively.

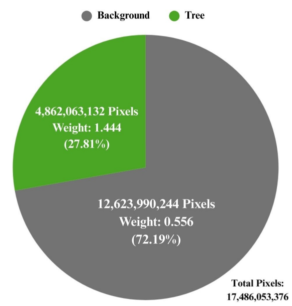
To ensure model generalization while preserving boundary characteristics essential for B-IoU evaluation, we apply a comprehensive augmentation pipeline exclusively to the training set. Geometric transformations include random crops to 1024×1024 pixels (maintaining full-resolution supervision compatibility), horizontal/vertical flips, and rotations up to ±180° that preserve crown boundary integrity. Color augmentations include brightness/contrast/saturation/hue jittering (P=0.7), Gaussian blur (P=0.5), and atmospheric/shadow effects (P=0.2 each) that simulate real-world imaging conditions without compromising edge definition quality. This strategy maintains spatial dimensions required for external full-resolution loss computation while enhancing model robustness across diverse environmental conditions.



(a) Global distribution [14]



(b) Biome diversity



(c) Class distribution

Fig. 4: OAM-TCD dataset characteristics.

4.2 Implementation Details and Model Configurations

We evaluate BARE across three reduced-resolution architectures: SegFormer (MiT-B5 backbone), PSPNet (ResNet-50 backbone), and SETR (Vision Transformer base model). All architectures utilize pre-trained ImageNet weights when available. The BARE framework applies external upsampling to decoder outputs during training: SegFormer logits at $\frac{H}{4} \times \frac{W}{4}$ resolution, PSPNet at $\frac{H}{8} \times \frac{W}{8}$, and SETR at $\frac{H}{16} \times \frac{W}{16}$ are bilinearly upsampled to $H \times W$ for loss computation, while maintaining original inference paths unchanged.

All models use identical standardized optimization settings: AdamW optimizer (lr=1×10^{-5}, weight decay=0.01), cosine annealing schedule with 10% warmup, mixed precision (FP16), and gradient accumulation (effective batch size 8). Training spans 50 epochs on NVIDIA L4 GPUs using PyTorch 2.6.0 and Transformers 4.51.3. Class weights of [0.556, 1.444] address the dataset imbalance when enabled.

For systematic evaluation, we implement four configurations per architecture: (1) Baseline (standard training with reduced-resolution loss), (2) Class Weighting only (CW), (3) Full-Resolution supervision only (Full-Res), and (4)

BARE (our complete approach combining full-resolution supervision with class weighting).

4.3 Evaluation Metrics and Boundary Assessment

Precise boundary delineation is crucial in TCD applications. We employ B-IoU [2]—to our knowledge, the first application of this boundary-focused metric in TCD research. B-IoU evaluates segmentation quality within a narrow band around object contours, providing direct measurement of boundary precision that complements traditional pixel-wise metrics. While B-IoU is not our novel contribution, its adoption enables rigorous assessment of boundary quality specifically relevant to tree crown delineation accuracy.

Boundary IoU Definition. Given ground truth mask G and prediction mask P, B-IoU computes IoU for mask pixels within distance d from the respective contours. Formally, B-IoU is defined as:

$$\text{B-IoU}(G, P) = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|} \tag{5}$$

where G_d and P_d represent boundary regions—the set of pixels within distance d from the ground truth and prediction contours, respectively. Unlike standard IoU that considers all pixels equally, B-IoU focuses exclusively on boundary regions, making it significantly more sensitive to boundary quality than traditional pixel-wise metrics. This boundary-focused evaluation is particularly valuable for TCD applications where precise crown separation is critical. The distance parameter d controls the sensitivity: we set d to 2% of the image diagonal following established guidelines [2], which corresponds to approximately 15 pixels for our 2048×2048 input images. This parameter choice ensures that B-IoU captures meaningful boundary errors while remaining robust to minor annotation ambiguities inherent in crown boundary delineation.

We also evaluate using comprehensive standard segmentation metrics: Intersection over Union (IoU), F1-Score (Dice coefficient), Precision, Recall, and overall Pixel Accuracy. These provide complementary perspectives on model performance, with IoU measuring region overlap, F1-Score offering balanced precision-recall assessment, and individual precision/recall metrics quantifying model exactness and completeness in tree crown detection.

4.4 Results and Analysis

Performance Analysis Table 1 presents our comprehensive evaluation across three architectures and multiple training configurations. The results reveal architecture dependent effectiveness of BARE training strategies, with significant boundary quality improvements for SegFormer and PSPNet, but unexpected negative interaction effects for SETR.

SegFormer demonstrates the strongest response to complete BARE, with boundary quality improving 5.1% from baseline (B-IoU: 0.590) to BARE (B-IoU:

Table 1: Performance metrics across configurations and architectures.

Config	Architecture	IoU	F1	Prec	Rec	Acc	B-IoU
Baseline	SETR	0.885	0.939	**0.949**	0.930	**0.971**	0.644
Baseline	SegFormer	0.817	0.899	0.931	0.870	0.953	0.590
Baseline	PSPNet	0.724	0.840	0.861	0.820	0.924	0.391
CW	SETR	0.881	0.937	0.925	0.949	0.969	0.615
CW	SegFormer	0.845	0.916	0.884	0.950	0.959	0.606
CW	PSPNet	0.727	0.842	0.782	0.911	0.916	0.378
Full-Res	SETR	**0.890**	**0.942**	0.945	0.938	**0.971**	**0.667**
Full-Res	SegFormer	0.828	0.906	0.931	0.882	0.957	0.610
Full-Res	PSPNet	0.724	0.840	0.859	0.822	0.924	0.392
BARE	SETR	0.878	0.935	0.908	**0.964**	0.967	0.624
BARE	SegFormer	0.848	0.918	0.885	0.953	0.960	0.620
BARE	PSPNet	0.729	0.844	0.806	0.884	0.921	0.395

0.620), while overall IoU advances from 0.817 to 0.848. PSPNet shows modest but consistent gains, with BARE enhancing B-IoU from 0.391 to 0.395. SETR reveals distinct architecture-dependent behavior: while full-resolution supervision alone achieves optimal performance (IoU: 0.890, B-IoU: 0.667), adding class weighting in the complete BARE approach creates negative interaction effects, reducing performance to IoU: 0.878 and B-IoU: 0.624.

These results clearly demonstrate that external full-resolution supervision benefits all architectures universally, but combining it with class weighting exhibits architecture-dependent behavior. SETR's superior baseline performance (IoU: 0.885, B-IoU: 0.644) and strong response to full-resolution supervision alone (achieving 0.890 IoU, 0.667 B-IoU) suggests that highly capable pure-transformer architectures may be sensitive to additional training constraints. This architecture-dependent response highlights the importance of empirical validation when combining multiple training optimizations, as the optimal configuration varies with architectural characteristics.

Qualitative Analysis Visual comparisons across architectural frameworks and landscape types demonstrate BARE's effectiveness. Figure 5 presents comprehensive qualitative comparisons showing consistent improvements in tree crown delineation quality across SegFormer, PSPNet, and SETR architectures, with notably cleaner boundaries and reduced artifacts compared to baseline approaches. BARE variants show enhanced boundary definition and reduced segmentation artifacts across diverse landscape types including rural (Sample 1), dense forest (Sample 2), urban mixed (Sample 3), and agricultural (Sample 4) environments. The analysis reveals systematic enhancement in boundary coherence, particularly evident in challenging scenarios with dense canopy coverage and complex urban-forest interfaces.

Figure 6 reveals error pattern effectiveness across architectures, where green regions indicate correct predictions, yellow shows false positives, and red repre-

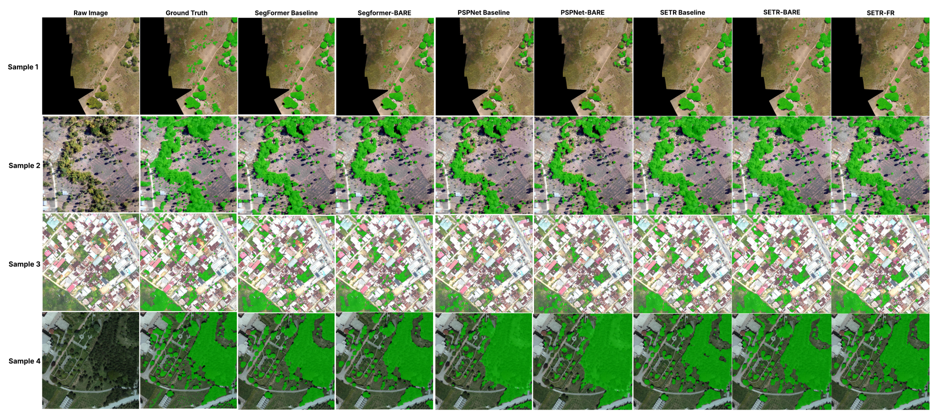


Fig. 5: Cross-architecture prediction comparison.

sents missed detections. BARE variants consistently exhibit reduced false positive artifacts (yellow regions) and improved boundary precision, with notably balanced error patterns that support superior boundary quality through external full-resolution supervision. The systematic reduction in yellow artifacts demonstrates BARE's effectiveness in reducing segmentation artifacts across all tested architectures.

Boundary-focused validation in Figure 7 provides direct evidence of BARE's enhancement capabilities, where green regions show correct boundaries. The progressive improvement in boundary quality from baseline to BARE variants is evident across all architectures, with BARE achieving superior boundary coherence across SegFormer, PSPNet, and SETR architectures compared to their respective baselines, particularly in regions with complex boundary structures.

Computational Efficiency Analysis BARE maintains efficiency by applying modifications only during training while preserving inference performance. Inference times remain largely unchanged across architectures (SegFormer: 274ms to 290ms, PSPNet: 195ms to 201ms, SETR: 184ms), while training overhead remains acceptable given boundary quality improvements. Parameter counts stay constant within each architecture (84.6M SegFormer, 49.0M PSPNet, 92.1M SETR), as BARE introduces no additional model parameters. Memory increases are minimal (SegFormer: 14.4GB to 14.7GB, PSPNet: 12.1GB to 12.3GB), making BARE practical for deployment scenarios requiring both accuracy and computational efficiency.

4.5 Discussion and Implications

External full-resolution supervision consistently outperforms architectural modifications by enabling decoders to learn robust low-resolution representations that produce cleaner boundaries when upsampled.

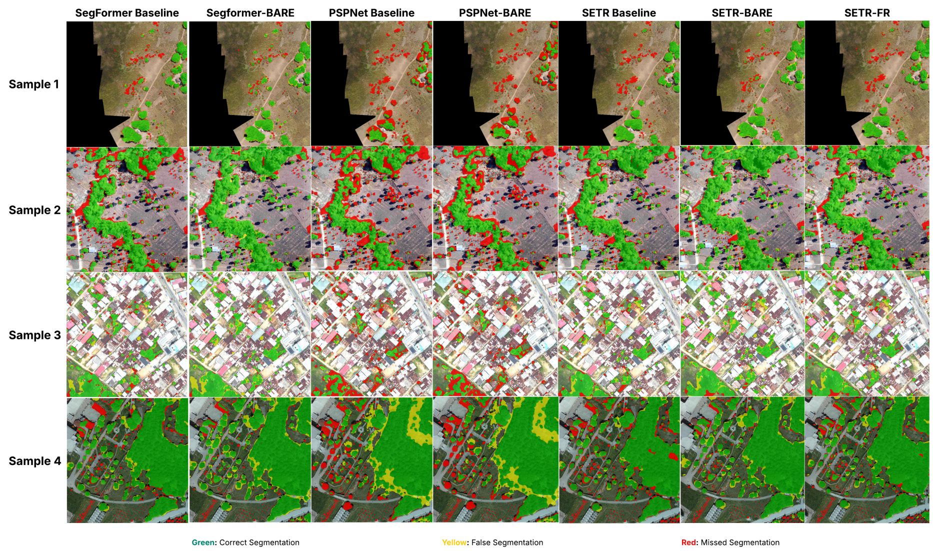


Fig. 6: Error pattern analysis across architectures.

Architecture-Dependent Effectiveness. The two BARE components exhibit different generalization characteristics. External full-resolution supervision benefits all tested architectures universally, with SETR Full-Res achieving highest performance (IoU: 0.890, B-IoU: 0.667). However, combining it with class weighting shows architecture-dependent behavior: while SegFormer and PSPNet benefit from complete BARE, SETR experiences negative interactions (IoU: 0.890 to 0.878). This suggests pure-transformer architectures already achieve near-optimal precision-recall balance, making class weighting redundant, while hybrid and CNN-based models require both components to handle class imbalance effectively.

Practical Implications. These findings provide actionable guidance: (1) apply external full-resolution supervision broadly across reduced-resolution architectures, (2) validate class weighting empirically per architecture as highly capable models may not benefit, and (3) for pure-transformer architectures, apply full-resolution supervision alone first. The training-only nature makes empirical validation computationally feasible.

Limitations. Our evaluation uses only the OAM-TCD dataset; generalizability across different remote sensing domains and TCD datasets requires validation. The study examines three architectural paradigms; effectiveness on newer transformer architectures (e.g., Swin Transformer) remains unexplored. We employ only bilinear upsampling; learned upsampling or alternative interpolation strategies may yield different results. Finally, deeper analysis of the mechanisms driving architecture-specific responses would strengthen understanding of optimal training strategies.

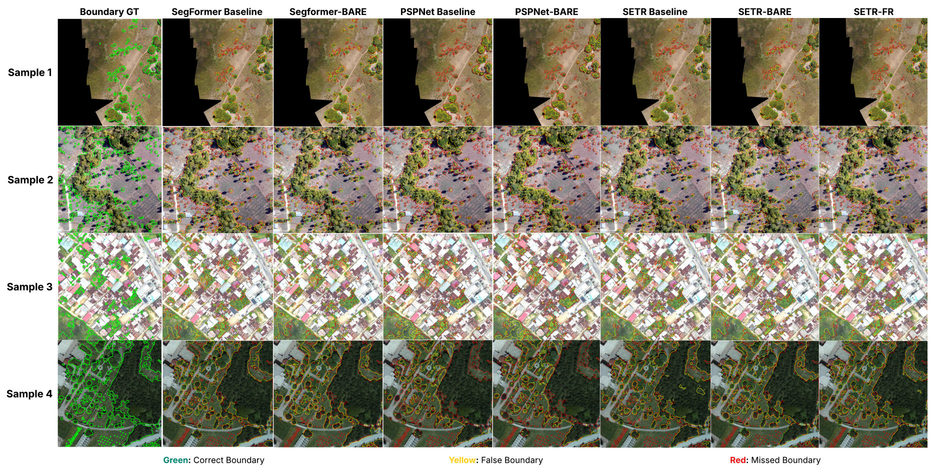


Fig. 7: Boundary-specific analysis results.

Conclusion

This paper introduces BARE, an architecture-preserving training strategy combining external full-resolution loss supervision with class weighting to enhance tree crown delineation boundary precision. Our evaluation reveals that external full-resolution supervision consistently benefits all tested architectures (SegFormer, PSPNet, and SETR), achieving boundary IoU improvements up to 5.3%, while class weighting exhibits architecture-dependent behavior—benefiting hybrid and CNN-based models but showing negative interactions with pure-transformer architectures. The training-only nature maintains computational efficiency, enabling practitioners to enhance existing models without structural changes or inference costs.

We introduce the B-IoU evaluation to TCD research, establishing the first boundary-focused assessment methodology in this domain. Future work should validate BARE across additional TCD datasets, explore learned upsampling strategies, and investigate mechanisms driving architecture-specific responses to inform refined guidelines for applying training optimizations across different architectural paradigms.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Ruzouq, R., Gibril, M.B.A., Shanableh, A., Bolcek, J., Lamghari, F., Hammour, N.A., El-Keblawy, A., Jena, R.: Spectral–Spatial transformer-based se-

mantic segmentation for large-scale mapping of individual date palm trees using very high-resolution satellite data. Ecological Indicators **163**, 112110 (2024). https://doi.org/10.1016/j.ecolind.2024.112110
2. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15334–15342. IEEE (2021). https://doi.org/10.1109/cvpr46437.2021.01508
3. Deng, G., Wu, Z., Xu, M., Wang, C., Wang, Z., Lu, Z.: Crisscross-Global Vision Transformers Model for Very High Resolution Aerial Image Semantic Segmentation. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–19 (2023). https://doi.org/10.1109/tgrs.2023.3276172
4. Dersch, S., Schöttl, A., Krzystek, P., Heurich, M.: Towards complete tree crown delineation by instance segmentation with Mask R–CNN and DETR using UAV-based multispectral imagery and lidar data. ISPRS Open Journal of Photogrammetry and Remote Sensing **8**, 100037 (2023). https://doi.org/10.1016/j.ophoto.2023.100037
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929
6. Freudenberg, M., Magdon, P., Nölke, N.: Individual tree crown delineation in high-resolution remote sensing images based on U-Net. Neural Computing and Applications **34**(24), 22197–22207 (2022). https://doi.org/10.1007/s00521-022-07640-4
7. Georges Gomes, F.D., Furuya, M.T.G., Marcato Junior, J., Gonçalves, D.N., Martins, J.A.C., Silva, P.A., Gonçalves, W.N., Osco, L.P., Ramos, A.P.M.: Urban Trees Mapping Using Multi-Scale Rgb Image and Deep Learning Vision Transformer-Based. SSRN Electronic Journal (2022). https://doi.org/10.2139/ssrn.4167085
8. Gibril, M.B.A., Shafri, H.Z.M., Al-Ruzouq, R., Shanableh, A., Nahas, F., Al Mansoori, S.: Large-Scale Date Palm Tree Segmentation from Multiscale UAV-Based and Aerial Images Using Deep Vision Transformers. Drones **7**(2), 93 (2023). https://doi.org/10.3390/drones7020093
9. Gominski, D., Kariryaa, A., Brandt, M., Igel, C., Li, S., Mugabowindekwe, M., Fensholt, R.: Benchmarking Individual Tree Mapping with Sub-meter Imagery. arXiv.org (2023). https://doi.org/10.48550/arXiv.2311.07981
10. Liu, Y., Zhang, Y., Wang, Y., Mei, S.: Rethinking Transformers for Semantic Segmentation of Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023). https://doi.org/10.1109/tgrs.2023.3302024
11. Ren, D., Li, F., Sun, H., Liu, L., Ren, S., Yu, M.: Local-enhanced multi-scale aggregation swin transformer for semantic segmentation of high-resolution remote sensing images. International Journal of Remote Sensing **45**(1), 101–120 (2023). https://doi.org/10.1080/01431161.2023.2292550
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015). https://doi.org/10.48550/arXiv.1505.04597
13. Tao, Y., Wang, Z., Zhao, G.: Shadow-Resilient Tree Crown Detection in UAV Remote Sensing Images Using Deep Learning. In: 2024 4th International Confer-

ence on Computer Science and Blockchain (CCSB). pp. 318–322. IEEE (2024). https://doi.org/10.1109/ccsb63463.2024.10735679
14. Veitch-Michaelis, J., Cottam, A., Schweizer, D., Broadbent, E.N., Dao, D., Zhang, C., Zambrano, A.A., Max, S.: Oam-TCD: A globally diverse dataset of high-resolution tree cover maps. arXiv.org (2024). https://doi.org/10.48550/arXiv.2407.11743
15. Wang, D., Chen, Y., Naz, B., Sun, L., Li, B.: Spatial-Aware Transformer (SAT): Enhancing Global Modeling in Transformer Segmentation for Remote Sensing Images. Remote Sensing **15**(14), 3607 (2023). https://doi.org/10.3390/rs15143607
16. Wang, Y., Yang, G., Lu, H.: Domain adaptive tree crown detection using high-resolution remote sensing images. Journal of Applied Remote Sensing **16**(04) (2022). https://doi.org/10.1117/1.jrs.16.044505
17. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021). https://doi.org/10.48550/arXiv.2105.15203
18. Xu, Z., Zhang, W., Zhang, T., Yang, Z., Li, J.: Efficient Transformer for Remote Sensing Image Segmentation. Remote Sensing **13**(18), 3585 (2021). https://doi.org/10.3390/rs13183585
19. Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C.: Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–20 (2022). https://doi.org/10.1109/tgrs.2022.3144894
20. Zhang, J., Shao, M., Wan, Y., Meng, L., Cao, X., Wang, S.: Boundary-Aware Spatial and Frequency Dual-Domain Transformer for Remote Sensing Urban Images Segmentation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–18 (2024). https://doi.org/10.1109/tgrs.2024.3430081
21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017). https://doi.org/10.48550/arXiv.1612.01105
22. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021). https://doi.org/10.48550/arXiv.2012.15840
23. Zhu, F., Chen, Z., Li, H., Shi, Q., Liu, X.: Cedanet: Individual Tree Segmentation in Dense Orchard via Context Enhancement and Density Prior. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 7040–7051 (2024). https://doi.org/10.1109/jstars.2024.3378167