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. Accurate automated tree crown delineation (TCD) requires highly precise boundary segmentation, yet reduced-resolution architectures face limitations from decoder outputs at lower spatial resolutions. We propose BARE (Boundary-Aware with Resolution Enhancement), a simple architecture-preserving training strategy combining external full-resolution loss supervision with class weighting. BARE upsamples decoder output solely during training, maintaining inference efficiency while improving boundary precision. Through comprehensive evaluation on SegFormer, PSPNet, and SETR using the OAM-TCD dataset, we demonstrate that external full-resolution supervision universally benefits all tested architectures, achieving significant boundary quality improvements. We introduce B-IoU (Boundary-Intersection over Union) to TCD research, enabling rigorous boundary quality assessment. Our systematic evaluation reveals architecture-dependent optimization characteristics, providing actionable guidelines for practitioners seeking to enhance boundary precision in reduced-resolution segmentation architectures via training-only modifications. 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 downstream analyses in dense forest canopies where individual tree discrimination is paramount. Traditional computer vision approaches face particular

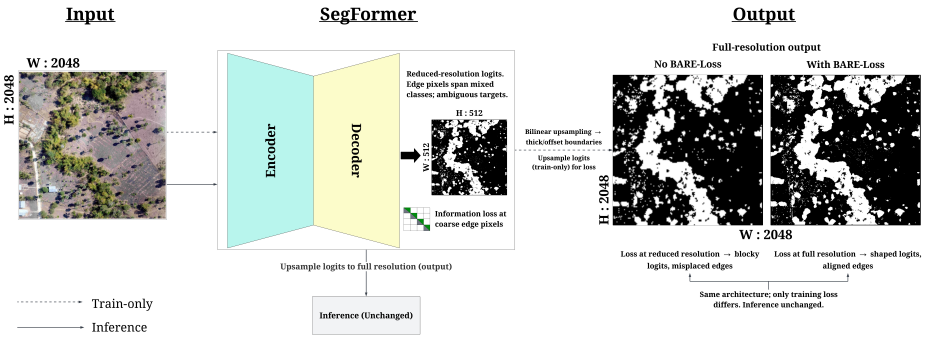


Fig. 1: Reduced-resolution decoder challenge.

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. Through systematic evaluation on the OAM-TCD dataset, we demonstrate that methodical optimization of training supervision can be more effective than architectural modifications for boundary-sensitive

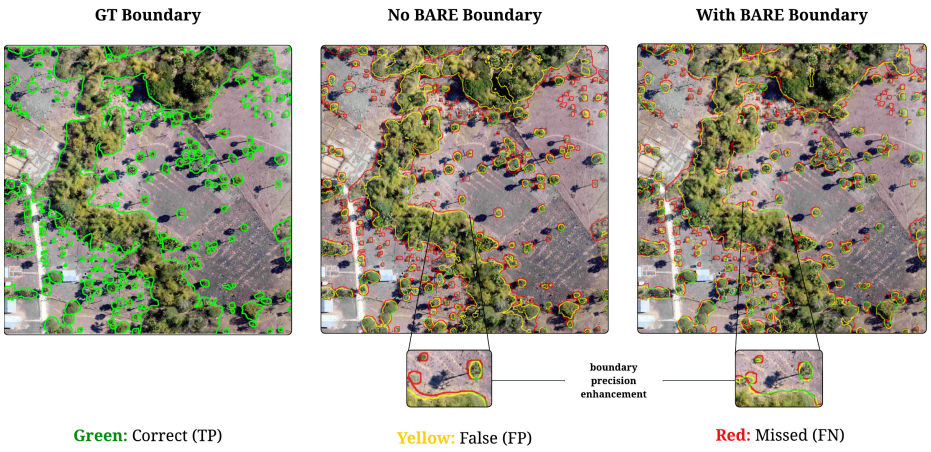


Fig. 2: B-IoU boundary precision evaluation.

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 universally benefits multiple reduced-resolution architectures (SegFormer, PSPNet, and SETR), while revealing that combining it with class weighting exhibits highly architecture-dependent effectiveness—only hierarchical transformers benefit from the complete approach, while pure-transformer and CNN-based architectures perform best with full-resolution supervision alone.

2 Related Work

We review the evolution toward Transformer-based TCD architectures, challenges in preserving spatial details, and existing approaches for high-resolution semantic segmentation and boundary refinement.

Deep learning for TCD has evolved from CNNs like U-Net [6,12] to self-attention models such as SegFormer [1,7,8,14]. Despite F1 scores exceeding ninty-five percent [8,16], precise boundary delineation remains challenging due to dense canopies [4,14,23], small instances [6], shadows [13], and overlapping crowns [20]. Advanced models often introduce significant computational costs [3,14], necessitating accuracy-efficiency trade-offs [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 spatial detail recovery through learned upsampling, multi-scale fusion, and attention mechanisms [10,11,19], though increasing complexity. Alternative approaches include interpolation-based upsampling and boundary refinement modules [15,18,20] at increased computational cost. Unlike dedicated boundary refinement modules adding inference overhead, BARE uses a training-only strategy to enhance boundary learning while preserving efficiency. Classical CNNs like PSPNet [21] face similar challenges through pyramid pooling at reduced resolution. We investigate external upsampling strategies requiring careful optimization for meaningful boundary improvements.

3 Methodology

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

3.1 BARE Framework Overview

BARE targets architectures producing decoder outputs at reduced resolution: 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 architectures, BARE applies two complementary training strategies. Panel (a) shows standard training with loss computed at reduced resolution, while panel (b) demonstrates BARE with external upsampling for full-resolution supervision. The approach applies across different reduction factors while maintaining inference efficiency through training-only modifications. BARE combines (1) external full-resolution supervision upsampling logits solely during training, and (2) class weighting addressing dataset imbalance. While external full-resolution supervision universally benefits all tested architectures, combining both components shows architecture-dependent effectiveness—only hierarchical transformers (SegFormer) benefit from the complete approach, while pure-transformer (SETR) and CNN-based (PSPNet) architectures show negative interactions.

3.2 Reduced-Resolution Processing Across Architectures

Contemporary segmentation architectures produce decoder outputs at reduced resolution for computational efficiency. For input $I \in \mathbb{R}^{H \times W \times 3}$, these architectures generate logits $L_{reduced} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times N_{cls}}$ where R is the reduction factor and N_{cls} the number of classes. The factor varies by architecture: SegFormer uses

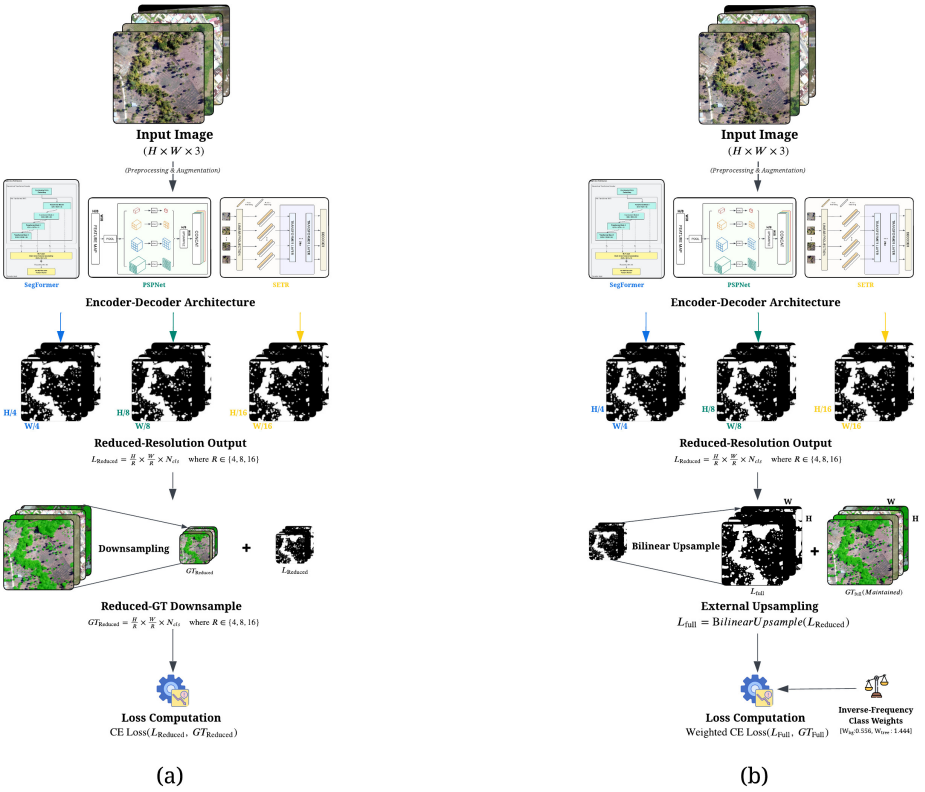


Fig. 3: BARE framework: (a) Standard vs (b) BARE training.

$R = 4$, PSPNet $R = 8$, and SETR $R = 16$. Standard training computes loss on $L_{reduced}$ or its upsampled version, potentially compromising boundary quality due to resolution mismatch with full-resolution ground truth masks.

3.3 External Full-Resolution Supervision Strategy

BARE's core component is external full-resolution supervision, bridging the resolution gap between decoder outputs and ground truth during training. As depicted in Figure 3(b), this strategy externally upsamples reduced-resolution logits for loss computation while maintaining the architecture unchanged. For input size $H \times W$ and logits $L_{reduced} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times N_{cls}}$ (R: 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 ground truth dimensions. This upsampling is applied during training only, preserving inference efficiency. Predicted

probabilities are computed via per-pixel softmax:

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

where i indexes pixels and c and j index classes.

3.4 Class Weighting Strategy

TCD datasets exhibit class imbalance, with tree crowns typically a minority class. BARE uses inverse-frequency class weights with unit-mean normalization. For a dataset with N_{ds} total pixels and n_c pixels in class c, the class frequency is $f_c = \frac{n_c}{N_{ds}}$ and the weight:

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

where N_{cls} is the number of classes (2 for binary TCD), j indexes classes, and f_j is the frequency of class j. This normalization keeps weights inversely proportional to frequency while ensuring unit mean. For OAM-TCD with 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

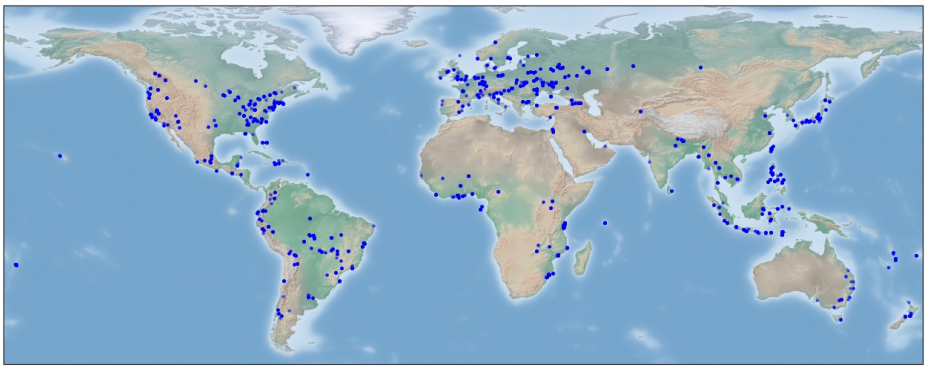
BARE employs Weighted Cross-Entropy (WCE) loss computed on full-resolution predictions. The complete loss formulation is:

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

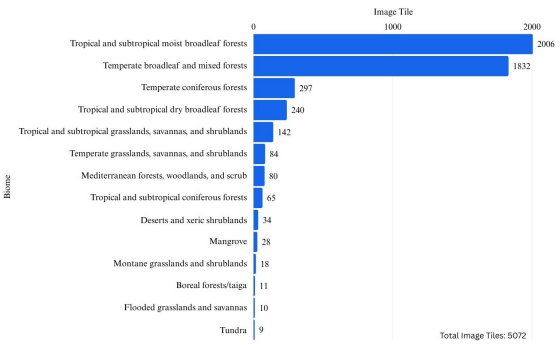
where $N = H \times W$ is the total pixels, N_{cls} the number of classes (2 for binary TCD), w_c the class weight from Equation 3, $y_{i,c} \in \{0, 1\}$ the ground truth label for pixel i and class c, and $P_{i,c}$ the predicted probability after softmax. For numerical stability, we add $\epsilon = 1 \times 10^{-7}$ to prevent $\log(0)$ undefined values while having negligible effect on typical probabilities. This formulation ensures boundary pixels receive full-resolution supervision while addressing class imbalance.

4 Experiments

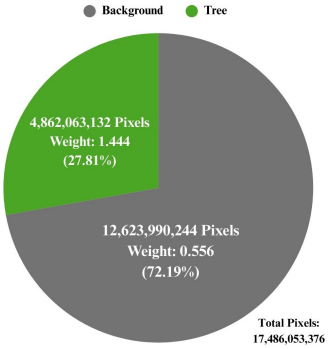
We present comprehensive experimental evaluation of BARE across multiple reduced-resolution architectures, beginning with dataset characteristics and experimental setup, followed by implementation details and results demonstrating BARE's effectiveness in improving boundary precision across different architectural paradigms.



(a) Global distribution [14]



(b) Biome diversity



(c) Class distribution

Fig. 4: OAM-TCD dataset characteristics.

4.1 Dataset and Experimental Setup

We use the OpenAerialMap Tree Crown Delineation (OAM-TCD) dataset [14], sourced from Hugging Face Hub as `restor/tcd`. The dataset provides global coverage across diverse forest environments spanning multiple continents and biomes, ideal for evaluating BARE's generalizability. Figure 4 shows 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. The binary segmentation task exhibits class imbalance with tree crowns representing 27.8% of training pixels, motivating inverse-frequency class weighting [0.556, 1.444] for background and tree crown classes.

We apply comprehensive augmentation exclusively to training. Geometric transformations include random crops to 1024×1024 pixels (maintaining full-resolution supervision compatibility), horizontal/vertical flips, and rotations up to ±180° preserving boundary integrity. Color augmentations include bright-

ness/contrast/saturation/hue jittering (P=0.7), Gaussian blur (P=0.5), and atmospheric/shadow effects (P=0.2 each) simulating real-world conditions without compromising edge quality. This maintains spatial dimensions for external full-resolution loss while enhancing robustness.

4.2 Implementation Details and Model Configurations

We evaluate BARE across three architectures: SegFormer (MiT-B5), PSPNet (ResNet-50), and SETR (ViT-base), using pre-trained ImageNet weights when available. BARE applies external upsampling to decoder outputs during training: SegFormer logits at $\frac{H}{4} \times \frac{W}{4}$, 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.

All models use consistent optimization settings: AdamW (lr=1×10^{-5}, weight decay=0.01), cosine annealing with 10% warmup, and FP16 precision. Due to memory constraints, batch configurations differ while maintaining effective batch size 8: SegFormer uses per-device batch 1 (eval batch 2) with 8 gradient accumulation steps, while PSPNet and SETR use larger per-device batches with 1 accumulation step. Training spans 50 epochs on NVIDIA L4 GPUs using PyTorch 2.6.0 and Transformers 4.51.3. Class weights [0.556, 1.444] address 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 for TCD. We employ B-IoU [2]—to our knowledge, its first application in TCD research. B-IoU evaluates segmentation within narrow bands around contours, providing direct boundary precision measurement complementing traditional pixel-wise metrics.

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 are boundary regions—pixels within distance d from ground truth and prediction contours. Unlike standard IoU treating all pixels equally, B-IoU focuses on boundary regions, making it more sensitive to boundary quality—particularly valuable for TCD where precise crown separation is critical. We set d to 2% of image diagonal following established guidelines [2], corresponding to approximately 15 pixels for 2048×2048 images. This captures meaningful boundary errors while remaining robust to minor annotation ambiguities.

Table 1: Performance metrics across configurations and architectures.

Config	Architecture	IoU	F1	Prec	Rec	Acc	B-IoU
Baseline	SETR	0.881	0.937	**0.948**	0.926	**0.970**	0.643
Baseline	SegFormer	0.817	0.899	0.931	0.870	0.953	0.590
Baseline	PSPNet	0.743	0.853	0.847	0.859	0.928	0.415
CW	SETR	0.879	0.935	0.908	0.964	0.968	0.627
CW	SegFormer	0.845	0.916	0.884	0.950	0.959	0.606
CW	PSPNet	0.727	0.842	0.775	0.922	0.916	0.399
Full-Res	SETR	**0.882**	**0.938**	**0.948**	0.927	**0.970**	**0.644**
Full-Res	SegFormer	0.828	0.906	0.931	0.882	0.957	0.610
Full-Res	PSPNet	0.743	0.852	0.852	0.853	0.928	0.418
BARE	SETR	0.879	0.936	0.908	**0.965**	0.968	0.625
BARE	SegFormer	0.848	0.918	0.885	0.953	0.960	0.620
BARE	PSPNet	0.728	0.843	0.776	0.922	0.916	0.404

We also evaluate using standard segmentation metrics: IoU, F1-Score (Dice), Precision, Recall, and Pixel Accuracy. These provide complementary perspectives, with IoU measuring region overlap, F1-Score offering balanced precision-recall assessment, and Precision/Recall quantifying exactness and completeness.

4.4 Results and Analysis

Performance Analysis. Table 1 presents comprehensive evaluation across three architectures and training configurations, revealing architecture-dependent effectiveness of BARE strategies.

SegFormer shows strongest response to complete BARE, with B-IoU improving 5.1% from 0.590 to 0.620 and IoU from 0.817 to 0.848. Notably, SegFormer alone benefits from combining full-resolution supervision with class weighting. PSPNet reveals that full-resolution supervision alone (B-IoU: 0.418) outperforms BARE (B-IoU: 0.404), indicating class weighting creates negative interactions—reducing boundary quality 3.3% despite similar overall IoU. SETR exhibits similar behavior: full-resolution supervision alone achieves optimal performance (IoU: 0.882, B-IoU: 0.644), while adding class weighting reduces performance (IoU: 0.879, B-IoU: 0.625).

These results demonstrate that external full-resolution supervision universally benefits all architectures, but combining it with class weighting shows architecture-dependent behavior. PSPNet and SETR achieve optimal boundary quality with full-resolution supervision alone—class weighting degrades performance. Only SegFormer's hierarchical transformer design benefits from complete BARE. This highlights the importance of empirical validation when combining training optimizations, as optimal configurations vary architecturally—pure-transformer and CNN-based models generally prefer full-resolution supervision alone, while hierarchical transformers benefit from the combined approach.

Qualitative Analysis. Figure 5 shows detailed visual comparisons demonstrating BARE's effectiveness across SegFormer, PSPNet, and SETR, with cleaner

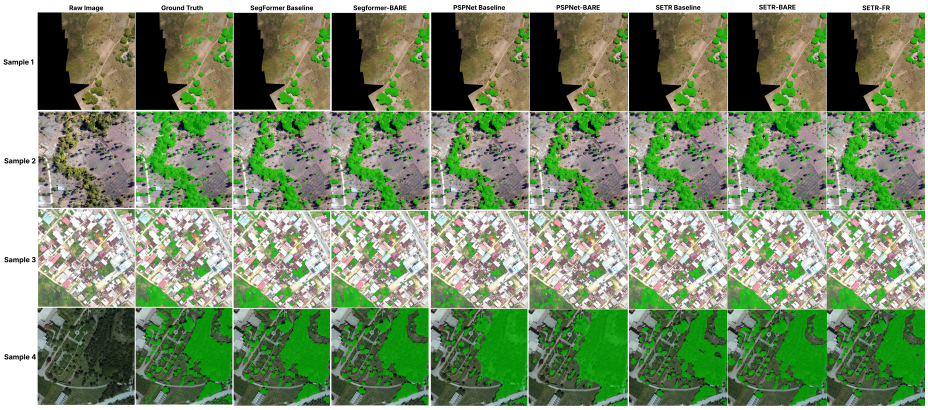


Fig. 5: Cross-architecture prediction comparison.

boundaries and reduced artifacts versus baselines. BARE variants show enhanced boundary definition across diverse landscapes: rural (Sample 1), dense forest (Sample 2), urban mixed (Sample 3), and agricultural (Sample 4). The analysis reveals systematic boundary coherence improvements, particularly in challenging scenarios with dense canopies and complex urban-forest interfaces.

Figure 6 shows error patterns where green indicates correct predictions, yellow shows false positives, and red shows missed detections. BARE variants exhibit reduced false positive artifacts and improved boundary precision with balanced error patterns. The systematic reduction in artifacts demonstrates BARE's effectiveness across all architectures.

Figure 7 shows boundary-focused validation where green indicates correct boundaries. Progressive improvement from baseline to BARE variants is evident across all architectures, with BARE achieving superior boundary coherence, particularly in complex boundary regions.

Computational Efficiency Analysis. BARE maintains efficiency by applying modifications only during training. Inference times remain largely unchanged (SegFormer: 274ms to 290ms, PSPNet: 195ms to 201ms, SETR: 184ms). Parameter counts stay constant (84.6M SegFormer, 49.0M PSPNet, 92.1M SETR), as BARE adds no parameters. Memory increases are minimal (SegFormer: 14.4GB to 14.7GB, PSPNet: 12.1GB to 12.3GB), making BARE practical for accuracy-constrained and efficiency-constrained deployments.

4.5 Discussion and Implications

Our evaluation reveals that external full-resolution supervision universally improves boundary precision across tested architectures through training-only modifications.

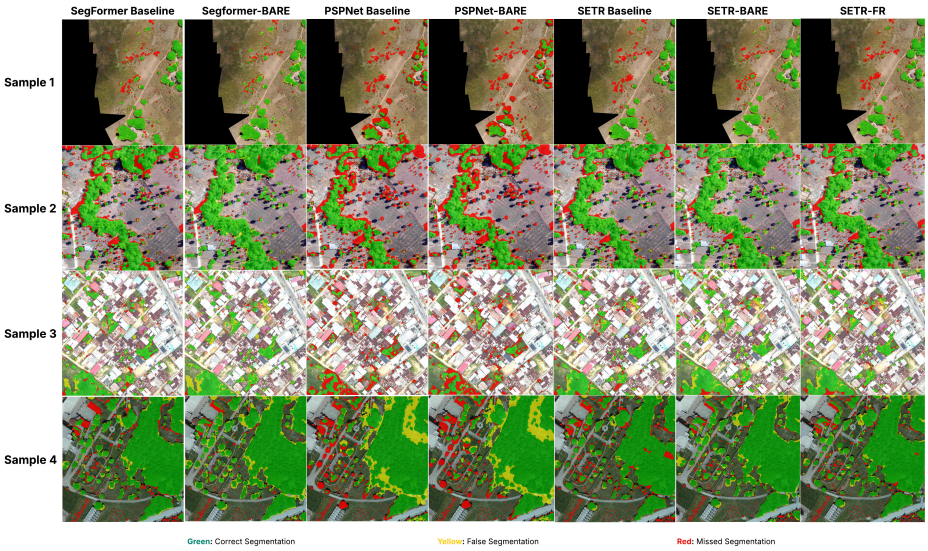


Fig. 6: Error pattern analysis across architectures.

Architecture-Dependent Effectiveness. BARE's two components show different generalization characteristics. External full-resolution supervision universally benefits all architectures, with SETR Full-Res achieving highest performance (IoU: 0.882, B-IoU: 0.644) and PSPNet Full-Res showing strongest boundary improvement (B-IoU: 0.415 to 0.418). However, combining full-resolution supervision with class weighting shows architecture-dependent behavior: only SegFormer benefits from complete BARE, while PSPNet (B-IoU: 0.418 to 0.404) and SETR (B-IoU: 0.644 to 0.625) experience negative interactions. This suggests pure-transformer and CNN-based architectures achieve near-optimal precision-recall balance with full-resolution supervision alone, making class weighting counterproductive. Only hierarchical transformer design (SegFormer) benefits from the combined approach, likely due to its multi-scale feature fusion requiring explicit class balance guidance.

Practical Implications. These findings provide actionable guidance: (1) apply external full-resolution supervision broadly as it universally benefits all tested models, (2) use class weighting cautiously—only hierarchical transformers (SegFormer) benefit from combining it with full-resolution supervision, while pure-transformer (SETR) and CNN-based (PSPNet) architectures perform best with full-resolution supervision alone, and (3) for pure-transformer and CNN-based architectures, avoid adding class weighting unless empirical validation shows benefits. The training-only nature makes empirical validation computationally feasible.

Limitations. Our evaluation uses only OAM-TCD; generalizability across remote sensing domains requires validation. The study examines three architectural paradigms; effectiveness on newer transformers (e.g., Swin) remains un-

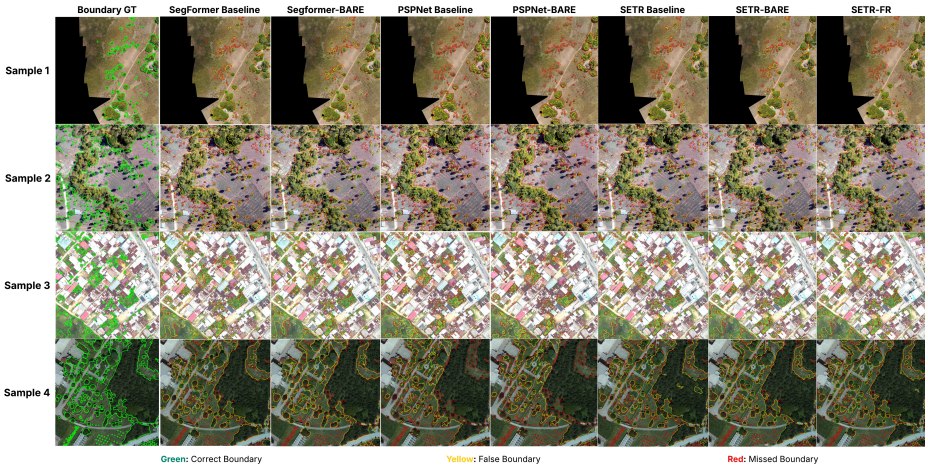


Fig. 7: Boundary-specific analysis results.

explored. We employ only bilinear upsampling; learned upsampling may yield different results. While maintaining effective batch size 8, memory constraints necessitated different gradient accumulation strategies, which could theoretically influence dynamics. Finally, deeper analysis of mechanisms driving architecture-specific responses would strengthen understanding.

Conclusion

We introduce BARE, an architecture-preserving training strategy combining external full-resolution supervision with class weighting to enhance boundary precision in tree crown delineation. Evaluation results reveal that external full-resolution supervision consistently benefits all tested architectures (SegFormer, PSPNet, SETR), achieving boundary IoU improvements, while class weighting shows architecture-dependent behavior—benefiting only hierarchical transformers (SegFormer) but degrading performance in pure-transformer (SETR) and CNN-based (PSPNet) 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 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 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 semantic 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 Conference 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