Grenze International Journal of Engineering and Technology, January Issue

Multi-Modal Transformer Framework for Accurate Detection and Counting of Standing Dead Trees in Forest

C.Roopa[1], K. Suvalakshmi[2], B. Bizu[3], V.M. Jashwanth[4], C.J. Jayavijaay[5] and D. Mayilanandhan[6]
[1-6]Department of CSE, Kongu Engineering College, Erode, India
Email: roopavivin@gmail.com, suva1798@gmail.com, bizucse@gmail.com, jashwanthvm.22cse@kongu.edu, jayavijaaycj.22cse@kongu.edu, mayilanandhand.22cse@kongu.edu

***Abstract*— Forest health monitoring is critical for ecological stability and carbon management. Dead trees serve as indicators of forest decline and increase the risk of wildfires. Traditional surveys are costly and limited in scale, making automated aerial detection vital. This paper presents a Multi-Modal Transformer Framework integrating RGB and Near Infrared data for accurate detection and counting of standing dead trees. The system employs three deep learning architectures U-Net, Feature Pyramid Network and U-Net++ with modality-specific encoders and transformer fusion layers to jointly optimize segmentation and counting tasks. Experimental evaluation using the Aerial Imagery for Standing Dead Tree Segmentation dataset demonstrates that the U-Net++ (ResNet-50) model achieved superior performance with an F1-score of 0.518, highlighting its capability for scalable and accurate forest health assessment.**

***Keywords*— Dead Tree Detection, Deep Learning, UAV Imagery, Transformer Networks, Semantic Segmentation, Multi-Modal Fusion, Environmental Monitoring.**

I. INTRODUCTION

Forests play a vital role in maintaining global ecological balance, yet tree mortality from climate stress, pests and wildfires poses serious challenges. Detecting standing dead trees accurately enables timely forest management and conservation planning. Traditional field surveys are slow and limited, while remote sensing using drones or UAVs provides large-scale imagery for automated analysis. However, identifying dead trees from aerial images remains difficult due to occlusion, varying canopy density, and spectral similarity. Deep learning-based semantic segmentation has shown promise [7], but single-modal approaches often fail to capture fine-grained and contextual features. This paper introduces a transformer-based multi-modal framework combining RGB and NIR data for accurate dead tree detection and counting. The architecture leverages attention mechanisms to fuse features across modalities [2],[4],[6] achieving improved performance in complex forest environments.

II. LITERATURE REVIEW

Recent research highlights deep learning's success in remote sensing segmentation [3], [7]. ADA-Net employed contrastive learning for multispectral imagery adaptation.

Grenze ID: 01.GIJET.12.1.253

Dual-Task Learning U-Net used self-attention and watershed segmentation to enhance boundaries, achieving 41% accuracy improvement [12]. YOLOv4 + MobileNetV3 offered efficient UAV detection with 97.3% accuracy. MSSCN introduced spatial attention with Gaussian kernel labeling for improved precision. LLAM-MDCNet utilized multi-density connections for vegetation anomaly detection [16]. In 2024, Junttila et al. analyzed boreal forest canopy mortality in Finland using long-term satellite data, confirming increased canopy decline due to climate stress. Hence, our proposed system aims to bridge this gap by leveraging multiple modalities and self-attention fusion layers [12], [13].

III. Proposed System

The proposed system (Fig 4) integrates three core components: (1) modality-specific encoders, (2) transformer-based feature fusion, and (3) dual-task prediction heads for segmentation and counting.

A. Dataset Preparation

The dataset, sourced from Kaggle's 'Aerial Imagery for Standing Dead Tree Segmentation', contains UAV images and corresponding masks [9]. Images were resized to 256×256 pixels, normalized, and augmented using flips, rotations, and brightness adjustments [3].

B. Model Architecture

U-Net (Fig. 1) and U-Net++ (Fig. 2) are encoder–decoder networks with skip connections [7], [11]. FPN (Fig. 3) enhances multi-scale feature learning through pyramid fusion [13]. The proposed framework adds transformer blocks between encoder and decoder layers to capture long-range dependencies across modalities [2]. [6]. Feature fusion is achieved via self-attention mechanisms integrating RGB and NIR features into a unified latent space [4], [12].

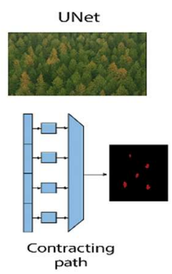

Figure 1. U-Net Architecture

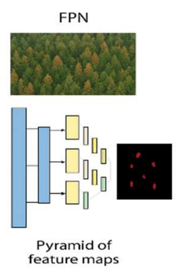

Figure 2. FPN Architecture

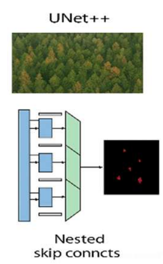

Figure 3. U-Net++ Architecture

C. Training and Optimization

Each model was trained using Binary Cross-Entropy (BCE) and Dice losses for segmentation, combined with L1 loss for counting accuracy [3]. Training was performed on GPU-enabled environments with Adam optimizer, learning rate 1e-4, batch size 8, and 12 epochs.

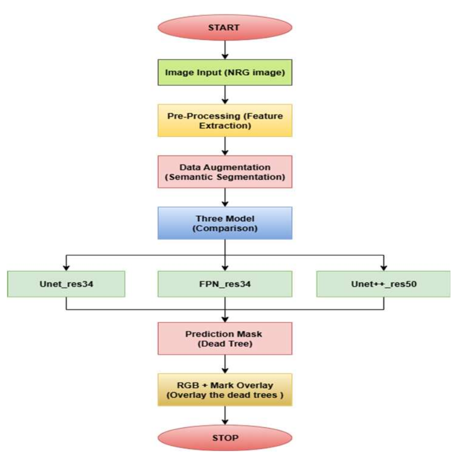

Figure 4. Flow of work

D. Dataset Description

The dataset used for this study is the Aerial Imagery for Standing Dead Tree Segmentation dataset sourced from Kaggle [9]. It consists of high-resolution aerial or UAV images of forested regions; each paired with pixel-level annotations identifying standing dead trees. The dataset was pre-processed to standardize image dimensions, normalize pixel values, and convert annotations into a binary mask format (dead tree = 1, background = 0). Data augmentation techniques, including rotations, flips, scaling, and brightness adjustments, were applied to increase the dataset size and diversity [3]. The data was then split into training (70%), validation (15%), and test (15%) subsets to ensure unbiased evaluation.

IV. RESULTS AND DISCUSSION

A. Parameters For Evaluation

The models were evaluated using multiple metrics commonly applied in segmentation tasks [3], [7]. These metrics (Fig. 5-8) provide a comprehensive assessment of the model's performance in accurately detecting and segmenting dead trees.

B. Validation Loss

Validation loss measures how well the model predicts the segmentation mask on unseen validation data during training. A lower validation loss indicates that the model's predictions closely match the ground-truth masks. Among the models, FPN_Res34 achieved the lowest validation loss (0.0585), followed by U-Net++_Res50 (0.0905) and UNet_Res34 (0.1152).

$$VAL\ LOSS = \frac{\text{Binary Cross Entropy} + \text{Dice Loss}}{2} \qquad (1)$$

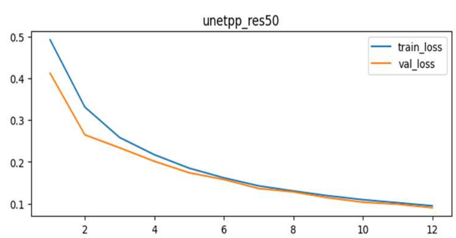

Figure 5. U-Net Architecture

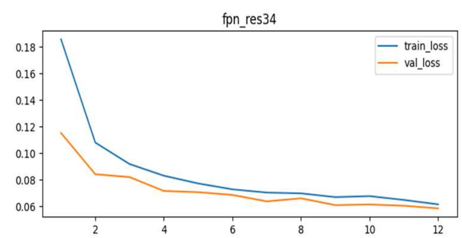

Figure 6. FPN Architecture

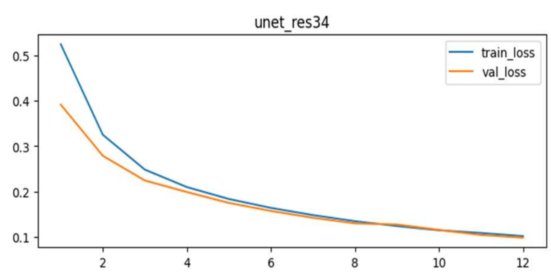

Figure 7. U-Net++ Architecture

C. Intersection over Union (IoU)

IoU quantifies the overlap between the predicted and ground-truth masks [7]. A higher IoU indicates better alignment. U-Net++_Res50 achieved the highest IoU (0.368), indicating its superior ability to localize dead trees in images compared to UNet_Res34 (0.313) and FPN_Res34 (0.288).

$$IOU = \frac{\text{TP}}{\text{TP+FP+FN}} \qquad (2)$$

D. Dice Coefficient

Dice coefficient measures similarity between predicted and true masks [3], emphasizing both false positives and false negatives. U-Net++_Res50 again outperformed other models with a Dice score of 0.518, highlighting its capability to capture detailed tree boundaries effectively.

$$Dice = \frac{2\,\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \qquad (3)$$

E. Precision and Recall

Precision represents the proportion of correctly predicted dead tree pixels among all predicted pixels, while recall measures the proportion of actual dead tree pixels correctly predicted. FPN_Res34 achieved the highest precision (0.645), suggesting it was highly confident in its positive predictions but had lower recall (0.342), indicating it missed many true dead tree pixels. U-Net++_Res50 achieved a balance between precision (0.642) and recall (0.482), making it more reliable overall.

$$Precision \;= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{F}} \qquad (4)$$

$$Recall \;= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{F}} \qquad (5)$$

F. F1 Score

The F1 score is the harmonic mean of precision and recall [3], providing a single measure of overall performance. U-Net++_Res50 achieved the highest F1 score (0.518), making it the best-performing model among the three evaluated. This indicates that it not only predicts dead tree pixels accurately but also captures a significant portion of all actual dead trees in the dataset.

$$F1\;Score = 2 * \left(\frac{\mathrm{Precision}+\mathrm{Recall}}{\mathrm{Precision}*\mathrm{Recall}}\right) \qquad (6)$$

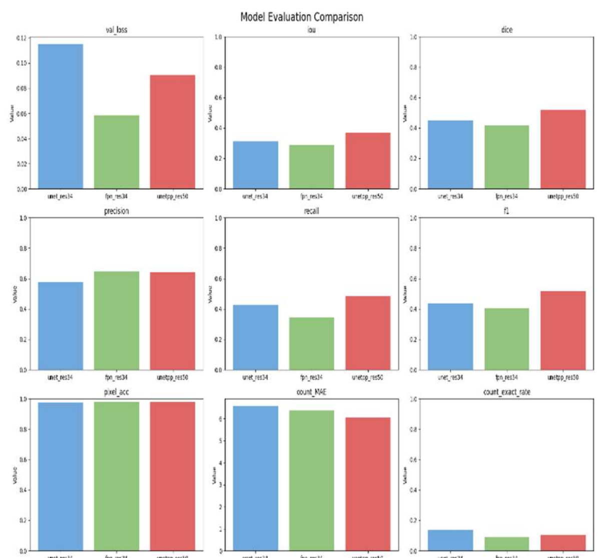

Figure 8. Evaluation metrics

G. Experimental Results

Performance was evaluated using metrics including Intersection over Union (IoU), Dice Coefficient, Precision, Recall, and F1-score [7]. Table I and Fig. 9 presents comparative performance of U-Net_Res34, FPN_Res34, and U-Net++_Res50. The U-Net++ model achieved the highest IoU and F1-score, proving its superior segmentation ability, particularly for small and occluded trees. Visual inspection indicates that U-Net++ effectively distinguishes between dead and healthy canopies with minimal noise. Transformer fusion improved the consistency of predictions across spectral domains [2], [6], [12], showing robustness under varying lighting and forest density

TABLE I. EVALUATION METRICS

Model	Val Loss	IoU	Dice	Precision	Recall	F1	Accuracy
UNet_Res34	0.1152	0.3132	0.4480	0.5766	0.4282	0.4366	0.9640
FPN_Res34	0.0584	0.2884	0.4170	0.6453	0.3421	0.4056	0.9478
UNet++_Res50	0.0904	0.3680	0.5178	0.6419	0.4822	0.5178	0.9776

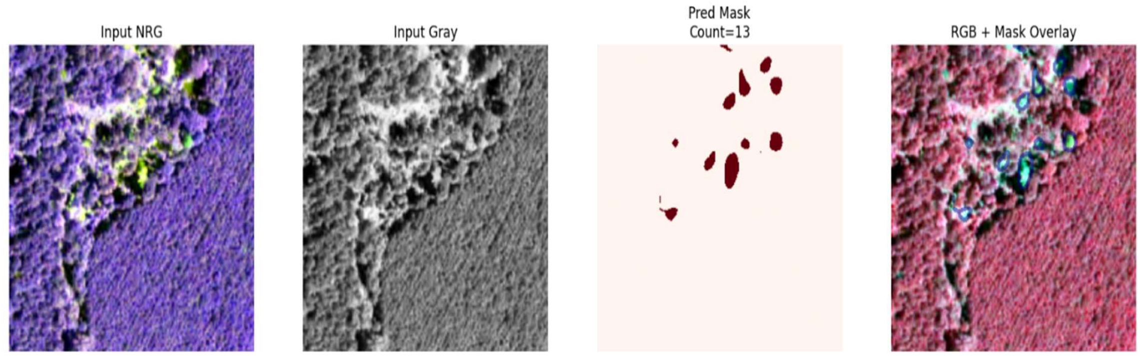

Figure 9. Result Structure

V. Conclusion and Future Work

This study developed and evaluated a Multi-Modal Transformer Framework for detecting and counting standing dead trees in UAV imagery. Among the compared architectures, U-Net++ (ResNet-50 backbone) demonstrated the best overall accuracy and segmentation fidelity. Future work includes expanding the dataset with hyperspectral imagery, incorporating attention modules [4], and deploying the model for real-time UAV forest monitoring. Integration with GIS systems and semi-supervised learning can further improve scalability and reduce annotation costs.

References

[1] Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Forster, Florencia Arias, ¨Jorn Hees, Beg ¨ um Demir, and Birgit Kleinschmit. Treesatai ¨ benchmark archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. Earth System Science Data Discussions, 2022:1–22, 2022.

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[3] Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Huttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke ¨ Schmeink, and Dorit Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. arXiv preprint arXiv:2312.05391, 2023.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1209–1218, 2018.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-toend object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818, 2018.

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Perpixel classification is not all you need for semantic segmentation. Advances in neural information processing systems,34:17864–17875, 2021.

[9] N Lakmal Deshapriya, Dan Tran, Sriram Reddy, and Kavinda Gunasekara. Centroid-unet: Detecting centroids in aerial images. arXiv preprint arXiv:2112.06530, 2021.

[10] Klaas Dijkstra, Jaap van de Loosdrecht, Lambert RB Schomaker, and Marco A Wiering. Centroidnet: A deep neural network for joint object localization and counting. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18, pages 585–601. Springer, 2019.

[11] Klaas Dijkstra, Jaap van de Loosdrecht, Waatze A Atsma, Lambert RB Schomaker, and Marco A Wiering. Centroidnetv2: A hybrid deep neural network for small-object segmentation and counting. Neurocomputing, 423:490–505,2021.

[12] Zhining Gu, Wenwen Li, Chia-Yu Hsu, Sizhe Wang, Yili Yang, Brendan M Rogers, and Anna Liljedahl. A multi- scale vision transformer-based multimodal geoai model for mapping arctic permafrost thaw. Available at SSRN 4762408.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Gir- shick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[14] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprintarXiv:1704.04861, 2017.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[16] Yu Ji, Enping Yan, Xianming Yin, Yabin Song, Wei Wei, and Dengkui Mo. Automated extraction of camellia oleifera crown using unmanned aerial vehicle visible images and the resu-net deep learning model. Frontiers in Plant Science, 13:958940, 2022.

[17] Jiange Jiang, Chen Chen, Yang Zhou, Stefano Berretti, LeiLiu, Qingqi Pei, Jianming Zhou, and Shaohua Wan. Heterogeneous dynamic graph convolutional networks for enhanced spatiotemporal flood forecasting by remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.