A MULTIMODAL ATTENTION-BASED MODEL FOR TREE SPECIES CLASSIFICATION USING LIDAR AND SATELLITE IMAGERY

Hadrien, Sablon & Rajen, Bajgain
Pacific Gas and Electric Company
Oakland, CA, USA

ABSTRACT

Accurate mapping of tree species is crucial for wildfire mitigation, biodiversity conservation, and sustainable forest management under climate change. While advances in remote sensing and deep learning have improved species classification, scarcity of high-quality ground truth data, low-resolution sensors, and small study areas with limited species diversity hinder scalability and generalization. To address these limitations, we assembled a dataset of half a million data points from five distinct level-III ecoregions in California. Ground truth labels across more than 20 species were obtained from arborist-supported tree inventories. We developed three deep learning models: a LiDAR-derived depth-view model (DVM) that exploits structural characteristics, a satellite-based Surface Reflectance model (SRM) that leverages spectral information, and a novel multimodal framework (MXAT) built with attention mechanisms to learn interdependencies between these complementary data modalities. Tested across 20 tree taxa, DVM achieves strong overall performance (mean sensitivity = 57.5%). In contrast, SRM exhibits limited predictive capabilities across most species but excels in identifying specific species such as Monterey Pine and Coast Redwood. Despite this performance gap, our findings show that LiDAR-based representations benefit substantially from integrating multispectral data: MXAT surpasses DVM baseline by nearly 5%. These results demonstrate the effectiveness of well-structured multimodal architectures in leveraging the complementary strengths of LiDAR and satellite imagery at scale.

1 INTRODUCTION, MOTIVATION AND SCOPE

Advancements in tree species identification using multi-scale data sources, such as LiDAR and multispectral reflectance, combined with deep learning, are proving invaluable for wildfire mitigation, forest management, and climate modeling Sun et al., 2019, Kwan et al., 2020, Hartling et al., 2019. To date, most studies focus on small-scale datasets, limiting model generalizability across diverse landscapes Allen et al., 2023, Ma et al., 2021, Fricker et al., 2019. While some models achieve high accuracy in specific ecoregions, performance declines in heterogeneous environments due to variations in species composition and canopy structure Fricker et al., 2019, Seidel et al., 2021. Expanding datasets across broader geographic areas is crucial for improving robustness, yet many studies remain constrained to specific species or forest types Lassalle et al., 2023, Welle et al., 2022, Blickensdörfer et al., 2024. Multimodal architectures show promise in developing scalable classification models trained on datasets spanning multiple ecoregions Kwon et al., 2023, Neyns et al., 2024. However, challenges remain in efficiently fusing diverse data sources Hussain et al., 2024, Wang and Zhang, 2018. Differences in resolution, scale, and dimensionality hinder optimal fusion, limiting the effectiveness of multimodal approaches. This study presents a deep learning multimodal framework that integrates airborne LiDAR data and satellite Surface Reflectance (SR) data across five level-III ecoregions in California.

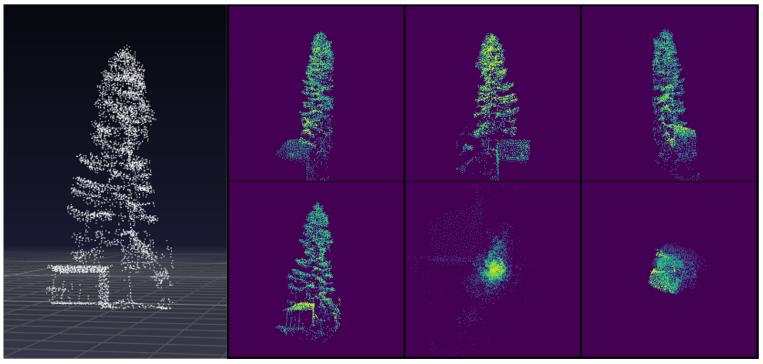

Figure 1: LiDAR point cloud and projected depth-views (4r1t1b). From top-left to bottom-right: 1st radial view to 4th radial view (4r), followed by top-down view (t) and bottom-up view (b)

2 DATASETS

We utilized the PlanetScope SuperDove SR basemaps with a resolution of 4.7 m, covering eight bands from blue to near infrared, and corrected for atmospheric effects to ensure spectral consistency across conditions. PlanetScope data have been used successfully in detecting canopy-scale tree mortality and survival via deep learning (Dixon et al., 2023). The August 2020 monthly basemap was selected to align with the LiDAR acquisition period (2019). Planet Labs' Usable Data Mask 2 (UDM2) was applied to retain only "clear sky" pixels.

We used airbone LiDAR data acquired by Sharper Shape and collected from April to November 2019, covering 26,000 miles across five level-III ecoregions in California. Individual tree point clouds were extracted by selecting high vegetation-classified points (12+ feet) within tree crowns delineated by Sharper Shape's segmentation algorithm.

Between 2019 and 2022, more than seven million trees were inventoried as part of a large-scale vegetation management initiative to improve electrical reliability and mitigate wildfire risk in California. Species and locations were documented using a GPS-enabled mobile application. To label our training and testing sets, we developed a strict proximity-based geospatial mapping algorithm to pair inventoried trees with LiDAR-detected counterparts. This approach, which prioritizes quality over coverage, yielded approximately 450,000 matched data points — 6.5% of the original dataset (Appendix A). Each labeled point was paired with its corresponding SR patch, centered on the LiDAR-derived treetop coordinates, forming the basis of our supervised training approach presented in Section 3. We then performed a region-based stratified split ensuring that each class contributed at least 1,000 samples or 20% of its data to the test set.

3 EXPERIMENTS AND MODEL ARCHITECTURE

We developed three deep learning models for tree species classification: (1) DVM, a Convolutional Neural Network (CNN)-based architecture for LiDAR-derived depth-views, (2) SRM, a lightweight CNN for PlanetScope Surface Reflectance imagery, and (3) MXAT, a multimodal model that integrates both modalities.

3.1 DVM: DEPTH-VIEW BASED MODEL

DVM is a deep learning framework based on the architecture introduced by SimpleView (Goyal et al., 2021), an approach leveraging six orthogonal camera projections (Figure 1) that has shown promising performance on limited tree point cloud samples (Allen et al., 2023). Our ablation experiments (Table 1b) indicate that while radial views are essential for achieving optimal performance, the bottom-up view does not contribute additional benefits. Additionally, resolution is a critical factor: performance improves as image resolution increases (Table 1a). The final model features a resolution of 256 as an effective trade-off between accuracy and computational cost. As expected, oversampling minority classes, combined with random scaling, rotation, and translation, significantly improves overall mean sensitivity (+10%).

Table 1: **DVM ablation** on the DMV dataset (slightly larger than MXAT), reporting mean sensitivity (MS, %). Default is 4r1t1b configuration at 256 resolution. MXAT settings are highlighted in gray.

Resolution	MS (%)
128	51.3
172	54.3
224	56.4
256	58
386*	**59.1**

(a) Resolution of depth-view projection. *with lower training batch size

DV configuration	MS (%)
4r1t1b	58
4r1t0b	**58.5**
4r0t0b	57.8
3r1t1b	54.3

(b) Depth-view (DV) configuration: default is 4 equally spaced radial views (4r), 1 top-down view (1t), and 1 bottom-up view (1b)

Table 2: **MXAT ablation experiment** reporting mean sensitivity (MS, %). For fully-connected post-fusion layers (a), we test different DV embedding treatment before fusion with SR embeddings. For transformer post-fusion blocks (b), we test cross-attention (X-att) and co-attention (Co-att) fusion mechanisms. MXAT settings are highlighted in **gray**.

Fusion type	DV pre-fusion treatment	
	Flatten	Global mean-pool
Add	61.3	61.6
Concat	61.4	61.6
X-att	**61.9**	61.4

(a) Post-fusion **fully-connected layers** (x2)

Fusion type	MS (%)
X-att	**62.2**
Co-att	62

(b) Post-fusion **transformer blocks** (x2)

3.2 SRM: Surface Reflectance-Based Model

SRM is a shallow CNN comprising three convolutional layers, followed by batch normalization, a dense layer, and a classification head (see Figure 2). The network processes 7×7 pixel patches extracted from Planet Lab's SR imagery, centered on LiDAR-derived tree tops, with the eight spectral bands serving as input channels. We assessed multiple patch sizes (3×3, 5×5, and 7×7) and observed a consistent performance improvement with increasing patch size: up to 2.6% from the 3x3 baseline. Accordingly, we adopted the 7×7 patch size in our final configuration and normalized each band before input to the SRM pipeline.

3.3 MXAT: Fusing LiDAR and Surface Reflectance Data within a Multimodal Architecture

To effectively integrate complementary information from LiDAR and satellite data, we developed a transformer-based multimodal fusion model. LiDAR depth views capture fine-grained 3D structural details, while spectral imaging provides crucial spectral signatures for species differentiation.
Building on recent advancements in transformer-based multimodal learning (Wadekar et al., 2024), we introduced an attention-driven fusion strategy, and evaluated it against multiple fusion approaches, including simple concatenation, element-wise operations, and co-attention-based mech-

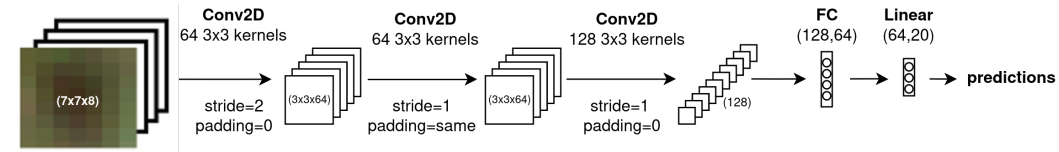

Figure 2: Overview of the proposed Surface Reflectance Model (SRM) architecture. Each Conv2D and fully connected (FC) layer is followed by batch normalization and a ReLU activation function.

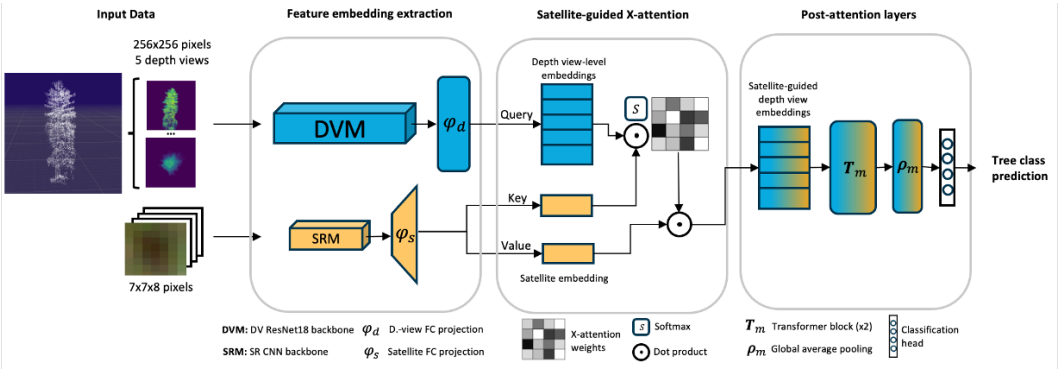

Figure 3: Overview of the proposed Multimodal Cross-Attention model (MXAT) architecture.

anisms. Experimental results indicate that attention-based fusion achieves incremental improvements over non-attentive methods (Table 4). Our best-performing model, MXAT, employs a cross-attention-based deep fusion framework. Depth-view and spectral embeddings are first processed independently before being integrated via cross-attention layers. Additional transformer blocks apply self-attention to refine the fused representation before final classification (Figure 3).

4 Results and Discussion

Table 3: Sample size per custom taxon and sensitivity % on testing set for DVM, SRM and MXAT. Best performance across models is highlighted in **gray**.

Tree custom taxa	Sample size (% total)	DVM (lidar)	SRM (satellite)	MXAT (multimodal)
Monterey pines	1.0	88	93	**97**
Ponderosa pines	16.3	**80**	10	**80**
Gray pines	6.0	87	52	**89**
Eucalyptus trees	1.1	81	72	**89**
True redwoods	3.8	88	78	**93**
Other oaks	14.7	61	24	**65**
Live oaks	13.0	53	20	**57**
Douglas firs	16.4	**45**	13	34
Incense cedars	5.8	**72**	13	68
Liquidambar trees	0.2	**63**	37	**63**
Sugar pines	0.8	62	42	**67**
Black oaks	6.4	58	39	**60**
Umbellularia trees	2.1	46	35	**59**
Walnuts	0.4	43	18	**50**
Tan oaks	2.7	46	60	**61**
Fir trees	1.0	37	22	**44**
Valley oaks	4.0	60	52	**63**
Poplars	0.2	30	9	**38**
Arbutus trees	2.0	22	12	**30**
Other	2.2	28	7	**39**
Mean sensitivity		**57.5**	**35.4**	**62.2**

All models were evaluated on a consistent testing set, focusing on classification accuracy for over 16,000 data points spanning 20 custom tree taxa (Table 3). The impact of sample size on classification performance deserves attention, as underrepresented classes tend to exhibit lower recall. For instance, poplars (recall = 38%), fir trees (recall = 44%) and arbutus trees (recall = 30%) accounts for 0.2%, 1%, and 2% of the overall dataset, respectively. The three models exhibit varying levels of

sensitivity across species (Appendix B contains the MXAT confusion matrix). DVM demonstrates robust overall performance, with five species surpassing 80% sensitivity and ten exceeding 60%. In contrast, SRM, trained exclusively on moderate-resolution satellite imagery, achieves limited overall accuracy but excels in detecting specific species such as Monterey pines (93%), true redwoods (78%), eucalyptus (73%), and, to a lesser extent, tan oaks (60%).

Despite this notable disparity in overall performance, we find that LiDAR-based representations significantly benefits from multispectral fusion. Specifically, MXAT surpassed the DVM baseline by nearly 5% in mean sensitivity across all predicted taxa. These results demonstrate that even with moderate-resolution imagery, synergy between LiDAR and satellite data can be actively exploited to considerably improve classification performance.

REFERENCES

[Allen et al.] Allen, M. J., Grieve, S. W. D., Owen, H. J. F., and Lines, E. R. Tree species classification from complex laser scanning data in mediterranean forests using deep learning. 14(7):1657–1667. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13981.

[Blickensdörfer et al.] Blickensdörfer, L., Oehmichen, K., Pflugmacher, D., Kleinschmit, B., and Hostert, P. National tree species mapping using sentinel-1/2 time series and german national forest inventory data. 304:114069.

[Dixon et al.] Dixon, D. J., Zhu, Y., Brown, C. F., and Jin, Y. Satellite detection of canopy-scale tree mortality and survival from california wildfires with spatio-temporal deep learning. 298:113842.

[Fricker et al.] Fricker, G. A., Ventura, J. D., Wolf, J. A., North, M. P., Davis, F. W., and Franklin, J. A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. 11(19):2326. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[Goyal et al.] Goyal, A., Law, H., Liu, B., Newell, A., and Deng, J. Revisiting point cloud shape classification with a simple and effective baseline. (arXiv:2106.05304).

[Hartling et al.] Hartling, S., Sagan, V., Sidike, P., Maimaitijiang, M., and Carron, J. Urban tree species classification using a WorldView-2/3 and LiDAR data fusion approach and deep learning. 19(6):1284. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[Hussain et al.] Hussain, M., O'Nils, M., Lundgren, J., and Mousavirad, S. J. A comprehensive review on deep learning-based data fusion. 12:180093–180124. Conference Name: IEEE Access.

[Kwan et al.] Kwan, C., Ayhan, B., Budavari, B., Lu, Y., Perez, D., Li, J., Bernabe, S., and Plaza, A. Deep learning for land cover classification using only a few bands. 12(12):2000. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[Kwon et al.] Kwon, R., Ryu, Y., Yang, T., Zhong, Z., and Im, J. Merging multiple sensing platforms and deep learning empowers individual tree mapping and species detection at the city scale. 206:201–221.

[Lassalle et al.] Lassalle, G., Ferreira, M. P., Cué La Rosa, L. E., Del'Papa Moreira Scafutto, R., and de Souza Filho, C. R. Advances in multi- and hyperspectral remote sensing of mangrove species: A synthesis and study case on airborne and multisource spaceborne imagery. 195:298–312.

[Ma et al.] Ma, M., Liu, J., Liu, M., Zeng, J., and Li, Y. Tree species classification based on sentinel-2 imagery and random forest classifier in the eastern regions of the qilian mountains. 12(12):1736. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[Neyns et al.] Neyns, R., Efthymiadis, K., Libin, P., and Canters, F. Fusion of multi-temporal PlanetScope data and very high-resolution aerial imagery for urban tree species mapping. 99:128410.

[Seidel et al.] Seidel, D., Annighöfer, P., Thielman, A., Seifert, Q. E., Thauer, J.-H., Glatthorn, J., Ehbrecht, M., Kneib, T., and Ammer, C. Predicting tree species from 3d laser scanning point clouds using deep learning. 12. Publisher: Frontiers.

[Sun et al.] Sun, Y., Huang, J., Ao, Z., Lao, D., and Xin, Q. Deep learning approaches for the mapping of tree species diversity in a tropical wetland using airborne LiDAR and high-spatial-resolution remote sensing images. 10(11):1047. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[Wadekar et al.] Wadekar, S. N., Chaurasia, A., Chadha, A., and Culurciello, E. The evolution of multimodal model architectures.

[Wang and Zhang] Wang, W. and Zhang, M. Tensor deep learning model for heterogeneous data fusion in internet of things. 4(1):32–41. Publisher: IEEE.

[Welle et al.] Welle, T., Aschenbrenner, L., Kuonath, K., Kirmaier, S., and Franke, J. Mapping dominant tree species of german forests. 14(14):3330. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

A Appendix A

The training dataset encompasses six distinct level-III ecoregions in Northern California, with significant representation across five of these ecoregions (see Figure 4).

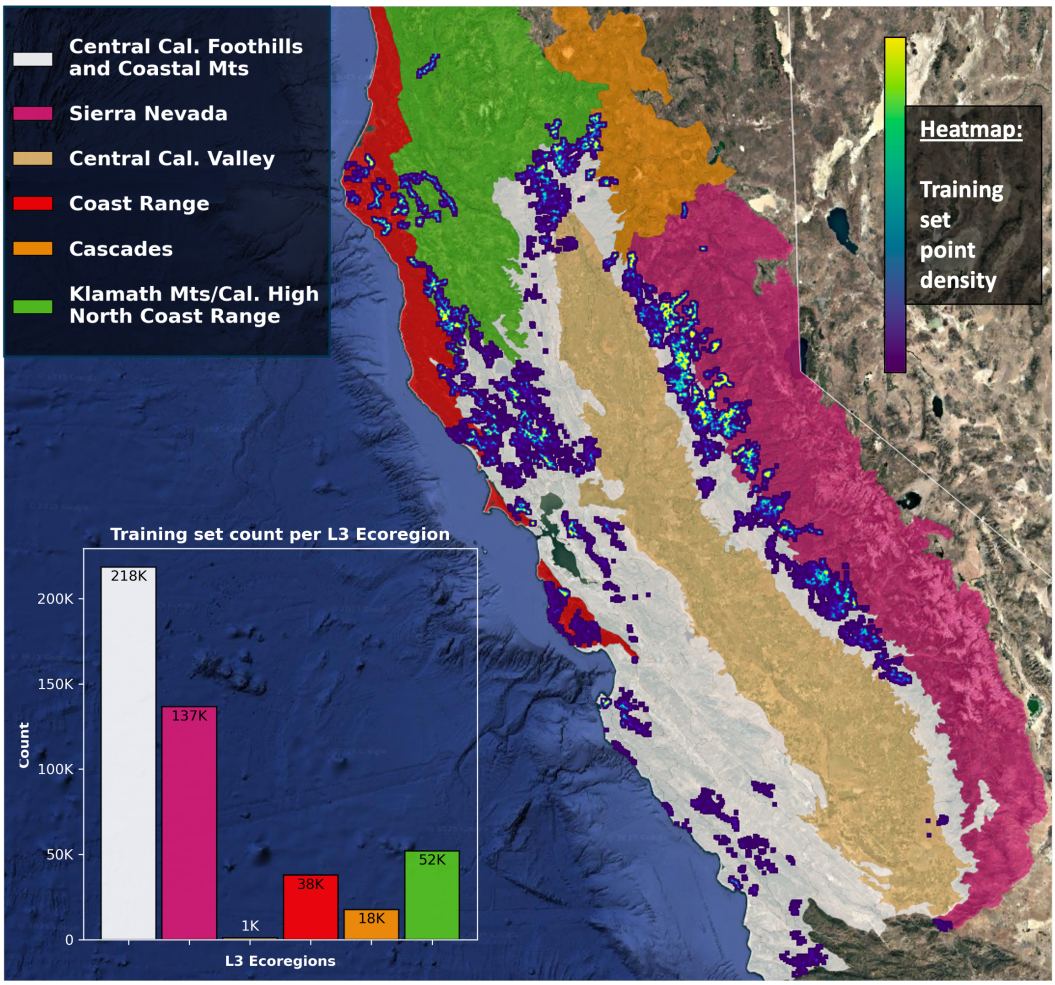

Figure 4: Training set distribution across California visualized as a heatmap, with embedded histogram detailing the training set distribution per level III ecoregion.

A APPENDIX B

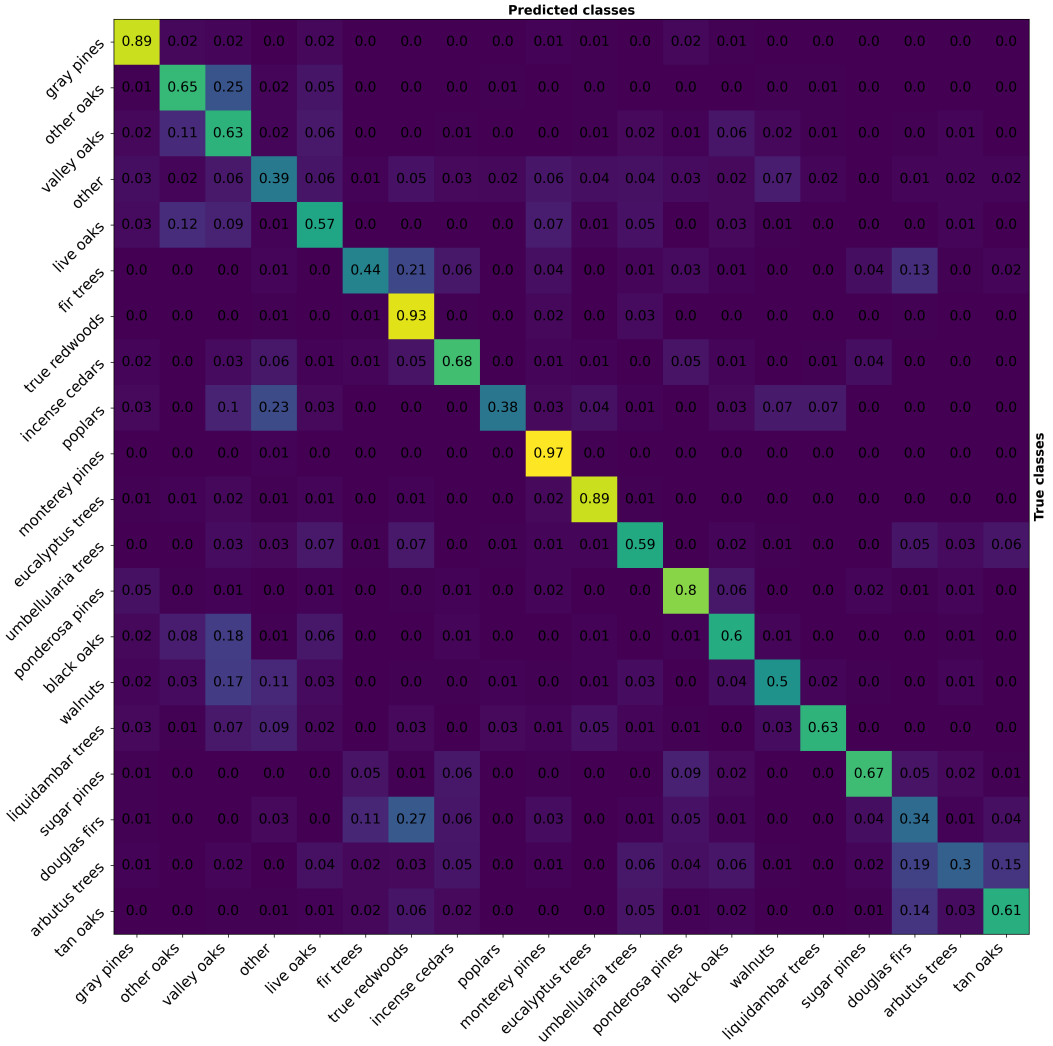

Figure 5: MXAT Confusion Matrix on test set (normalized)