Multi-Label Aerial Image Classification Using Transfer Learning: Progress Report

Tunahan Yazar Cansu Temizkan Defne Koçulu Yavuz Can Atalay
CS415 - Deep Learning
Sabancı University
Date: 23 November 2025

***Abstract*—This report presents our work on multi-label aerial image classification using the AID_MultiLabel dataset. We implemented a transfer learning approach using EfficientNet-B4 as the backbone architecture, achieving a macro F1 score of 0.8097 and micro F1 score of 0.9165 on the test set. Our approach addresses the challenges of multi-label classification in aerial imagery, including class imbalance, label co-occurrence, and the need for rotation invariant feature extraction. This report details our theoretical design decisions, experimental methodology, results, and future directions for improvement.**

I. Introduction and Motivation

A. Problem Definition

Multi-label aerial image classification is the task of assigning multiple semantic labels to overhead imagery captured from aerial or satellite platforms. Unlike traditional single-label classification where each image belongs to exactly one category, aerial scenes typically contain multiple objects simultaneously (e.g., an airport image may contain airplanes, buildings, pavement, and cars). This multi-label nature makes the task significantly more challenging than conventional image classification.

B. Motivation

The accurate multi-label classification of aerial imagery serves a pivotal role across diverse sectors, including urban planning, where it facilitates automated land use analysis and infrastructure monitoring. In the realm of environmental monitoring, these systems are essential for tracking deforestation, water bodies, and land degradation, while also proving critical in disaster response scenarios by enabling rapid damage assessment and resource allocation. Additionally, applications extend to agriculture for crop monitoring and precision farming, as well as defense for intelligence gathering and surveillance. However, the inherent complexity of aerial scenes, characterized by high intra-class variation and inter-class similarity, necessitates the deployment of advanced deep learning approaches capable of effectively modeling label dependencies and extracting discriminative features.

II. Related Work

A. Aerial Image Datasets

AID Dataset Foundation: Xia et al. [1] introduced the Aerial Image Dataset (AID), a large-scale benchmark containing 10,000 images across 30 scene categories collected from Google Earth imagery. The dataset was specifically designed to address the limitations of earlier aerial image datasets by providing higher intra-class diversity and inter-class similarity, making it more challenging and realistic for evaluating classification algorithms. The images are 600×600 pixels and cover diverse geographic locations and imaging conditions.

AID Multi-Label Dataset: Hua et al. [2] extended the AID dataset to create AID_MultiLabel, containing 3,000 images with 17 object-level labels. This dataset addresses the fundamental limitation that aerial scenes inherently contain multiple semantic categories. The authors proposed a Relation Network that models label dependencies through three modules: label-wise feature parcel learning, attentional region extraction, and label relational inference. Their work demonstrated that explicitly modeling label relationships significantly improves multi-label classification performance compared to treating labels independently.

B. Multi-Label Classification Methods

Deep Learning for Multi-Label Learning: The field of multi-label classification has evolved significantly with deep learning. Traditional approaches treated multi-label problems as multiple independent binary classification tasks, but this ignores valuable label correlations. Modern deep learning methods leverage CNNs for feature extraction combined with specialized mechanisms for capturing label dependencies [3].

Graph-Based Label Modeling: Chen et al. [4] introduced ML-GCN (Multi-Label Graph Convolutional Network), which constructs a directed graph over object labels where each node is represented by word embeddings. The GCN learns to map this label graph into inter-dependent object classifiers, enabling the model to exploit label co-occurrence patterns. Their approach achieved state-of-the-art results by using a novel re-weighted scheme to create an effective label correlation matrix. This work is particularly relevant for aerial imagery where certain labels frequently co-occur (e.g., harbor with water and ships).

C. Transfer Learning for Remote Sensing

ResNet and Deep Residual Learning: He et al. [5] introduced Deep Residual Learning with skip connections, enabling the training of very deep networks (50-152 layers) without degradation. ResNet architectures have become the foundation for transfer learning in computer vision, including remote

sensing applications. The residual connections allow gradients to flow directly through the network, mitigating the vanishing gradient problem and enabling effective feature learning.

Transfer Learning Challenges in Remote Sensing: Transfer learning from ImageNet-pretrained models to remote sensing domains presents unique challenges. Aerial imagery differs from natural images in perspective (overhead vs. ground-level), scale variations, and spectral characteristics. Despite these differences, research has shown that transfer learning significantly outperforms training from scratch, particularly when labeled aerial data is limited. Fine-tuning strategies that adapt pre-trained features to the aerial domain have proven effective [6].

D. Handling Class Imbalance

Binary Cross-Entropy for Multi-Label: Standard multi-label classification employs Binary Cross-Entropy (BCE) loss, which treats each label as an independent binary classification problem. The loss is computed as:

$$BCE = -[y \cdot \log(\sigma(x)) + (1 - y) \cdot \log(1 - \sigma(x))] \quad (1)$$

where σ is the sigmoid function, y is the binary ground truth, and x is the model's logit output. This formulation is suitable for multi-label scenarios because sigmoid outputs are independent (unlike softmax), allowing multiple labels to have high probabilities simultaneously.

Advanced Techniques: Secondly, weighted sampling is implemented to handle class imbalance. For this, we first quantified the global distribution of the dataset to derive class specific weights inversely proportional to their occurrence frequency, thereby assigning higher importance values to underrepresented labels. For every training image, a scalar sampling weight was computed as the arithmetic mean of the inverse-frequency weights associated with its ground-truth labels. This mechanism directs the data loader to sample instances with replacement based on these calculated probabilities, effectively increasing the representation of rare classes within each mini-batch and preventing the optimization process from converging toward a local minimum dominated by majority background categories.

Other Techniques Tried: Lastly, we tried focal loss but this was not helpful in terms of getting better macro f1 scores.

E. Research Gap

While significant progress has been made in multi-label classification, most state-of-the-art methods focus on natural images, often failing to account for the distinct complexities of remote sensing. Aerial imagery presents unique challenges, most notably the requirement for rotation invariance, as overhead views lack a canonical orientation and can be captured from any angle. Furthermore, these datasets typically exhibit significant class imbalance, typified by the disparity between ubiquitous structures like buildings and rare objects like mobile homes, as well as extreme scale variations where objects appear at differing resolutions. Finally, aerial scenes differ from natural images due to specific label co-occurrence patterns inherent to geographic layouts. Our work addresses these challenges through rotation invariant augmentation, over-sampling strategies, and transfer learning with modern CNN architectures optimized for efficiency.

III. Dataset Description

A. AID_MultiLabel Dataset

We utilize the AID_MultiLabel dataset [2], which is derived from the standard AID benchmark through the addition of manual multi-label annotations. This dataset comprises 3,000 high-resolution aerial images (600×600 pixels) covering 17 distinct object-level categories: airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water. Reflecting the complex semantic content of overhead imagery, the dataset features a dense label distribution with an average of 3.2 labels per image, ranging from a single label to a maximum of seven concurrent categories.

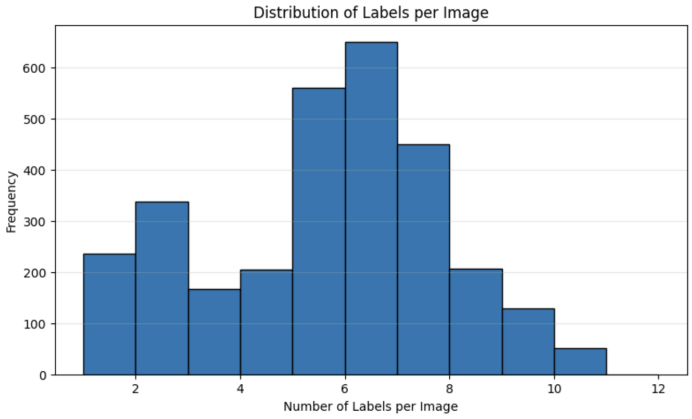

Fig. 1. Distribution of the Number of Labels per Image (AID Multi-Label Dataset).

B. Dataset Statistics

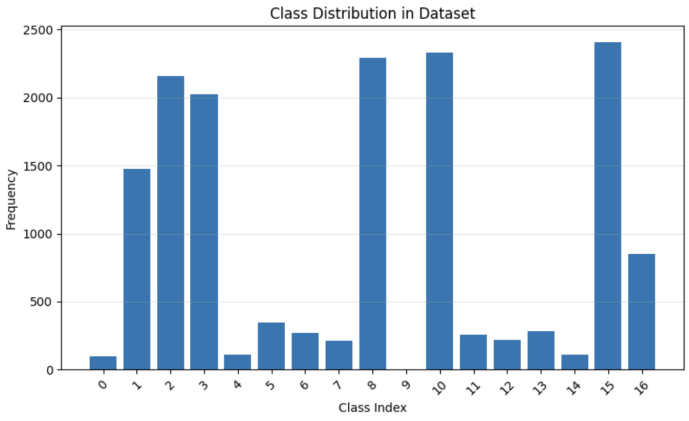

Fig. 2. Class Frequency Distribution Across 17 Object Categories.

Our analysis of the dataset revealed significant class imbalance across the 17 object categories. The distribution is dominated by common features, with Trees and Pavement appearing

most frequently, in 1,018 (33.9%) and 1,017 (33.9%) images respectively, followed closely by Grass (32.6%), Buildings (30.7%), and Cars (28.6%). In sharp contrast, specific object categories are severely underrepresented; Mobile home is the rarest class with only 29 occurrences (0.97%), while Airplane (1.23%), Chaparral (1.37%), Tanks (1.63%), and Sea (3.07%) also display very low frequencies. This disparity results in an approximate 35:1 imbalance ratio between the most and least frequent classes, presenting a significant challenge for training balanced models.

This 35:1 imbalance ratio between the most and least frequent classes presents a significant challenge for training balanced models. The class distribution analysis motivated our decision to implement weighted sampling strategies.

C. Dataset Split

We divided the dataset as follows: training set as 2,100 images (70%), validation set as 450 images (15%), and lastly, test set as 450 images (15%). The stratified split ensures similar label distributions across all sets, with random shuffling (seed=42) for reproducibility.

IV. Methodology

A. Overall Architecture

Our approach follows a transfer learning paradigm with the following pipeline:

Input Image (600×600 RGB)
$\downarrow$
Data Augmentation & Preprocessing
$\downarrow$
EfficientNet-B4 Backbone (Pre-trained on ImageNet)
$\downarrow$
Global Average Pooling
$\downarrow$
Classification Head (FC: $1536 \to 512 \to 17$)
$\downarrow$
Sigmoid Activation
$\downarrow$
Multi-Label Predictions

B. Model Architecture

1) Backbone Selection: EfficientNet-B4: We selected EfficientNet-B4 as our backbone architecture, driven by its superior balance of computational efficiency and feature extraction capability. Central to this decision is the model's utilization of compound scaling, a technique that uniformly scales network width, depth, and resolution to achieve higher efficiency than traditional methods of scaling individual dimensions [7]. In terms of model complexity, EfficientNet-B4 demonstrates remarkable parameter efficiency, achieving competitive accuracy with only 19 million parameters compared to ResNet50's 25 million, rendering it significantly more suitable for deployment constraints. Furthermore, this architecture has consistently outperformed ResNet baselines on ImageNet and transfer learning benchmarks, particularly when fine-tuned on limited datasets. Finally, the model provides high feature richness; its deeper architecture generates 1536-dimensional representations that, despite being more compact than ResNet50's 2048 dimensions, offer more expressive features for the classification task.

Architecture Details: The implementation utilizes the EfficientNet-B4 architecture sourced from the `timm` library, initialized with weights pre-trained on the ImageNet-1K dataset. This backbone extracts 1536-dimensional feature vectors, which are subsequently processed by a custom classification head. The head is structured as a multi-stage sequence: it begins with a dropout layer ($p = 0.25$) to mitigate overfitting, followed by a linear transformation mapping the 1536 input features to a 512-dimensional hidden layer equipped with ReLU activation. A second dropout layer ($p = 0.25$) is applied before the final linear layer, which projects the hidden features onto the 17 output units corresponding to the target classes.

2) Multi-Label Classification Head: The multi-label classification head is structured to balance model capacity with regularization. It initiates with a dropout layer ($p = 0.25$) to prevent overfitting by randomly deactivating neurons during the training phase. This is immediately followed by a dense hidden layer of 512 units, which provides the network with additional non-linear transformation capacity. The architecture culminates in an output layer comprising 17 units, one for each target class, that produces raw logits. A pivotal design choice for this final layer is the application of sigmoid activation instead of softmax. Since the target labels are non-mutually exclusive, the sigmoid function is required to compute class probabilities independently, thereby enabling the simultaneous prediction of multiple high-confidence labels for a single input.

C. Training Strategy

1) Loss Function: For the optimization objective, we employ Binary Cross-Entropy with Logits (BCEWithLogitsLoss). This formulation integrates the sigmoid activation function and the Binary Cross-Entropy loss into a single, numerically stable operation. By treating each label as an independent binary classification problem, this loss function is uniquely suited for non-mutually exclusive multi-label scenarios, allowing the model to effectively learn multiple concurrent category assignments for a given input without the constraints imposed by softmax-based objectives.

2) Addressing Class Imbalance: Weighted Sampling: To mitigate the severe 35:1 class imbalance, we implemented a weighted random sampling strategy. This process begins by calculating rarity weights for each class, defined as:

$$weight_j = 1/(frequency_j + \epsilon) \tag{2}$$

where the weight is inversely proportional to the label's frequency. Subsequently, a specific sampling weight is assigned to each image, computed as the average weight of its positive labels. These weights guide the `WeightedRandomSampler`, which ensures that images containing rare labels are selected more frequently during training. This strategy successfully increases exposure to

minority classes without discarding majority class samples, effectively helping the model learn balanced representations.

3) Optimization and Training Configuration: The network optimization was performed using the Adam optimizer, initialized with a learning rate of 0.001 and a weight decay of 0.0001 for L2 regularization, while maintaining standard β parameters ($\beta_1 = 0.9, \beta_2 = 0.999$). To ensure stable convergence, we employed a `ReduceLROnPlateau` scheduler that dynamically reduces the learning rate by a factor of 0.5 upon observing three consecutive epochs of validation loss stagnation, with a lower bound set at 1×10^{-5}. The training procedure was executed on an NVIDIA L4 GPU with a batch size of 32 for a maximum of 30 epochs, incorporating an early stopping mechanism with a patience of 7 epochs to mitigate overfitting.

D. Data Augmentation

Aerial imagery has unique properties requiring specialized augmentation strategies:

1) Training Augmentation: Aerial imagery possesses unique characteristics that necessitate specialized augmentation strategies, particularly regarding perspective. To induce rotation invariance, critical since overhead imagery lacks a canonical orientation, we applied random $\pm 90^\circ$ rotations alongside random horizontal and vertical flips, each with a probability of $p = 0.5$. Geometric diversity was introduced by resizing images to 256×256 followed by a random crop to 224×224, a process that effectively simulates scale variations. Furthermore, to ensure robustness against varying lighting conditions and atmospheric effects, we employed ColorJitter to perturb brightness, contrast, and saturation by $\pm 20\%$, and hue by $\pm 10\%$. Finally, all inputs were normalized using standard ImageNet statistics (Mean: $[0.485, 0.456, 0.406]$, Std: $[0.229, 0.224, 0.225]$) to facilitate transfer learning.

2) Validation/Test Augmentation: To ensure consistent and reproducible evaluation results, the validation and test pipelines employ a deterministic preprocessing strategy distinct from the training phase. In this protocol, no random augmentations are applied; instead, images are resized directly to 224×224 pixels without cropping to preserve the full spatial context of the scene. Subsequently, the data is normalized using the identical ImageNet mean and standard deviation statistics employed during training. This methodological distinction is critical: while strong augmentation during training enhances the model's generalization capabilities, deterministic preprocessing during evaluation guarantees that performance metrics remain stable and comparable across experimental runs.

E. Evaluation Metrics

Multi-label classification requires specialized metrics that account for partial correctness:

1) F1 Scores: Micro F1 (overall performance):

$$Precision_{micro} = \frac{TP_{all}}{TP_{all} + FP_{all}} \tag{3}$$

$$Recall_{micro} = \frac{TP_{all}}{TP_{all} + FN_{all}} \tag{4}$$

$$F1_{micro} = 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{5}$$

Aggregates true positives, false positives, and false negatives across all classes. Dominated by frequent classes.

Macro F1 (per-class average):

$$F1_{macro} = \frac{1}{C} \sum_j F1_j \tag{6}$$

Computes F1 for each class independently, then averages. Treats all classes equally regardless of frequency. Primary metric for imbalanced datasets.

Weighted F1 (frequency-weighted average): Weights each class's F1 by its support (number of true instances).

2) Other Metrics: Hamming Loss: Fraction of incorrectly predicted labels (lower is better)

$$Hamming\ Loss = \frac{1}{N \times C} \sum_i \sum_j [y_{ij} \neq \hat{y}_{ij}] \tag{7}$$

Subset Accuracy: Percentage of samples with all labels correct (strictest metric)

$$Subset\ Accuracy = \frac{1}{N} \sum_i [y_i = \hat{y}_i] \tag{8}$$

Per-Class Precision/Recall/F1: Individual class performance for identifying problematic categories.

V. Experiments and Results

A. Training Dynamics

1) Training Progress: The model was trained for 30 epochs with specific convergence behavior. Regarding the Loss Curves, the training loss decreased from 0.322 (epoch 1) to 0.023 (epoch 30). The validation loss decreased from 0.203 (epoch 1) to a minimum of 0.128 (epoch 9), then gradually increased to 0.208 (epoch 30). Overfitting was observed after epoch 10, as the validation loss increased while training loss continued decreasing.

In terms of F1 Score Evolution, the validation micro F1 improved from 0.857 (epoch 1) to 0.917 (epoch 23). The validation macro F1 improved from 0.502 (epoch 1) to 0.807 (epoch 23) [best]. Early stopping did not trigger (patience=7) because the macro F1 continued fluctuating. Finally, the Learning Rate Schedule began with an initial LR of 0.001. LR reduction was triggered at epochs 13, 16, 20, 24, and 27. The final LR was 0.000031 (after 5 reductions by a factor of 0.5).

2) Best Model Selection: The best model was selected based on validation macro F1 (not micro F1) to prioritize balanced performance across all classes. The model from epoch 23 achieved a validation macro F1 of 0.8074 and a validation micro F1 of 0.9171.

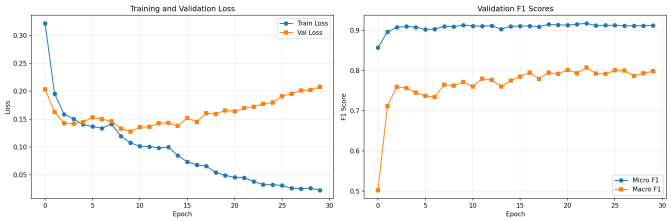

Fig. 3. Training and validation loss curves (left) and F1 score progression (right) over 30 epochs. The model achieves best validation macro F1 at epoch 23.

B. Test Set Performance

1) Overall Metrics (Threshold = 0.5): The final model achieved the following performance on the held-out test set (Table I).

TABLE I
TEST SET PERFORMANCE METRICS

Metric	Score
Macro F1	0.8097
Micro F1	0.9165
Weighted F1	0.9165
Macro Precision	0.8108
Micro Precision	0.9039
Macro Recall	0.8114
Micro Recall	0.9293
Hamming Loss	0.0505
Subset Accuracy	0.4667

Key Observations regarding the metrics indicate a Strong Overall Performance, with a Micro F1 of 0.9165 indicating excellent aggregate performance. The Balanced Class Performance is reflected in a Macro F1 of 0.8097, showing the model performs reasonably well across all classes, including rare ones. The Micro-Macro Gap of 0.107 (0.9165 - 0.8097) indicates some classes perform significantly worse than others, as expected with class imbalance. Furthermore, the High Recall (Micro recall of 0.9293) shows the model successfully detects most positive labels (low false negative rate). Finally, the Strict Accuracy (Subset accuracy of 46.67%) means the model predicts all labels exactly correct for nearly half the test samples.

2) Per-Class Performance Analysis: Detailed per-class results reveal which categories are challenging (Table II).

Regarding the Best Performing Classes, Pavement (F1: 0.9771) was the most frequent class and visually distinctive (gray textured surfaces). Trees (F1: 0.9582) had high support (357 samples) and consistent visual appearance (green foliage), while Buildings (F1: 0.9590) featured distinctive geometric structures and high support (312 samples).

In contrast, the Worst Performing Classes highlighted specific difficulties. Mobile home (F1: 0.0000) had Zero test samples; the class has no representation in the test set, making evaluation impossible. This indicates potential issues with the dataset split or extreme class rarity. Chaparral (F1: 0.4828) was very rare (14 test samples) and visually similar to other vegetation (field, grass). Ship (F1: 0.7654) had moderate support (44 samples) but was challenging due to scale variation and similarity to other water based objects.

TABLE II
PER-CLASS PERFORMANCE ANALYSIS

Class	Precision	Recall	F1 Score	Support
pavement	0.9743	0.9799	0.9771	348
trees	0.9529	0.9636	0.9582	357
buildings	0.9441	0.9744	0.9590	312
airplane	1.0000	0.9091	0.9524	11
grass	0.9227	0.9795	0.9502	341
tanks	0.9333	0.9333	0.9333	15
sea	0.9375	0.9375	0.9375	32
cars	0.9191	0.9342	0.9266	304
sand	0.9024	0.9487	0.9250	39
field	0.7941	0.8710	0.8308	31
bare soil	0.7711	0.8767	0.8205	219
dock	0.8158	0.7750	0.7949	40
water	0.7542	0.7946	0.7739	112
court	0.8571	0.7119	0.7778	59
ship	0.8378	0.7045	0.7654	44
chaparral	0.4667	0.5000	0.4828	14
mobile home	0.0000	0.0000	0.0000	0

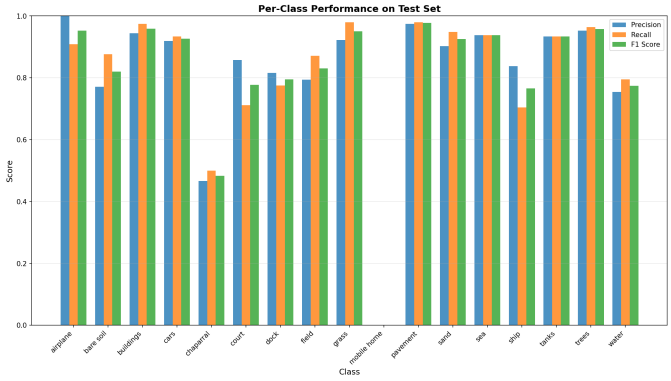

Fig. 4. Per-class precision, recall, and F1 scores on the test set.

The Class Imbalance Impact is evident when comparing sample sizes. Classes with > 300 samples achieved an average F1 of 0.9451, while classes with < 50 samples had an average F1 of 0.6754. This difference of 0.27 F1 points demonstrates significant imbalance effects.

C. Qualitative Analysis

Visual inspection of predictions reveals distinct patterns. Regarding Successful Predictions, the model correctly identifies common co-occurrences (buildings + pavement + cars). It handles rotation invariant recognition well (buildings identified regardless of orientation) and accurately separates similar classes with sufficient context (grass vs. field).

However, there are notable Failure Cases. Chaparral confusion occurs where the class is often mislabeled as field or grass due to visual similarity. Small object detection remains a challenge, as airplanes and tanks are sometimes missed when appearing at small scales. Label omission is observed when the model occasionally predicts a subset of ground truth labels

(e.g., predicting only buildings when water is also present). Additionally, False positives on rare classes occur, where low confidence predictions for mobile homes appear on images containing other residential structures.

D. Comparison with Related Work

While direct comparison is limited due to different dataset versions and experimental setups, our results are competitive. Hua et al. [2] (2020) utilized a Relation Network on the AID_MultiLabel dataset, employing a relation network with label correlation modeling (an advanced GCN-based approach). Their exact results were not specified in accessible materials.

In comparison, Our Baseline (2025) also utilized the AID_MultiLabel dataset but employed EfficientNet-B4 with transfer learning and weighted sampling. We achieved a Macro F1 of 0.8097 and a Micro F1 of 0.9165. Our transfer learning baseline achieves strong performance without explicit label correlation modeling, demonstrating the power of modern pre-trained architectures and effective imbalance handling.

VI. Theoretical Design Justifications

A. Architecture Choices

Regarding the choice of EfficientNet-B4 vs. ResNet50, EfficientNet's compound scaling provides better parameter efficiency (19M vs. 25M parameters). Furthermore, during the experimentation, EfficientNet consistently showed better performance in terms of macro f1 score compared to resnet.

The Transfer Learning Rationale relies on the fact that ImageNet pre-training provides low level features (edges, textures) transferable to aerial imagery. Fine-tuning adapts these features to overhead perspective and domain specific patterns. This strategy is critical given limited labeled aerial data (2,100 training images).

B. Loss Function Selection

In determining Why BCE over Alternatives, we considered distinct trade offs. When compared vs. Softmax + Categorical CE, Softmax enforces mutually exclusive predictions (inappropriate for multi-label). Additionally, vs. Focal Loss, Focal loss is tested during experimentations and it add overhead and also did not increased the performance of the model.

For the BCEWithLogitsLoss Implementation, the method is numerically stable by fusing sigmoid and BCE: log-sum-exp trick prevents overflow/underflow. It is standard in multi-label literature, enabling fair comparison with prior work.

C. Imbalance Handling

Weighted Sampling vs. Class-Weighted Loss: We chose weighted sampling over class-weighted loss for three primary reasons. First, regarding Training dynamics, oversampling exposes the model to rare classes more frequently throughout training, providing more gradient updates. Second, it offers Flexibility, as it works with any loss function without modification. Third, it ensures Batch diversity by maintaining diverse batches while increasing minority class frequency.

Regarding the Alternative (not used), specifically Class-weighted BCE, one can add a pos_weight parameter to BCE-WithLogitsLoss defined as:

$$pos\_weight = neg\_count/pos\_count \quad \text{(per class)} \tag{9}$$

The Trade-off is that while it is simpler to implement, the weights apply to loss gradients, potentially causing training instability. This approach is tried during the experiments but again, it did not increase the performance.

D. Augmentation Strategy

Rotation Invariance is addressed because aerial imagery lacks canonical orientation (unlike natural images where "up" matters). Therefore, $\pm 90^{\circ}$ rotation augmentation teaches the model orientation invariance, which is critical for generalization to images captured from different flight paths.

Color Jittering simulates atmospheric conditions (haze, pollution), time of day variations, and sensor differences. This is essential for robustness to imagery collected under varying conditions.

Finally, regarding Crop vs. Resize, random cropping ($256 \rightarrow 224$) introduces scale variation during training, while at test time, center crop (resize 224) ensures consistent evaluation. The Trade-off is that some boundary information lost, but model learns scale invariant features.

VII. Summary and Future Work

A. Work Completed

We have successfully completed the initial phase of development for the multilabel aerial image classification system, achieving several key technical and research milestones. Technically, the system was built upon a transfer learning pipeline utilizing the EfficientNet-B4 backbone, which established a strong performance baseline on the AID_MultiLabel dataset. Crucially, we implemented a rotation invariant augmentation pipeline and incorporated a weighted sampling strategy to effectively address the challenges posed by class imbalance. These core efforts resulted in meeting the minimum success criterion defined in the project proposal, yielding a macro F1 score of 0.8097 and a micro F1 score of 0.9165 on the test set. Our research contributions further include providing a detailed per-class performance analysis, which definitively demonstrated the effect of class imbalance and identified specific challenging categories chaparral, mobile home, and ship for future investigation and optimization.

B. Next Steps

Building on our current findings, several improvements can be implemented to enhance multi-label aerial scene classification performance. First, modeling label correlations can significantly strengthen prediction accuracy. Graph Convolutional Networks (GCN), particularly the ML-GCN [4] framework, can be used to construct a label co-occurrence graph from the training dataset and learn an adaptive adjacency matrix that captures dependencies such as harbor $\rightarrow$ water or stadium $\rightarrow$ parking. Additionally, attention mechanisms like CBAM

can be integrated into the EfficientNet backbone to enable the model to focus on the most relevant spatial regions and emphasize important feature channels for each label.

Architectural improvements can also be considered. Ensemble strategies that combine EfficientNet-B3, EfficientNet-B4, and ResNet50 predictions may offer higher robustness, with final outputs obtained by averaging class probabilities before thresholding.

Finally, data centric enhancements can further improve generalization. Advanced augmentation techniques such as Mixup, CutMix, and AutoAugment can introduce beneficial regularization and diversify the training distribution. Moreover, per class threshold optimization on the validation set can better handle class imbalance and refine the final decision boundaries without requiring additional training. Collectively, these improvements offer a practical and effective roadmap for future development.

VIII. Conclusion

This project successfully developed a multi-label aerial image classification system achieving competitive performance on the AID_MultiLabel dataset. Our transfer learning approach using EfficientNet-B4 combined with weighted sampling and rotation invariant augmentation yielded a macro F1 of 0.8097 and micro F1 of 0.9165, demonstrating strong overall performance while maintaining reasonable balance across classes.

The comprehensive experiments and analysis revealed both the strengths of modern transfer learning for aerial imagery and the persistent challenges posed by extreme class imbalance. Classes with sufficient training data (pavement, trees, buildings) achieve F1 scores above 0.95, while rare classes (chaparral, mobile home) remain challenging.

Our work establishes a strong foundation for future improvements. The roadmap of advanced loss functions, label correlation modeling, and ensemble methods provides a clear path to pushing performance toward the 0.85-0.90 macro F1 range. The detailed per-class analysis and identified limitations offer valuable insights for both immediate next steps and long term research directions in multi-label aerial scene classification.

References

[1] Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (2017). AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3965-3981.

[2] Hua, Y., Mou, L., & Zhu, X. X. (2020). Relation Network for Multilabel Aerial Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7), 4558-4572.

[3] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A Unified Framework for Multi-label Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2285-2294.

[4] Chen, Z.-M., Wei, X.-S., Wang, P., & Guo, Y. (2019). Multi-Label Image Recognition with Graph Convolutional Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5177-5186.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

[6] Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., Ahmed, A., & Dar, S. H. (2021). Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50. *Mathematical Problems in Engineering*, 2021, 5843816.

[7] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 6105-6114.