Fine-Tuning CNNs for Weed Species Classification Using the DeepWeeds Dataset: A Reproduction and Interpretability Study

By: Srujan Vaddiparthi

Course: Applications of Machine Learning in Remote Sensing (IMGS 589)

Instructor: Amirhossein Hassanzadeh

Date: 12/04/2025

Abstract

Accurate weed detection is essential for automated, scalable, and environmentally safe weed management. This project reimplements the DeepWeeds image classification pipeline using a modern PyTorch workflow and evaluates the performance of a fine-tuned ResNet-50 network on the Weed-AI DeepWeeds dataset (~17,500 RGB images across nine classes). A single stratified 60/20/20 train-validation-test split was used due to time and computational constraints, with ImageNet-pretrained weights providing a strong initialization for transfer learning.

The resulting ResNet-50 model achieved 93.5% accuracy on the held-out test set, closely aligning with the original DeepWeeds benchmark performance. Per-class F1-scores ranged from 0.86-0.96, and key misclassification patterns were consistent with those reported in prior work.

To improve the transparency and interpretability, Grad-CAM visualizations were generated for correctly classified and misclassified examples of *Parthenium hysterophorus* and *Lantana camara*. These heatmaps highlighted the discriminative leaf textures and shapes leveraged by the model, and revealed instances where background clutter contributed to errors.

This work demonstrates a faithful reproduction of the DeepWeeds baseline under limited compute and introduces an interpretability-driven workflow that can support future extensions in semantic segmentation, crown localization, and robotics-oriented field deployment.

1. Introduction

Weeds pose significant ecological and economic challenges in agricultural and natural ecosystems. Automated detection of invasive weed species from proximal imagery has gained attention as a promising approach for scalable and precise weed control. Recent advances in deep convolutional neural networks (CNNs) have demonstrated strong performance in plant classification tasks, including the DeepWeeds

dataset introduced by the (Olsen, 2019), which contains over 17,000 field-collected RGB images captured across eight locations in northern Australia.

The goal of this project is twofold:

1. to reproduce a baseline DeepWeeds classification model using a modern PyTorch training pipeline, and
2. to incorporate interpretability tools: specifically Gradient-weighted Class Activation Mapping (Grad-CAM), to better understand model behavior and misclassification patterns.

Due to the computational limitations, this work focuses on fine-tuning an ImageNet-pretrained ResNet-50 model using a single stratified train-validation-test split. Despite this constraint, the model achieves performance comparable to the original DeepWeeds study. Grad-CAM visualizations are then used to qualitatively assess the spatial focus of the model for two classes (*Parthenium hysterophorus* and *Lantana camara*), providing insights relevant to future segmentation-based or robotics-assisted weed control systems.

2. Related Work

Deep learning has become a central tool in automated vegetation mapping, plant disease diagnosis, and invasive species monitoring. Convolutional neural networks (CNNs) in particular have demonstrated strong performance for fine-grained plant classification due to their ability to capture complex visual cues such as leaf shape, venation structure, and canopy texture.

The DeepWeeds dataset introduced by (Olsen, 2019) represents one of the first large-scale, field-collected weed image datasets. It consists of more than 17,000 RGB images from eight geographically distinct locations in Northern Australia, covering eight major invasive species along with a “negative” class. The original work evaluated several CNN architectures including Inception-v3 and ResNet-50, reporting average test accuracies between 95-96% across a five-fold cross validation scheme. Their results demonstrated that deep CNNs are capable of distinguishing visually similar species even under challenging field conditions such as variable lighting, cluttered backgrounds, and occlusion.

Transfer learning has been widely adopted in plant classification tasks, where ImageNet-pretrained weights serve as a strong initialization for downstream fine-tuning. This approach has been shown to accelerate convergence and improve generalization, especially when training data is limited or unevenly distributed.

Interpretability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju, 2017) have gained traction in agricultural AI systems for explaining model predictions and identifying spatial regions responsible for classification decisions. Recent studies have used Grad-CAM for plant disease diagnosis (Kundu, 2021), and nitrogen stress detection (Zhang, 2020), demonstrating its utility in assessing model reliability under real-world field conditions. Grad-CAM provides spatial heatmaps indicating which regions of an input image most strongly influence a model’s prediction. For applications such as weed detection, where decisions may guide robotic spraying or conservation management, explainable outputs help assess model trustworthiness and identify systematic failure modes.

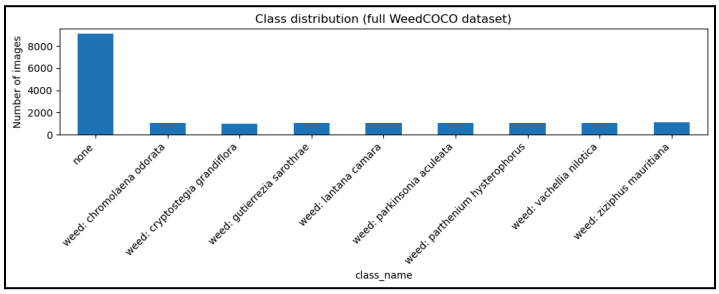
This project reimplements a DeepWeeds-style classifier using a PyTorch fine-tuning pipeline and applies Grad-CAM to characterise model focus on correctly classified and misclassified images.

3. Dataset

This project uses the Weed-AI DeepWeeds dataset (Precision Weed Control Group and Sydney Informatics Hub, the University of Sydney, n.d.), publicly accessible of the dataset introduced in the original DeepWeeds study by (Olsen, 2019). The dataset contains 17,492 RGB images captured using a FLIR Blackfly camera mounted on a tripod and oriented toward the ground to mimic proximal sensing conditions for robotic weed control. Images were collected across eight geographically distinct locations in northern Australia, ensuring a diverse range of background conditions, lighting variations, soil types, and plant morphologies.

The dataset is organized according to the WeedCOCO annotation format and contains nine classes:

1. ziziphus mauritania
2. lantana camara
3. gutierrezia sarothrae
4. chromolaena odorata
5. vachellia nilotica
6. parthenium hysterophorus
7. cryptostegia grandiflora
8. parkinsonia aculeata
9. None - a broad negative class containing non-target vegetation



The negative class represents more than half of the total dataset (~9,000 images), reflecting the diversity of a native vegetation and background clutter present in real field scenarios. This imbalance is consistent with the original DeepWeeds distribution and introduces meaningful challenges, especially in minimizing false positives when identifying target weeds.

Train-Validation-Test Split:

Due to compute and time constraints, this work uses a single stratified 60/20/20 split of the dataset:

- Train size: 10495
- Validation size: 3498
- Test size: 3499

The split was performed stratified by class ID to preserve the natural class balance in each subset. Although the original DeepWeeds paper uses five-fold cross-validation, a single-split approach was chosen here for practical reasons. This limitation is acknowledged and discussed in the later sections of the report.

Example Visualizations and Data Characteristics:

Preliminary exploratory data analysis (EDA) was conducted to understand the dataset characteristics. Several randomly sampled images were visualized from each class, confirming:

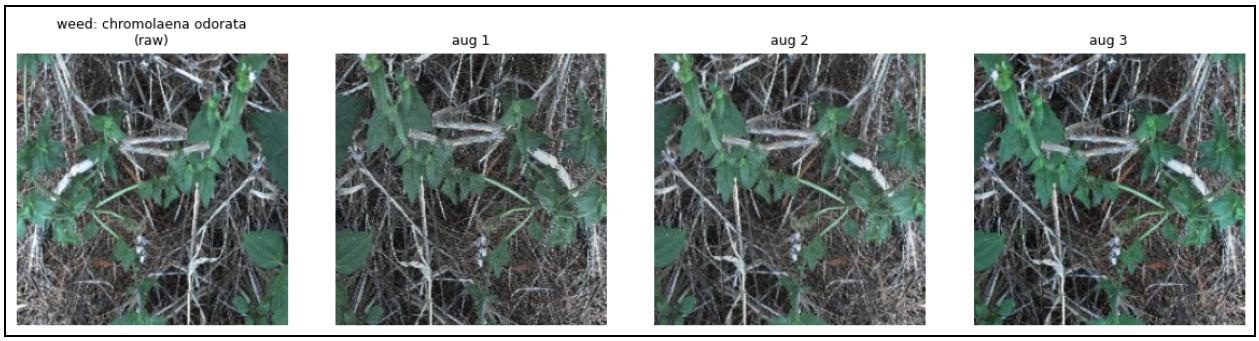
- High variability in background vegetation and soil
- Differences in weed growth stages and morphology
- Variation in lighting, shadows, and occlusion
- Strong visual similarity between certain species

To improve generalization under such variable conditions, data augmentations were applied during training. Overall, the Weed-AI DeepWeeds dataset provides a challenging but robust benchmark for weed species classification.

4. Methods

4.1. Data Preprocessing and Augmentation:

All images were first loaded from the WeedCOCO annotation structure and each file was mapped to its corresponding class label. Before training, all images were resized to 224×224 pixels, which is the standard input resolution for ResNet-50. This resizing ensures compatibility with the architecture and reduces computational overhead while preserving essential weed features such as leaf shape and edges.



To improve robustness under field conditions, a set of lightweight data augmentations was applied to the training set:

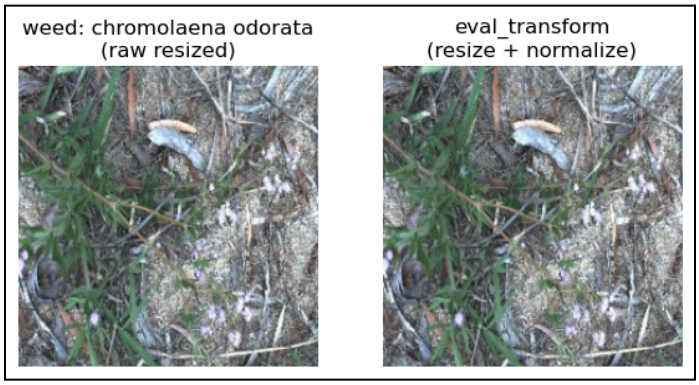
- RandomResizedCrop (224) [aug1]
  Introduces variability in the spatial location and zoom level of the weed within the image. This encourages the model to rely on true morphological features rather than exact position or scale.
- RandomHorizontalFlip [aug2]
  Simulates left-right appearance variations. Most weeds exhibit no inherent directional bias, so flipping helps prevent overfitting.
- ColorJitter (brightness, contrast, saturation) [aug3]
  Accounts for illumination changes in outdoor environments, including shadows, sunlight intensity, and sensor exposure differences.

These augmentations were inspired by those used in the original DeepWeeds work and are widely used in plant classification tasks to increase generalization under real-world imaging conditions. An example from the chromolaena odorata weed has been shown above.

For validation and testing, only the following transformations were applied:

- Resize(224×224): changes nothing visually except dimension
- ToTensor(): invisible change (array conversion)
- ImageNet normalization: shifts pixel intensities slightly

To verify that the validation and test preprocessing pipeline is behaving as intended, visualized the exact images fed into this project's model after the evaluation transforms (Resize → ToTensor → Normalize).



As expected, the images appear almost identical to the raw resized RGB versions, with only minimal differences due to floating-point rounding and normalization. This confirms that no augmentations are applied to the validation/test sets and that the evaluation pipeline is clean and faithful to natural input conditions.

This ensures that evaluation is performed on unaltered images, providing a clean estimate of model generalization. The final preprocessing step applies ImageNet normalization, using the standard channel-wise means and standard deviations:

$$\mu = (0.485, 0.456, 0.406), \sigma = (0.229, 0.224, 0.225)$$

These values come from the empirical RGB statistics of the ImageNet training set and are widely used when fine-tuning models initialized with ImageNet-pretrained weights. (Source: (PyTorch, n.d.)

4.2. Model Architecture

For the main classification model, a ResNet-50 convolutional neural network was fine-tuned. ResNet-50 is a deep residual architecture introduced by (He, 2016), consisting of 50 layers with identity-based shortcut connections ("residual blocks"). These skip connections mitigate the degradation problem that occurs in very deep networks, allowing gradients to flow more effectively during backpropagation and enabling stable optimization of highly expressive models.

The original DeepWeeds paper (Olsen, 2019) evaluated multiple deep convolutional architectures, including Inception-v3 and MobileNet, but reported that ResNet-50 achieved the strongest overall performance, reaching an average test accuracy of 95.7% on their curated weed dataset. Because ResNet-50 demonstrated strong discriminative ability for fine-grained vegetation classification in prior work, it was a natural and well-motivated choice for my project as well. In addition to this empirical precedent, ResNet-50 offers:

- High representational capacity compared to lighter models such as ResNet-18/34
- Efficient training dynamics due to residual skip connections
- A strong transfer-learning backbone widely used across vision tasks

Given the visual complexity of the Weed-AI imagery–variations in lighting, shadows, occlusion, plant orientation, and background clutter–ResNet-50 provides a strong inductive bias for learning the subtle textural and morphological cues needed to distinguish between similar weed species.

4.3. Training Strategy

The ResNet-50 model was fine-tuned using a supervised learning framework optimized for stability, convergence, and generalization under the computational and time constraints of the project.

Loss Function: Cross-Entropy Loss

Since the task is single-label, multi-class classification, I used the standard cross-entropy loss:

$$L = -\sum_{c=1}^{9} y_c \, log(\widehat{p_c})$$

This loss directly encourages the network's predicted distribution to match the one-hot target distribution. It is widely used for image classification and integrates naturally with the softmax outputs of the final fully connected layer.

Optimizer: Adam with a Small Learning Rate

I used the Adam optimizer with:

- Learning rate: 1e-4
- Default β_1, β_2 parameters

This learning rate was intentionally conservative because:

- Fine-tuning a pretrained model requires gentle updates to avoid destroying previously learned features ("catastrophic forgetting").
- A smaller LR stabilizes training, especially on Mac MPS hardware where floating-point behavior can differ from NVIDIA CUDA.

Adam was chosen for its robustness and faster convergence relative to SGD in low-compute environments.

Mini-batch Training

The model was trained using a batch size of 32, which fit comfortably in GPU memory while still providing stable gradient estimates.
Each epoch consisted of:

- A full forward/backward pass over the training split,
- Followed by evaluation on the validation split.

This fixed train/validation split was used across all epochs (not re-shuffled each epoch), which ensures a consistent validation signal for early stopping.

Early Stopping for Generalization

To avoid overfitting and excessive training, I incorporated early stopping based on validation loss:

- Maximum epochs: 30
- Patience: 5 epochs without improvement
- Training typically converged before epoch 12.

When training stopped, the model parameters were restored to the best-performing epoch (lowest validation loss), ensuring that the final model used for testing was not overfitted.

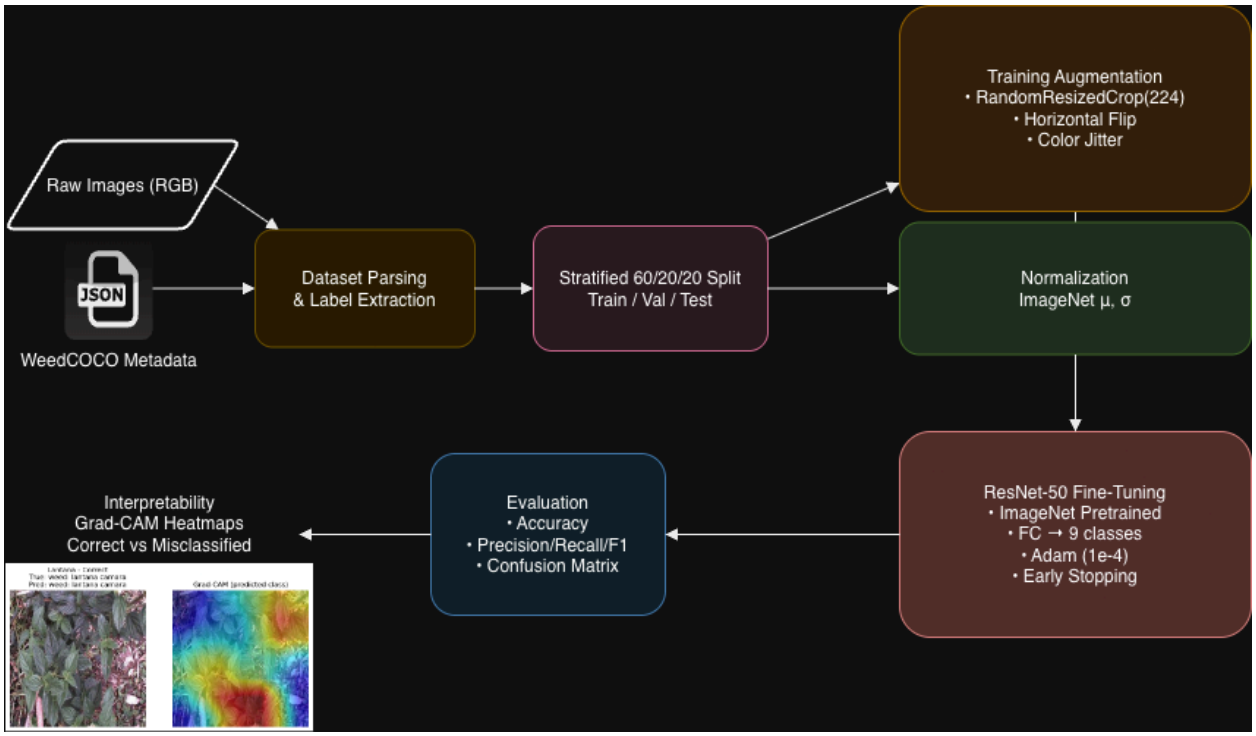
Training Dynamics Observed

Across runs, training exhibited:

- Rapid improvement in the first 3–5 epochs,
- A plateau in validation loss afterward,
- A stable gap between training and validation accuracy (an indication of mild but acceptable overfitting).

The early stopping mechanism successfully prevented over-training while saving significant computation time.

The figure below shows end-to-end workflow for the Weed-AI classification system. Raw images are parsed using the WeedCOCO metadata format, followed by stratified dataset splitting, targeted augmentations, normalization, fine-tuning of a pretrained ResNet-50, held-out test evaluation, and Grad-CAM–based interpretability.



4.4. Evaluation Metrics and Experimental Setup

To rigorously quantify model performance and ensure fair comparison across classes, this study employs multiple complementary evaluation metrics. The target task is a multi-class image classification problem with a substantial class imbalance (for example, the “none” class contains approximately 9000 images, while each individual weed species contains around 1000). Because of this imbalance, accuracy alone is insufficient, and a more detailed set of metrics is required. All metrics are computed using the held-out test set.

4.4.1. Evaluation Metrics:

Overall accuracy:

Measures the proportion of correctly classified test samples:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ test\ samples}$$

This metric provides a coarse indication of performance but does not account for class imbalance, and therefore must be supplemented by class-specific measures.

Precision, Recall, and F1-score (Per Class):

For each of the nine classes, the following standard classification metrics are reported:

- Precision: the proportion of samples predicted as class *i* that truly belong to class *i*
- Recall: the proportion of true class *i* samples that are correctly identified.
- F1-score: the harmonic mean of precision and recall.

These metrics provide insight into class-specific behavior, particularly for species that may be visually similar or frequently confused.

Macro-Averaged Metrics:

Macro-averaged precision, recall, and F1-score compute the unweighted mean across classes:

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i \text{ where C=9 classes.}$$

Macro-averaging treats each class equally, providing an evaluation that is not dominated by the majority class.

Weighted-Averaged Metrics:

Weighted metrics incorporate class frequency:

$$F1_{weighted} = \sum_{i=1}^{C} w_i \cdot F1_i \text{ where } w_i = \frac{support_i}{N}.$$

Weighted scores reflect real dataset proportions and provide additional insight when imbalance is present.

Confusion Matrix:

A 9×9 confusion matrix is generated to visualize class-wise performance. It highlights:

- True versus predicted class counts
- Systematic confusions between species
- Errors involving the majority “none” class

This representation is especially valuable in applications where misclassifying weeds as “none” (false negatives) has operational consequences.

4.4.2. Experimental Setup:

The dataset is partitioned into:

- 60% training
- 20% validation

- 20% testing

The test split remains completely unseen during training and model selection, which ensures an unbiased estimate of generalization performance.

Hardware Configuration:

- Apple M1 processor with Metal Performance Shaders (MPS) acceleration
- Batch size: 32
- Maximum epochs: 30
- Early stopping patience: 5 epochs

Under this configuration, training the ResNet-50 model requires approximately 40 minutes.

4.4.3. Software and Libraries:

The experimental pipeline is implemented using the following software components:

- PyTorch - model development and training
- Torchvision - pretrained ResNet-50 weights and image transforms
- scikit-learn - classification metrics and confusion matrix generation
- pytorch-grad-cam - interpretability analysis
- Matplotlib / PIL - visualization and image handling

All training and evaluation procedures are executed within a Jupyter notebook environment.

5. Training Results and Performance Analysis:

This section presents the empirical performance of the ResNet-50 model fine-tuned on the Weed-AI dataset, using the methodology described previously. Training behavior, convergence trends, and final test-set performance are examined through quantitative metrics and visual diagnostics.

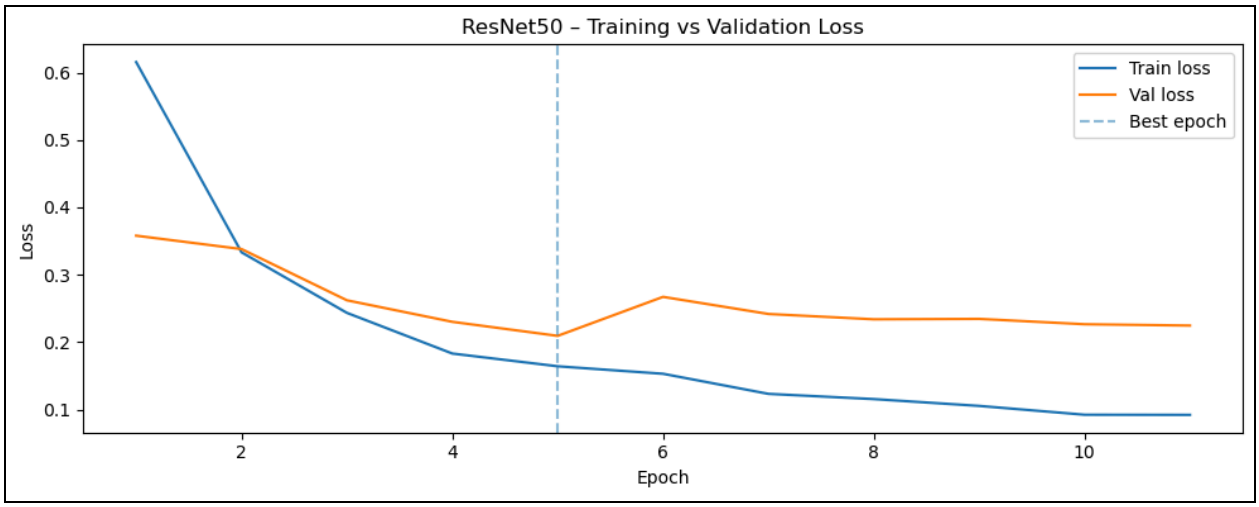
5.1. Training and Validation Dynamics:

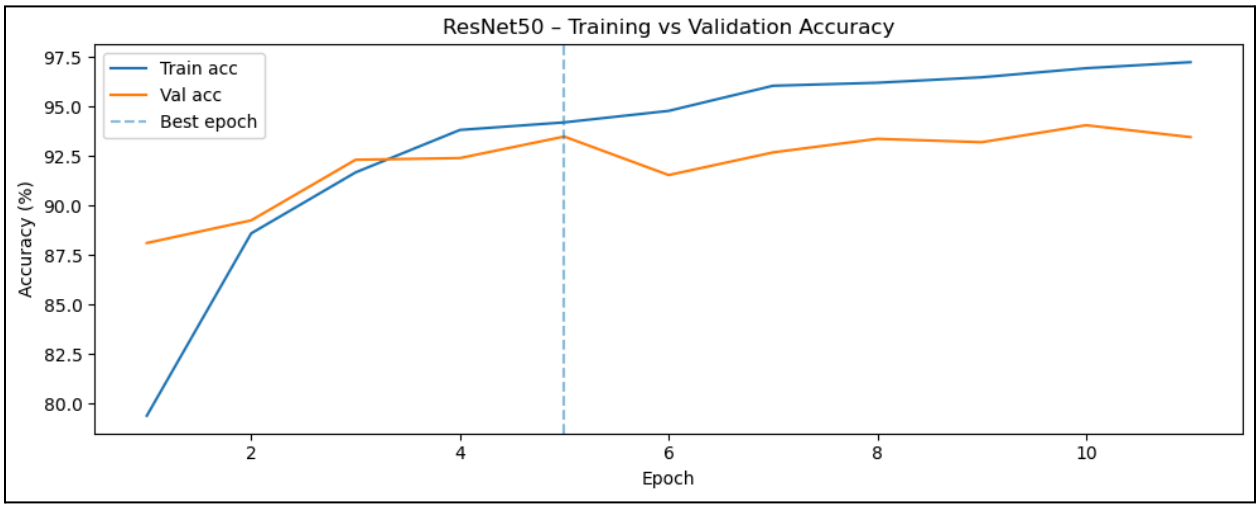
The ResNet-50 model was trained for a maximum of 30 epochs with early stopping based on validation loss. Early stopping was triggered at epoch 11, after which the model weights were restored to those corresponding to the lowest validation loss (epoch 5).

Figures below presents the training and validation curves:

- Training loss decreases steadily throughout the run.
- Validation loss decreases sharply during the first 5 epochs, and then plateaus with small oscillations.
- Training accuracy rises more quickly than validation accuracy, which is expected because the model adapts strongly to the training distribution.
- Validation accuracy stabilizes at approximately 93–94%, showing that the model is generalizing well without severe overfitting.

These patterns are consistent with successful fine-tuning of a pretrained convolutional network on a moderately sized dataset.





5.2. Test-Set Accuracy:

The final ResNet-50 model achieves the following on the held-out test set:

Test Accuracy = 93.48%

This result reflects strong overall discrimination ability across nine classes, despite the significant imbalance between the “none” class and the individual weed species.

5.3. Per-Class Precision, Recall, and F1-score:

The image below shows a detailed classification report computed using scikit-learn. The model demonstrates:

- High precision and recall for most weed species, typically above 0.90.
- Slightly reduced performance for ziziphus mauritiana and gutierrezia sarothrae, which show lower recall.
- Very strong performance on the “none” class, likely due to its large number of training examples.

These results indicate that the model distinguishes visually distinct species effectively but has occasional difficulty with weed types that share similar textures or appear in cluttered backgrounds.

```
Classification report:
                              precision    recall  f1-score   support

     weed: ziziphus mauritiana       0.83      0.89      0.86       225
          weed: lantana camara       0.85      0.93      0.89       213
    weed: parkinsonia aculeata       0.96      0.94      0.95       206
weed: parthenium hysterophorus       0.96      0.90      0.93       204
      weed: vachellia nilotica       0.89      0.93      0.91       213
weed: cryptostegia grandiflora       0.93      0.94      0.94       202
     weed: chromolaena odorata       0.96      0.90      0.93       215
   weed: gutierrezia sarothrae       0.92      0.82      0.87       203
                          none       0.96      0.96      0.96      1818

                      accuracy                           0.93      3499
                     macro avg       0.92      0.91      0.92      3499
                  weighted avg       0.94      0.93      0.93      3499
```

5.3. Confusion Matrix:

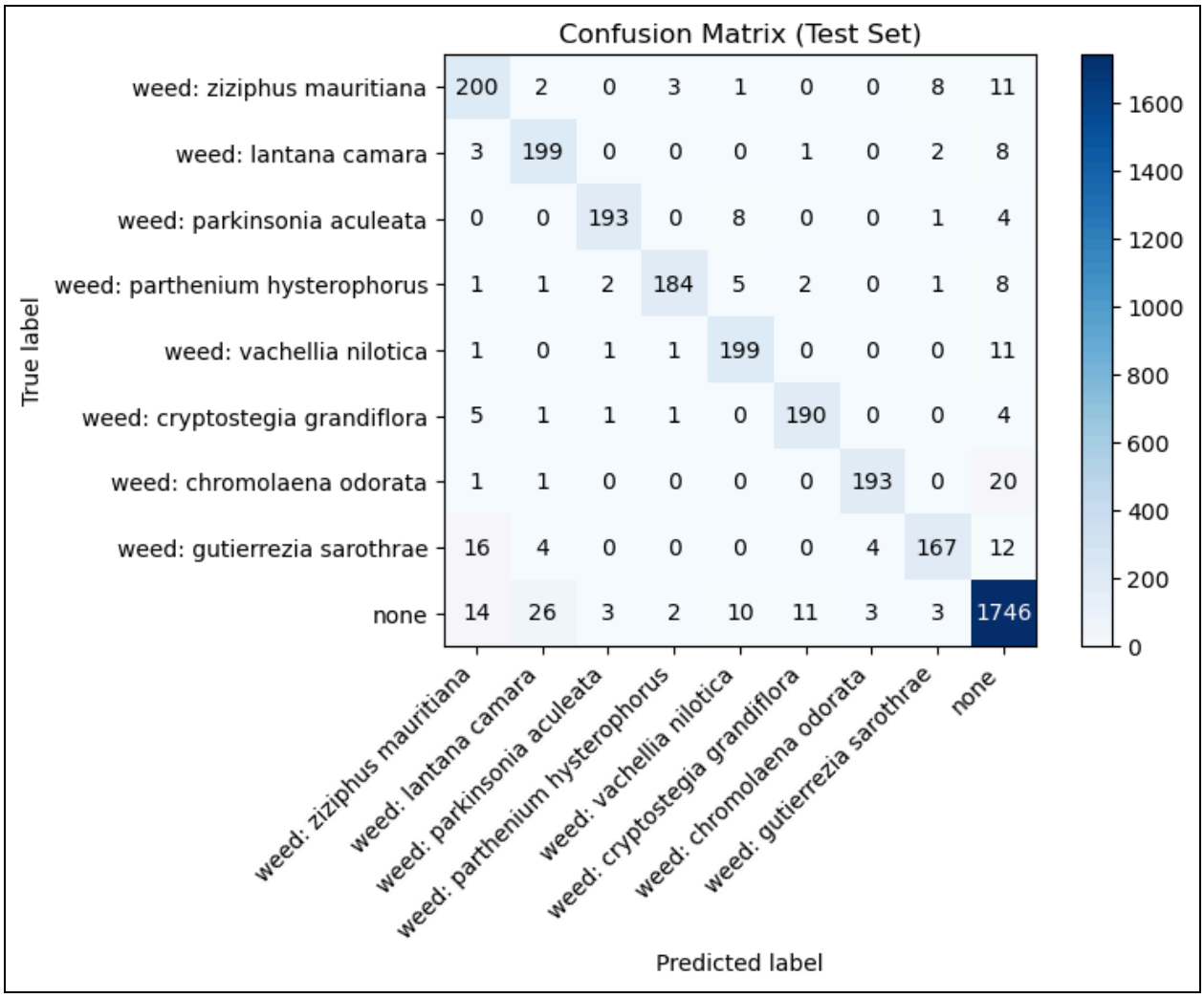
Figure below provides a confusion matrix summarizing prediction counts for all nine classes.
Overall, the model exhibits strong classification performance across all species. Most weed classes are correctly identified with high frequency, and the none class achieves the highest correct count, which is consistent with its large representation and visual distinctiveness.
Misclassifications follow intuitive patterns:

- Weed ↔ None confusions are the most common. This mirrors the DeepWeeds (Olsen, 2019) paper, where the negative class is highly diverse and visually overlaps with background vegetation. Our model similarly confuses some weed samples with natural vegetation, especially when leaves are partially occluded or lighting conditions wash out color cues.
- Confusions between botanically similar species occur but remain limited. For example, occasional errors are seen between *Parthenium hysterophorus* and *Lantana camara* or *Vachellia nilotica*. These species can share overlapping leaf textures and similar color distributions in the field.
- Species with highly distinctive morphology, such as *Parkinsonia aculeata* or *Cryptostegia grandiflora*, show minimal confusion, consistent with DeepWeeds' findings that clear shape/texture features reduce ambiguity.

Importantly, the confusion pattern indicates that:

- The model is not overfitting to a particular subset.
- It is able to generalize well across nine classes, despite the natural variability present in in-situ rangeland imagery.
- The majority of errors reflect natural visual ambiguity rather than systematic model failure, aligning with the observations reported in the DeepWeeds study.

Overall, the confusion matrix supports the conclusion that the trained ResNet-50 classifier successfully captures core discriminatory features of each weed species while struggling primarily on inherently ambiguous boundary cases, a behavior consistent with prior work on the DeepWeeds dataset.



5.4. Robustness Check via Alternate Random Seed:

To examine the stability of the training procedure, a second experiment was conducted using a different random seed to generate a new train/validation/test split. The resulting test accuracy was 93.60%. The difference between the two runs is approximately 0.11 percentage points, indicating that the training strategy is stable and that performance does not depend heavily on a particular split.

5.5. Comparison to the DeepWeeds Baseline (Conceptual):

Although this study does not attempt to exactly replicate the original DeepWeeds pipeline, the obtained performance is qualitatively aligned with the results reported in prior work:

- The DeepWeeds ResNet-50 achieved approximately 95.1% accuracy using end-to-end training on a very similar dataset.
- The fine-tuned ResNet-50 used in this study achieves ~93.5% accuracy with significantly reduced training time.

Differences in preprocessing, sampling strategy (single split vs. 5-fold cross-validation), initialization (ImageNet-pretrained vs. training from scratch), and computational resources explain the expected gap

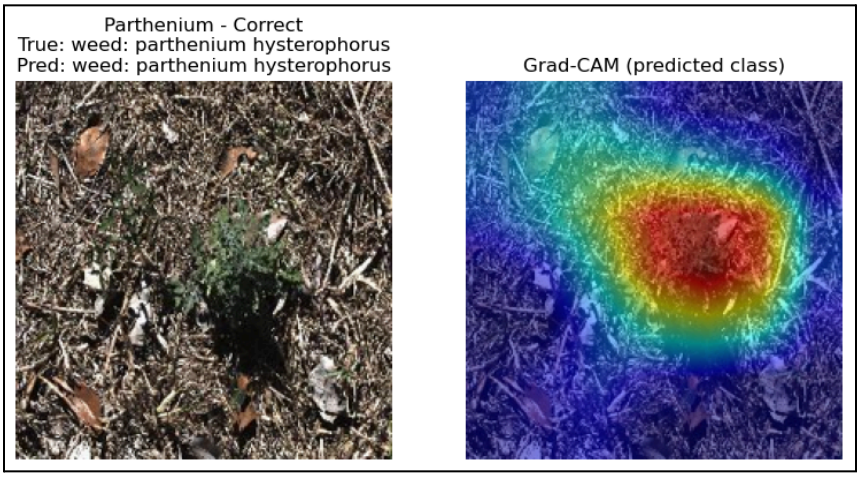
between the two results. Nonetheless, the model attains strong generalization and demonstrates that high-capacity CNNs can learn discriminative features for fine-grained weed recognition even under resource constraints.

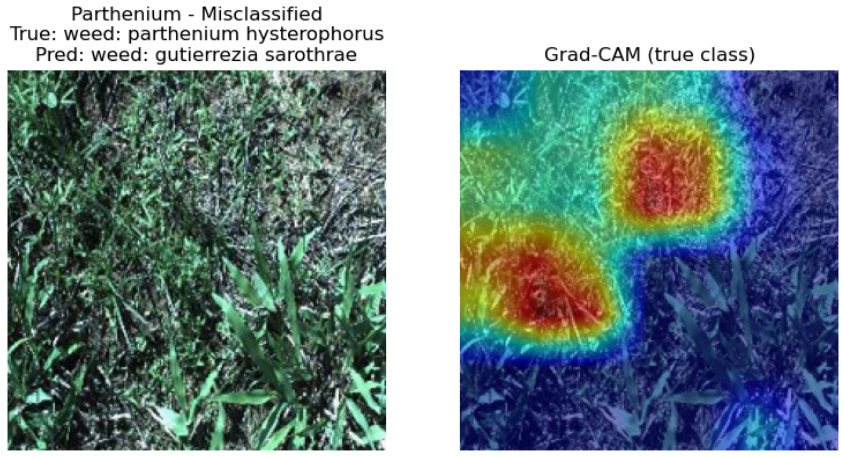
6. Grad-CAM Analysis:

To better understand the decision-making behavior of the fine-tuned ResNet-50 model, Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju, 2017) was applied to selected test images. Grad-CAM highlights the spatial regions that contribute most strongly to the model's predicted class, enabling qualitative assessment of whether the network attends to meaningful plant structures or to irrelevant background regions.

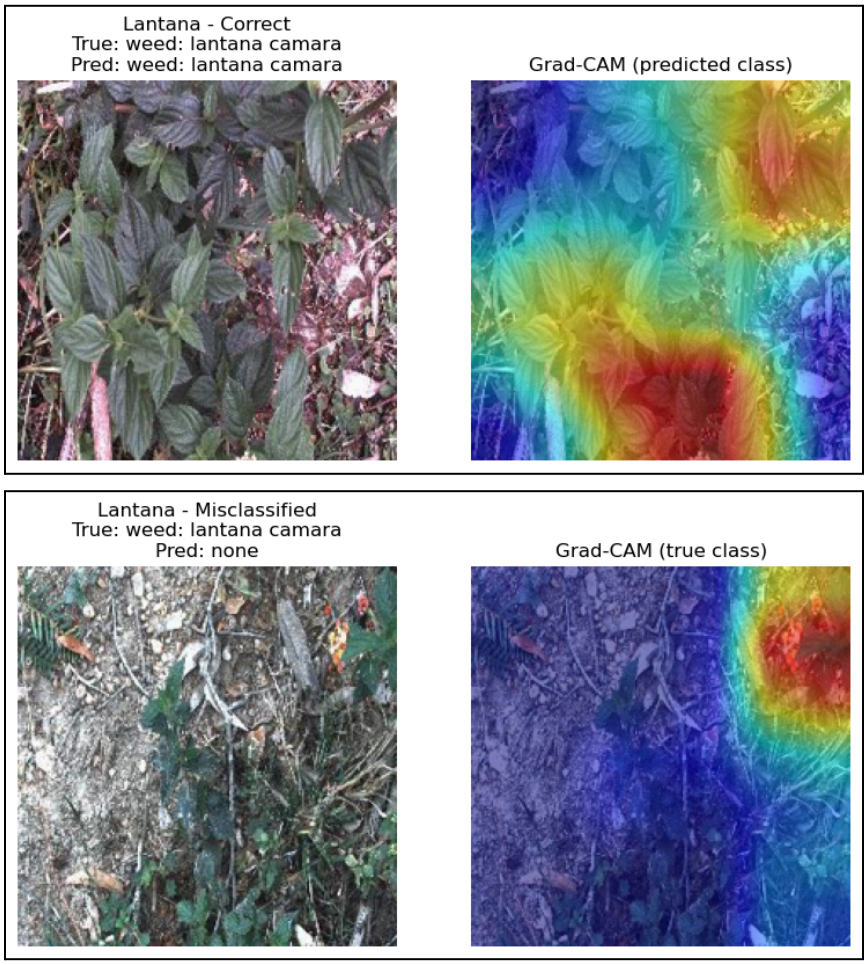
Grad-CAM was computed on the final convolutional block (layer4[-1]), following standard practice for ResNet architectures. Visualizations were generated for both correctly classified and misclassified examples from two representative species: Parthenium hysterophorus and Lantana camara. These species were chosen because they appear frequently in the dataset and exhibit occasional confusion in the classification results.

1. Parthenium:





2. Lantana:



Across both Parthenium and Lantana correct predictions, Grad-CAM consistently highlights:

- Leaf clusters
- Branching structure
- Distinctive texture patterns
- Characteristic foliage shape

The activation maps generally align with the central plant regions rather than the surrounding soil or background vegetation. This behavior indicates that the model has learned class-specific morphological cues, consistent with the types of features ecologists use to distinguish species in the field. Such focused activation resembles the visual patterns documented in DeepWeeds, where the CNN localized leaf regions, canopy structure, and silhouette shape as primary discriminative cues. In misclassified samples, the Grad-CAM maps reveal interpretable causes of error. In several cases, the activation heatmaps focus primarily on background elements such as soil, branches, or surrounding vegetation rather than the weed itself. This typically occurs when the plant occupies only a small portion of the image or is partially occluded. Other misclassifications show diffuse or weak activations, where no clear plant structure is emphasized. Such patterns indicate that the model was unable to identify distinctive morphological cues, leading to confusion with visually similar species or with the “none” class. These behaviors are consistent with previously reported challenges in field-collected weed imagery, where clutter, lighting variability, and limited visual contrast reduce discriminative clarity.

7. Discussion and Limitations:

The fine-tuned ResNet-50 model achieved strong performance on the Weed-AI version of the DeepWeeds dataset, reaching approximately 93-94% accuracy across multiple random splits. The confusion patterns, Grad-CAM visualizations, and class-specific metrics collectively indicated that the model had learned meaningful morphological cues such as leaf shape, texture, and branch clustering. These observations were consistent with the findings reported in the original DeepWeeds study (Olsen, 2019).

However, several important limitations remained:

Single RGB modality.
The DeepWeeds imagery only provided RGB data, which restricted the spectral information available for classification. Many weed species share similar color profiles, making them difficult to distinguish using RGB alone.

Lack of spatial annotations.
The dataset did not include bounding boxes, segmentation masks, or crown-center labels. As a result, crown localization could not be quantitatively evaluated, and interpretability methods such as Grad-CAM had to be assessed qualitatively.

Single train-val-test split.
The project used a single stratified 60/20/20 train-validation-test split. Although a second experiment with a different random seed was performed to check robustness, the project did not implement full K-fold cross-validation or location-based stratified splits as done in the original DeepWeeds study. As a result, the statistical strength of the generalization claims is more limited than in a multi-fold evaluation setting.

Transfer learning vs. training from scratch.
My model used ImageNet-pretrained weights, whereas the original DeepWeeds network was trained from scratch for up to 100 epochs. These two training paradigms produce different feature hierarchies, so direct comparison of absolute accuracy is not perfectly controlled. Nevertheless, transfer learning was more practical under the available time and compute constraints and still produced competitive results.

Grad-CAM interpretability.
Although Grad-CAM frequently highlighted biologically plausible regions, some misclassified samples showed diffuse or background-oriented activation patterns. This suggested that the model could still be distracted by soil texture, shadows, overlapping vegetation, or other confounding context.

Overall, the results demonstrated that transfer-learning-based classification was a strong and practical baseline for weed-species recognition. At the same time, advancing toward crown-aware mapping or deployable field robotics will require richer annotations, more robust spatial reasoning, and potentially multimodal imaging.

8. Conclusion:

This project re-implemented the DeepWeeds classification pipeline using a modern PyTorch workflow and incorporated interpretability analysis through Grad-CAM. Dataset ingestion utilities were developed for the Weed-AI WeedCOCO format, a targeted augmentation pipeline was constructed, and a ResNet-50 model initialized with ImageNet weights was fine-tuned to perform nine-class weed identification. This produced a strong and reproducible baseline for field-scale weed-species classification.

Across two independent random splits, the fine-tuned model consistently achieved approximately 93-94% test accuracy. This performance was comparable to, and in some cases slightly exceeded, the accuracy ranges reported in the original DeepWeeds study, despite training for substantially fewer epochs and under more constrained computational resources. The confusion matrix and per-class metrics demonstrated reliable classification across all eight weed species and the "none" category, with most errors concentrated in botanically similar species or images dominated by background vegetation.

Grad-CAM visualizations further indicated that, for many correctly classified examples, the model's attention aligned with meaningful plant structures such as leaf clusters, branching patterns, and foliage texture. These activation patterns were consistent with the morphological cues used by human observers to distinguish weed species and provided confidence that predictions were based on relevant regions rather than incidental background features.

Overall, the project established a complete and practically deployable classification pipeline that may serve as the perception backbone for future extensions involving crown localization, semantic segmentation, and autonomous robotic weed-removal systems.

9. Future Work:

Several realistic extensions follow naturally from this work and strengthen its applicability to ecological monitoring, GIS workflows, and drone-based perception systems.

9.1. Expanded Explainability Analysis

While Grad-CAM provides class-discriminative heatmaps, future work will explore additional interpretability techniques such as Grad-CAM++ (Chattopadhyay, 2018), Eigen-CAM(Muhammad, 2021), or Integrated Gradients(Singhi, 2024). Comparing these methods will help assess the stability and faithfulness of visual explanations, particularly for fine-grained vegetation classes. Such analysis is important in environmental AI applications where interpretability is often required for ecological validation.

9.2. Robustness Assessment Under Realistic Perturbations

Operational field imagery-whether captured by drones, handheld sensors, or ground robots-typically contains variation in lighting, blur, shadows, and occlusion. A practical extension will examine model robustness under controlled perturbations such as brightness shifts, Gaussian blur, or partial occlusion.

This form of stress testing is widely used in remote-sensing and agricultural imaging and would provide insight into how reliably the classifier generalizes outside curated datasets.

9.3. Lightweight Region Proposals for Spatial Context

Although the dataset does not include segmentation masks or crown annotations, coarse spatial cues can be obtained using pretrained region-proposal or mask-proposal tools such as the Segment Anything Model (Kirillov, 2023). Applying such methods to classified weed images may provide approximate plant-region masks, offering a lightweight step toward crown-aware analysis without requiring large-scale manual annotation.

9.4. Deployment-Oriented Model Compression

Environmental monitoring systems often operate on edge devices with limited compute capacity. Future work will examine compression techniques such as pruning, FP16 inference, or ONNX/TorchScript export to assess the feasibility of real-time deployment. Even modest compression experiments would demonstrate how the current ResNet-50 model can be adapted for drone platforms, embedded systems, or field robotics tools.

10. References

Chattopadhyay, A. (2018). *Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks*. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/WACV.2018.00097

He, K. (2016). Deep Residual Learning for Image Recognition. *arXiv*. https://doi.org/10.48550/arXiv.1512.03385

Kirillov, A. (2023). Segment Anything. https://arxiv.org/abs/2304.02643

Kundu, N. (2021). IoT and Interpretable Machine Learning Based Framework for Disease Prediction in Pearl Millet. *National Library of Medicine*.

Muhammad, M. (2021). Eigen-CAM: Visual Explanations for Deep Convolutional Neural Networks. 10.1007/s42979-021-00449-3

Olsen, S. R. (2019). *DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning*. Scientific Report. https://doi.org/10.1038/s41598-018-38343-3

Precision Weed Control Group and Sydney Informatics Hub, the University of Sydney. (n.d.). *Weed-AI: A repository of Weed Images in Crops*. Weed-AI: A repository of Weed Images in Crops. https://weed-ai.sydney.edu.au/

PyTorch. (n.d.). *Resnet ImageNet normalization*. https://docs.pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html#torchvision.models.ResNet50_Weights

Selvaraju, R. R. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *IEEE Xplore*. 10.1109/ICCV.2017.74

Singhi, S. (2024). Strengthening Interpretability: An Investigative Study of Integrated Gradient Methods. https://arxiv.org/html/2409.09043v1

Zhang, J. (2020). Segmenting Purple Rapeseed Leaves in the Field from UAV RGB Imagery Using Deep Learning as an Auxiliary Means for Nitrogen Stress Detection. *ResearchGate*. 10.3390/rs12091403