Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CAM-Threshold-Guided Weakly Supervised Lesion Segmentation and Relative Lesion-Burden Stratification of Wheat Stripe Rust Using CF2- WRD and Light-UNet

Bojian Guo

guobojian26@kstu.kg

Kyrgyz State Technical University after I Razzakov

Xin LI

Kyrgyz State Technical University after I Razzakov

Temirbaeva Nazgul Ismanovna

Kyrgyz State Technical University after I Razzakov

Asel Kartanova

Kyrgyz State Technical University after I Razzakov

Article

Keywords: wheat stripe rust, weakly supervised segmentation, Grad-CAM, pseudo-label, Light-UNet

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10959751/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CAM-Threshold-Guided Weakly Supervised Lesion Segmentation and Relative Lesion-Burden Stratification of Wheat Stripe Rust Using CF2-WRD and Light-UNet

Bojian Guo[1]*, Xin LI[2], Temirbaeva Nazgul Ismanovna[3], Asel Kartanova[4]
[1]Department of Information Systems in Economics, Kyrgyz State Technical University after I Razzakov, 720044, Bishkek, Kyrgyz Republic
[2]Department of Information Systems in Economics, Kyrgyz State Technical University after I Razzakov, 720044, Bishkek, Kyrgyz Republic
[3]Department of Information Systems in Economics, Kyrgyz State Technical University after I Razzakov, 720044, Bishkek, Kyrgyz Republic
[4]Department of Information Systems in Economics, Kyrgyz State Technical University after I Razzakov, 720044, Bishkek, Kyrgyz Republic

*Correspondence: guobojian26@kstu.kg

Abstract

Accurate localization and quantification of wheat stripe rust lesions are important for phenotyping and disease management, yet dense pixel-level annotation is costly. We present a Grad-CAM-guided weakly supervised framework that reuses a dual-branch image classifier, CF2-WRD, to generate pseudo-labels for a compact Light-UNet segmenter. The dataset contained 34,694 images in 87 image-level categories, split into 27,834 training, 3,430 validation, and 3,430 test images. CF2-WRD achieved validation and test accuracies of 0.8776 and 0.8711. Among four CAM thresholds, 0.40 produced the highest agreement with CAM-derived validation pseudo-labels (mIoU 0.6584, Dice 0.7101, lesion IoU 0.5555), whereas 0.45 was used as a conservative operating point. Raising the segmentation probability threshold from 0.50 to 0.80 reduced mean predicted lesion-to-leaf ratio from 55.02% to 29.63% in a 300-image subset. The 3,430-image burden distribution was stratified by empirical quartiles. Because no independent manual lesion masks or expert-calibrated severity labels were available, overlap metrics quantify agreement with weak-supervision targets and burden groups represent dataset-relative rankings rather than agronomic severity grades.

Keywords: wheat stripe rust; weakly supervised segmentation; Grad-CAM; pseudo-label; Light-UNet

1. Introduction

Wheat stripe rust, caused by Puccinia striiformis f. sp. tritici, is a major foliar disease of wheat. Epidemics can develop rapidly under cool and humid conditions, producing elongated yellow-orange uredinial stripes on susceptible leaves. Severe infection reduces the photosynthetic area, accelerates senescence, and threatens yield stability[1]. Effective management therefore requires more than a binary diagnosis. In breeding, field monitoring, and fungicide decision-making, it is also necessary to know where lesions occur and how much leaf area has been affected. Spatially explicit outputs are particularly valuable for field and mobile-based phenotyping systems, where visual assessment must be both scalable and interpretable.

Computer vision has markedly improved automated plant disease recognition. Convolutional neural networks trained on RGB images can distinguish disease categories with high image-level accuracy[2-4], and wheat-specific studies have reported effective stripe rust grading and detection under controlled, mobile, and field conditions. Classification, however, still produces a global category or score. It does not delineate lesion boundaries, separate scattered uredinial regions, or yield a transparent area-based severity estimate. Pixel-level segmentation can provide these outputs, but conventional supervised segmentation requires a manually delineated mask for every training image. Such annotation is much more expensive than image-level labeling, is sensitive to ambiguous lesion margins, and often depends on plant-pathology expertise.

Encoder-decoder networks, including FCN, U-Net, and DeepLab, have become standard tools for dense prediction[5-7]. U-Net is especially useful for agricultural imagery because skip connections preserve fine spatial detail while the encoder captures broader context. The annotation bottleneck nevertheless remains. Weakly supervised semantic segmentation offers a practical alternative by deriving approximate spatial supervision from cheaper labels such as image-level classes, points, scribbles, or bounding boxes. Class activation mapping (CAM) and gradient-weighted CAM are attractive in this setting: when a reliable classifier already exists, they reveal image regions that contribute strongly to the class decision[8,9].

CAM-based supervision also has an inherent limitation. A classifier is trained to find enough evidence for recognition, not to recover the full object or lesion extent. A CAM may therefore highlight only the most discriminative portion of an infected region. Converting a continuous heatmap into a binary pseudo-mask adds another source of variation: the threshold. Low thresholds preserve faint and peripheral lesion responses but may include healthy tissue or background. High thresholds suppress noise but can reduce lesions to their most confident cores. Expansion, affinity learning, self-equivariant attention, and full-resolution CAM refinement have been proposed to mitigate incomplete activation[10-12]. Even so, the operating threshold is often underreported, especially for small, elongated, and low-contrast plant lesions.

Here, we examine a practical classification-to-segmentation transfer strategy for wheat stripe rust. A previously trained CF2-WRD classifier is used as a frozen feature generator. CF2-WRD contains MobileNetV2-like and EfficientNet-B0-like branches and an adaptive complementary feature fusion module, producing a fused 7 × 7 feature map that combines lightweight global representation with texture-sensitive local features. Grad-CAM from this fused representation is refined into lesion pseudo-labels. These labels then supervise a compact three-level U-Net, referred to as Light-UNet. The framework separates image-level disease discrimination from pixel-level lesion-shape learning, while keeping the annotation requirement low.

The study is designed as a characterization of CAM-derived weak supervision rather than as an independent validation of biological lesion boundaries. Four CAM thresholds (0.40, 0.45, 0.50, and 0.55) are tested while the segmentation architecture, loss, and optimization settings are kept fixed. This design examines how pseudo-label granularity affects the ability of Light-UNet to reproduce and regularize the weak-label distribution. The segmentation output is then constrained by a leaf mask and converted into a lesion-to-leaf area ratio. Because neither independently hand-delineated lesion masks nor expert-calibrated absolute severity scores were available, this ratio is treated as a dataset-relative lesion-burden indicator rather than a validated disease-severity measurement.

The main contributions are as follows. First, the study provides a reproducible weakly supervised pipeline that transfers image-level knowledge from a dual-branch wheat disease classifier to pixel-level lesion cues. Second, it formalizes CAM pseudo-label construction through thresholding, morphology, component filtering, and leaf-region constraints. Third, it reports a controlled multi-threshold experiment that characterizes the learnability and agreement of different weak-label granularities under a fixed downstream segmenter. Finally, it provides a post hoc probability-threshold sensitivity analysis and quartile-based descriptive stratification of predicted lesion burden on a 3,430-image profiling split, while explicitly delimiting the conclusions that can be drawn in the absence of independent pixel-level and expert severity ground truth.

2. Related Work

2.1 Deep learning for plant disease recognition and lesion-based severity estimation

Plant disease recognition originally relied on hand-crafted color, texture, and shape descriptors. Deep convolutional networks have since improved robustness to background variation, growth-stage differences, and symptom heterogeneity. Mohanty et al. showed the feasibility of image-based plant disease classification with deep learning[2], and later work by Ferentinos and Too et al. confirmed that modern CNN architectures can perform well across multiple crop-disease combinations[3,4]. These studies established the value of transfer learning and deep feature extraction. Their outputs, however, remained image-level decisions.

Wheat stripe rust has been studied at several spatial scales. Mi et al. treated severity grading as fine-grained classification and introduced attention into DenseNet for mobile images[13]. Schirrmann et al. explored early stripe rust detection using deep residual networks[14,15]. Liu et al. proposed WSRD-Net for oriented detection of elongated stripe rust symptoms[16]. Together, these studies show that stripe rust provides learnable visual cues, although classification and object detection do not directly produce the dense lesion masks needed for area-based severity estimation.

Dense prediction methods offer that missing spatial detail. FCN reformulated semantic segmentation as end-to-end pixel prediction[5], U-Net introduced a symmetric encoder-decoder with skip connections[6], and DeepLab strengthened multi-scale context through atrous convolution and pyramid pooling[7]. In plant pathology, fully supervised segmentation can delineate lesions and calculate severity as the ratio of diseased to total leaf area. Its practical weakness is the cost of mask creation, especially when lesions are numerous, coalesced, translucent, or arranged as narrow stripes. Recent work has also evaluated an RSE-Swin U-Net segmentation framework on wheat disease imagery, including wheat stripe rust, illustrating the continued interest in dense symptom delineation[17].

For wheat stripe rust, Jiang et al. developed image-processing and machine-learning procedures that use lesion-area percentages for severity assessment[18,19]. StripeRust-Pocket further demonstrated the practical value of mobile deep learning for severity quantification under complex backgrounds [20]. These studies support lesion-area-based assessment as an interpretable route to disease quantification. Their supervision and evaluation settings, however, differ from weakly supervised segmentation; direct metric comparisons across datasets would therefore be misleading.

2.2 CAM-based weak supervision, pseudo-label learning, and lightweight segmentation

CAM uses the weights of a global-average-pooled classifier to locate discriminative spatial regions. Grad-CAM extends this idea by weighting convolutional feature maps with target-class gradients. Because it can be applied to a trained network without retraining the classifier, Grad-CAM is a natural tool for reusing classification models. The resulting maps are coarse and decision-driven: they indicate evidence for a class, not necessarily the full lesion boundary.

Weakly supervised segmentation methods have addressed this limitation through seed expansion and boundary constraints[10], semantic affinity propagation[11], and self-supervised equivariance [12]. Grad-CAM++, Score-CAM, and F-CAM improve localization or map resolution in different ways[21-23]. Plant-specific work has also combined classification with self-supervised segmentation or weak localization to reduce annotation

requirements[24,25]. Yet the heatmap-to-mask step remains sensitive to threshold choice, and agricultural studies rarely isolate this factor with a fixed downstream segmenter.

Pseudo-labeling treats model predictions as surrogate annotations[26]. In segmentation, it allows a dedicated dense predictor to be trained even when the original labels are only image-level. The advantage is that the segmentation network can learn local edges, texture continuity, and shape regularities that a thresholded CAM does not encode. The risk is confirmation bias: errors in the pseudo-labels can be learned and amplified. Confidence filtering, morphology, area constraints, and consistency regularization are therefore commonly used to improve reliability.

Lightweight backbones and compact decoders are also important for agricultural deployment. MobileNetV2 reduces computation through inverted residuals and depthwise separable convolution[27], whereas EfficientNet scales network depth, width, and resolution systematically[28]. The present CF2-WRD classifier uses complementary MobileNetV2-like and EfficientNet-B0-like branches, and the downstream Light-UNet uses only three downsampling stages with 32 base channels, giving approximately 1.93 million trainable parameters. The design favors reproducibility and computational accessibility rather than architectural complexity.

2.3 Research gap

Existing studies establish the feasibility of deep learning for stripe rust recognition, fully supervised severity assessment, and CAM-based localization. What remains less clear is how CAM threshold selection affects a lightweight segmenter trained entirely from image-level-derived pseudo-labels. This study addresses that gap by characterizing threshold-dependent weak-label behavior and downstream pseudo-label agreement. Because independent manual lesion masks and expert absolute severity scores are unavailable, the area-based output is deliberately interpreted as relative lesion burden rather than as validated agronomic severity.

3. Materials and Methods

3.1 Dataset and experimental split

The image collection comprised 34,694 RGB images organized into 87 image-level categories. The images covered seedling and adult growth stages and included greenhouse and field acquisition conditions in Gansu Province, China. The wheat plants imaged in this study were cultivated experimental materials grown at the experimental fields of Gansu Agricultural University and the Gansu Academy of Agricultural Sciences, Gansu Province, China. Permission to access the experimental fields and acquire plant images was granted by the responsible management authorities of the respective institutions. All field imaging procedures were conducted in accordance with relevant institutional and national regulations and guidelines. No wild or endangered plant species were involved in this study. The 87 categories represent the image-level class structure used by the classifier; the alphanumeric dataset codes are listed in Supplementary Table S8. The split contained 27,834 training images, 3,430 validation images, and 3,430 test images, corresponding approximately to an 8:1:1 partition. All 87 class codes occurred in each partition. The same split assignment was retained throughout the workflow. The training and validation sets were used to train CF2-WRD and Light-UNet and to select checkpoints. The 3,430-image test split was used for classifier evaluation. For relative lesion-burden analysis, however, a 300-image subset of this split was used in a post hoc probability-threshold sensitivity analysis; therefore, for burden-related analyses the 3,430-image split is treated only as a descriptive profiling split and not as a held-out severity-evaluation set. Split-level sample accounting is provided in Supplementary Table S1, and per-class counts are provided in Supplementary Table S8.

Images were converted to RGB, resized to 224 × 224, transformed into tensors, and normalized with ImageNet statistics (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]). Segmentation masks were resized by nearest-neighbor interpolation to preserve binary labels. No manual pixel-level annotation was used to supervise Light-UNet.

3.2 Overall workflow and CF2-WRD feature fusion

This framework consists of six interrelated operations, as shown in Figure 1. CF2-WRD is first trained for 87-class image recognition. The trained classifier is then frozen, and its fused convolutional feature map is exposed for Grad-CAM. The resulting class-evidence map is normalized, thresholded, and refined to

produce a binary lesion pseudo-label. Light-UNet learns to predict lesion and non-lesion pixels from RGB images. Finally, the predicted lesion mask is intersected with a leaf mask and converted into a lesion-to-leaf area ratio for descriptive relative lesion-burden analysis.

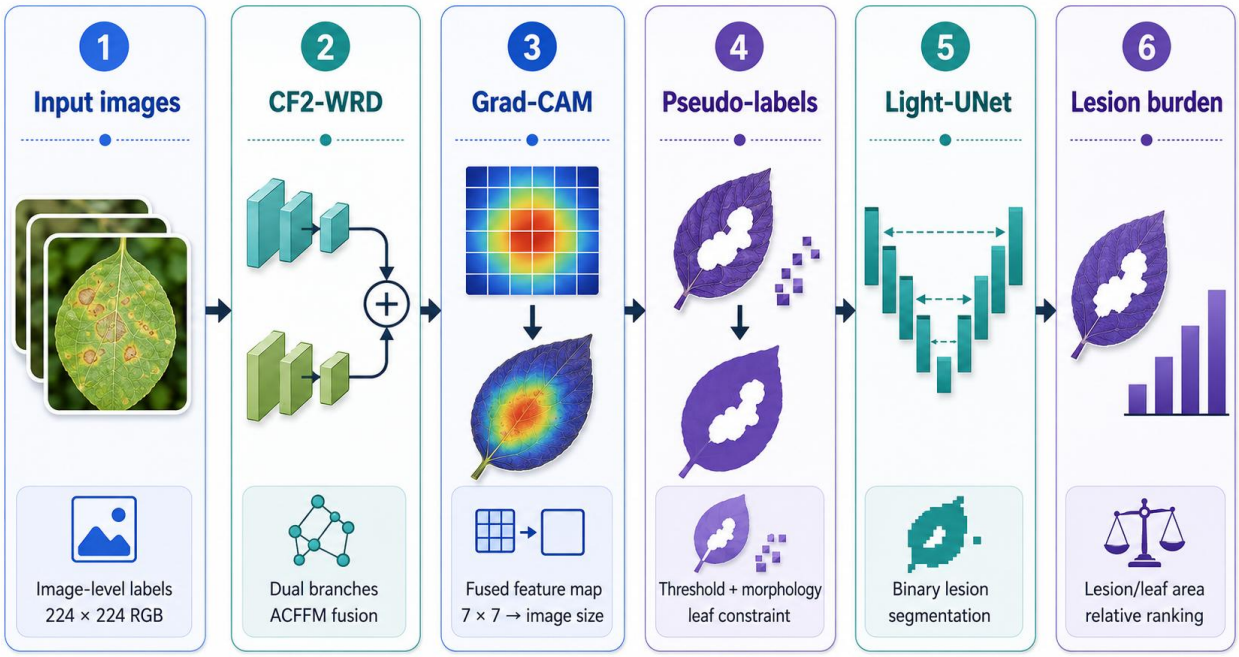


Fig. 1. Weakly supervised classification-to-segmentation framework and relative lesion-burden workflow.

CF2-WRD is a dual-branch classifier designed to combine complementary lightweight features. The first branch follows MobileNetV2 principles and uses inverted residual blocks with expansion ratios of 1 or 6. The second branch follows an EfficientNet-B0-like configuration, using mobile inverted bottleneck convolution and squeeze-excitation. For a 224 × 224 image, each branch outputs a 1,280-channel feature map with spatial dimensions of 7 × 7. A schematic of the classifier is provided in Supplementary Fig. S1.

The two branch outputs are independently projected to 512 channels by 1 × 1 convolution. For each projected branch, global average pooling produces a 512-dimensional vector. L2-normalized copies of these vectors are used to compute a cosine-similarity descriptor, $s = \left\langle \frac{v_m}{||v_m||_2}, \frac{v_e}{||v_e||_2} \right\rangle$. The two unnormalized pooled vectors and s are concatenated and passed through a two-layer gate (1025 → 256 → 2); a softmax produces sample-specific weights w_m and w_e. In the implemented model, these weights act as residual gains rather than direct multiplicative gates:

$$F_m^w = (1 + w_m)F_m, F_e^w = (1 + w_e)F_e \quad (1)$$

The reweighted features are concatenated and modulated by channel attention. A 1 × 1 convolution then reduces the concatenated representation to 512 channels, followed by a 3 × 3 convolution, yielding the fused feature map $F_f \in R^{512\times7\times7}$. Global average pooling, dropout (p = 0.3), and a fully connected layer generate 87-class logits. During CAM extraction, classifier parameters are kept frozen and F_f is used as the target convolutional representation. Principal component dimensions are summarized in Supplementary Table S2.

3.3 Grad-CAM generation and pseudo-label construction

For an input image x and target class c, the classifier produces a score y^c. Grad-CAM assigns a weight to each channel k of F_f by globally averaging the gradient of the class score with respect to that channel:

$$\alpha_k^c = \frac{\left(\frac{1}{Z}\right)\Sigma_{\{I,j\}}^{c}\partial y}{\partial F_f^{k(I,j)}}$$

(2)

where Z is the number of spatial locations. The class-specific activation map is then obtained as:

$$L_{Grad} - CAM^c = \mathrm{ReLU}\left(\Sigma_k \alpha_k^c F_f^k\right)$$

(3)

The 7 × 7 activation map is min-max normalized to [0,1] and bilinearly resized to the original image dimensions. For all CAMs reported in this study, the classifier-predicted image class was used as the Grad-CAM target. This preserves the end-to-end classifier-to-localization pathway but also permits classification errors to propagate into pseudo-label generation. Extracting Grad-CAM from the fused feature map, rather than from either branch alone, uses the representation produced by ACFFM.

Let H(x) denote the normalized CAM value at pixel x, and let τ denote the CAM binarization threshold. The initial lesion pseudo-label $M\tau(x)$ is defined as 1 when $H(x) \geq \tau$ and 0 otherwise. Four thresholds, $\tau \in \{0.40, 0.45, 0.50, 0.55\}$, were evaluated. All other data-processing and Light-UNet training settings were fixed, so downstream differences mainly reflect pseudo-label granularity.

Fixed CAM thresholds were used to isolate threshold effects; adaptive histogram thresholding methods such as Otsu thresholding[29] were not used.

The binary map was refined by 3 × 3 morphological opening to remove isolated noise and 5 × 5 closing to fill small holes. Eight-connected components smaller than 80 pixels were removed. A leaf mask was independently estimated in HSV space using a broad hue interval of 10-110 and saturation/value lower bounds of 20, followed by morphology. The lesion mask was intersected with the leaf mask so that activations outside the visible leaf could not become positive pseudo-labels. At the sample level, training masks with a total positive area below 50 pixels were excluded from segmentation training. The pseudo-label workflow is illustrated in Supplementary Fig. S2.

The training pool contained 27,834 images. For each CAM-threshold condition, training pseudo-label masks with fewer than 50 positive pixels were excluded before Light-UNet training. The retained training-sample counts were 27,803, 27,774, 27,744, and 27,695 for CAM thresholds 0.40, 0.45, 0.50, and 0.55, respectively, corresponding to exclusions of 31, 60, 90, and 139 masks. Empty masks, which form a subset of the excluded <50-pixel masks, numbered 20, 28, 45, and 78, respectively. Supplementary Table S9 provides the complete threshold-specific sample accounting.

Three thresholds are therefore deliberately separated. The CAM threshold τ controls pseudo-label construction. The connected-component and sample-area thresholds remove isolated artifacts. A separate segmentation probability threshold is used only when Light-UNet probabilities are converted into masks for relative lesion-burden calculation.

3.4 Light-UNet architecture and training objective

Light-UNet is a compact two-class U-shaped segmentation network derived from U-Net. Its encoder contains an initial double-convolution block and three downsampling blocks, with channel widths of 32, 64, 128, and 256. Each double-convolution block applies two 3 × 3 convolutions, each followed by batch normalization[30] and ReLU. Downsampling is performed by 2 × 2 max pooling. The architecture schematic is provided in Supplementary Fig. S3.

The decoder contains three transposed-convolution upsampling blocks. At each resolution, the upsampled decoder feature is concatenated with the corresponding encoder feature through a skip connection and then processed by a double-convolution block. A final 1 × 1 convolution produces two logits per pixel: background/non-lesion and lesion. The network has approximately 1.93 million trainable parameters, making it substantially smaller than a high-capacity encoder-decoder while retaining full-resolution output.

The segmentation objective combines pixel-wise cross-entropy with a lesion-class Dice loss[31]. For logits z and binary target mask y, cross-entropy is computed over the two classes. Let denote the softmax probability of the lesion class at pixel i. Dice loss is defined as:

$$L_{Dice} = 1 - \frac{2\Sigma_i p_i y_i + \varepsilon}{\Sigma_i p_i + \Sigma_i y_i + \varepsilon} \quad (4)$$

Here ε is a numerical smoothing constant used to avoid division by zero; the implemented Dice loss used $\varepsilon = 1 \times 10^{-6}$. The total loss is:

$$L = L_{CE} + L_{Dice} \quad (5)$$

Cross-entropy provides stable pixel-wise discrimination, whereas Dice loss reduces the dominance of the background class and directly optimizes lesion overlap.

3.5 Evaluation metrics and scope of relative lesion-burden analysis

Lesion IoU, Dice, mean IoU, and pixel accuracy were calculated from Light-UNet predictions and validation pseudo-labels. If TP, FP, and FN denote lesion-pixel counts, lesion IoU and Dice are:

$$IoU_{lesion} = \frac{TP}{TP + FP + FN} \quad (6)$$

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (7)$$

Background IoU was calculated analogously, and mIoU was the mean of background and lesion IoU. Pixel accuracy was the proportion of correctly classified pixels among all pixels. In the implementation, 1×10^{-6} was added to the IoU and Dice denominators for numerical stability. Because the validation masks were generated by the same CAM pipeline as the training pseudo-labels, the reported segmentation metrics quantify agreement with the CAM-derived weak-supervision distribution rather than accuracy against independently hand-delineated lesion masks. Moreover, the validation target changes with the CAM threshold; therefore, cross-threshold differences indicate how readily each pseudo-label distribution is learned and reproduced under the fixed Light-UNet setting, but they do not establish which threshold is closest to the true biological lesion boundary.

No independent manual lesion masks or expert-calibrated severity scores were used in model selection or evaluation. Accordingly, the study does not report sensitivity, specificity, boundary accuracy, or area error relative to biological ground truth, and it does not numerically rank the proposed method against fully supervised segmentation systems evaluated on manual masks. The 3,430-image lesion-burden analysis is descriptive. Because probability-threshold sensitivity was explored on 300 images from that same split, the burden analysis does not constitute a held-out test of agronomic severity accuracy; it summarizes only the distribution and ranking produced by the weakly supervised pipeline under the stated operating point.

For relative lesion-burden profiling, Light-UNet lesion probabilities were thresholded and constrained by the leaf mask. Let A_lesion be the number of predicted lesion pixels within the leaf and A_leaf the number of leaf-mask pixels. Relative lesion burden was calculated as:

$$S = 100 \times \frac{A_{lesion}}{A_{leaf}(\%)} \quad (8)$$

A post hoc sensitivity analysis compared two probability-threshold endpoints, 0.50 and 0.80, on 300 randomly selected images from the 3,430-image classifier test split. The 0.50 endpoint produced spatially more expansive lesion predictions, whereas 0.80 substantially reduced the mean and median predicted burden. The 0.80 endpoint was then used as an exploratory conservative operating point for descriptive burden profiling; it was not optimized against independent biological ground truth. Because this operating point was informed by images from the same 3,430-image split, that split is not treated as a held-out severity-evaluation set in the present study. Instead, it is referred to as a profiling split for burden-related analyses. Its images were divided by the empirical first, second, and third quartiles of S only to summarize dataset-relative ranking.

3.6 Implementation details

Experiments were implemented in PaddlePaddle 2.x with Python 3.10 on Baidu AI Studio. GPU resources were dynamically allocated as NVIDIA Tesla V100 or T4 devices; the host environment used Intel Xeon CPUs and approximately 32 GB memory. Light-UNet training was configured for a maximum of 30 epochs with a batch

size of 16. AdamW, a decoupled-weight-decay variant of Adam [32,33], was used with a fixed learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The training loader used shuffling and drop_last=True, whereas the validation loader was not shuffled. For each CAM threshold, the checkpoint with the highest validation mIoU was selected, and all metrics reported for that threshold were taken from that same checkpoint. Except for the CAM threshold, all ablation settings were held constant. Additional implementation settings are summarized in Supplementary Table S3.

CF2-WRD was trained for 30 epochs with AdamW (learning rate 1×10^{-4}; weight decay 1×10^{-4}), a batch size of 16, and unweighted cross-entropy loss. Training images were resized to 224 × 224, randomly horizontally flipped, randomly rotated within ±15°, converted to tensors, and normalized with channel means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Validation and test images used resizing and the same normalization without random augmentation. The training loader used shuffle=True and drop_last=True, whereas validation and test loaders were not shuffled. A last checkpoint was saved after every epoch, and cf2_wrd_best.pdparams was updated whenever validation accuracy exceeded the previous best. The highest validation accuracy, 0.8776, occurred at epoch 29; that checkpoint was used for subsequent test evaluation and CAM extraction.

4. Results

4.1 Classification performance and CAM behavior

The frozen CAM generator first needs stable image-level discrimination. During 30-epoch classifier training, CF2-WRD reached its highest validation accuracy of 0.8776 at epoch 29; the corresponding best checkpoint was then used for test evaluation and achieved an accuracy of 0.8711. The 0.65-percentage-point difference indicates limited degradation from validation to test partitions. This result supports the use of the fused feature representation for spatial explanation, although image-level accuracy alone cannot guarantee complete lesion localization. The classifier summary is provided in Supplementary Table S4.

Figure 2 compares the spatial responses obtained from the MobileNetV2-like branch, the EfficientNet-B0-like branch, and the ACFFM fused feature representation. The branch-specific maps show different response patterns. The MobileNetV2-like branch generally produces broader activation regions, suggesting stronger sensitivity to global plant configuration and large-scale context. The EfficientNet-B0-like branch more often generates localized or discontinuous responses, indicating greater sensitivity to local texture, organ boundaries, and high-contrast structures.

In the correctly classified examples in Fig. 2a-c, the fused-feature CAM integrates the dominant responses of the two branches and forms a more coherent activation pattern. This behavior is consistent with the intended role of the cosine-similarity-based complementary gate and ACFFM. Even so, the fused activation is not confined to visually identifiable lesion tissue. In images with overlapping leaves, spikes, shadows, or senescent structures, the response can extend to non-lesion plant regions. The CAM is therefore treated as a weak class-discriminative cue, not as a manually verified lesion mask. The low-confidence misclassified example in Fig. 2d illustrates the error-propagation risk of classifier-derived supervision. The sample was assigned to class 11b instead of the ground-truth class 8b, with a confidence of 0.610; the branch-specific and fused CAMs also focus on different canopy regions. This case illustrates error propagation and motivates the leaf-region, connected-component, and area constraints used during pseudo-label construction; confidence-aware sample weighting or filtering was not used in the reported experiments.

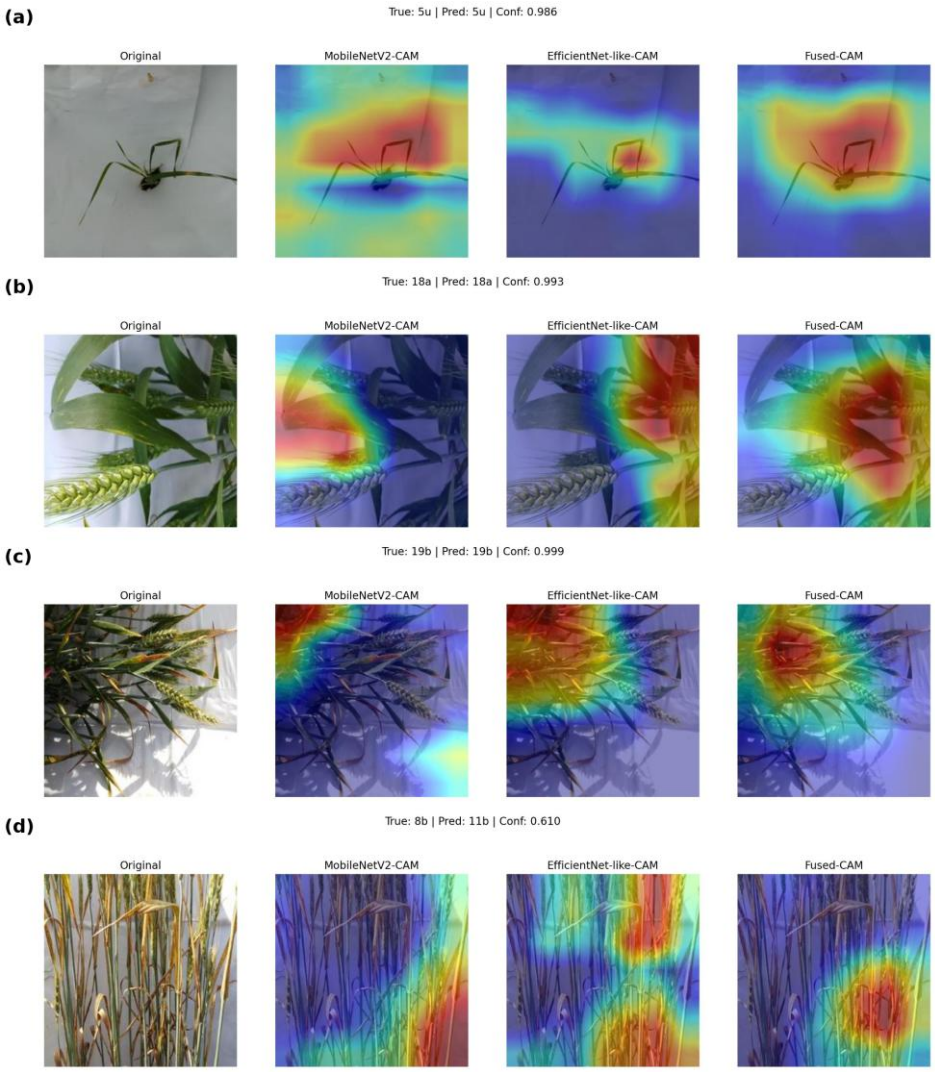


Fig. 2. Qualitative comparison of branch-specific and fused-feature Grad-CAM responses. Each row shows the original image, the MobileNetV2-like CAM, the EfficientNet-B0-like CAM, and the CAM extracted from the ACFFM fused feature map. Rows (a)-(c) are correctly classified examples under different growth stages and background conditions; row (d) is a low-confidence misclassified case. Red denotes relatively high class-discriminative response, and blue denotes low response. Heatmaps are independently normalized; warmer colors indicate higher relative class-discriminative activation.

4.2 CAM-threshold ablation of pseudo-label agreement

Table 1 summarizes the CAM-threshold control experiment. For each threshold, the reported row corresponds to the checkpoint with the highest validation mIoU, and all metrics in that row were taken from the same checkpoint. The highest pseudo-label agreement metrics on the validation split were obtained at $\tau = 0.40$. As τ increased, mIoU, Dice, and lesion IoU all declined. Validation mIoU decreased from 0.6584 at 0.40 to 0.6312 at 0.55, an absolute reduction of 0.0272. Over the same interval, Dice decreased by 0.0896 and lesion IoU by 0.0998. Because each threshold defines its own CAM-derived validation target, these differences describe threshold-dependent learnability and agreement rather than independently validated biological segmentation accuracy.

Table 1. Effect of CAM binarization threshold on Light-UNet pseudo-label validation performance.

CAM τ	Completed epochs	Selected epoch	Train mIoU	Valid mIoU	Valid Dice	Valid lesion IoU
0.40	30	14	0.7004	0.6584	0.7101	0.5555
0.45	30	17	0.6963	0.6490	0.6739	0.5136
0.50	30	20	0.6904	0.6415	0.6480	0.4855
0.55	30	18	0.6756	0.6312	0.6206	0.4557

For each threshold, the selected epoch is the checkpoint with the highest validation mIoU, and all reported metrics in that row are taken from the same checkpoint. Values are from a single training run per threshold; no between-run uncertainty estimates are available.

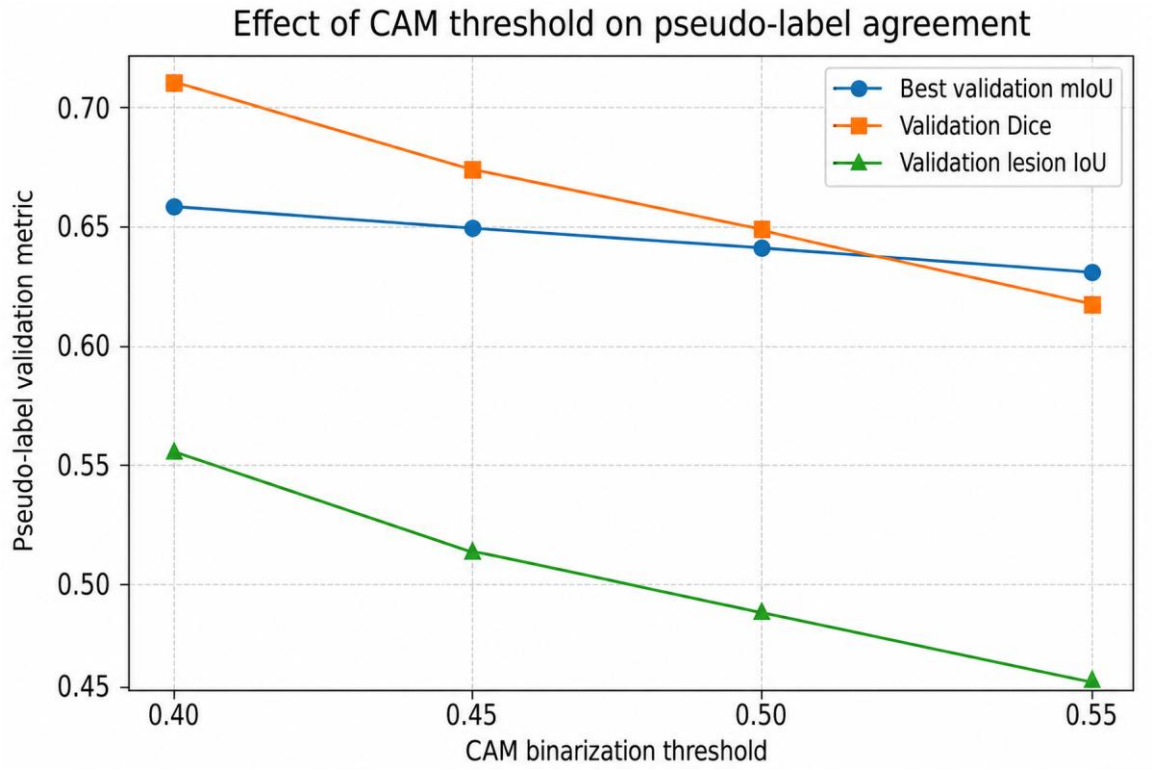


Fig. 3. CAM-threshold ablation of pseudo-label agreement using metrics from the validation-mIoU-selected checkpoint of each threshold run. All segmentation settings except the pseudo-label threshold were fixed. Each point represents a single training run; error bars are not shown because replicate-run estimates were not available.

The monotonic trend is consistent with the spatial role of τ within the weak-label system. At 0.40, lower-confidence peripheral responses remain in the pseudo-labels, increasing spatial coverage and agreement with a segmenter that can smooth local regions. At 0.50 and 0.55, the pseudo-labels increasingly represent only high-response cores, providing less supervision for peripheral regions and scattered responses. The experimental runs for τ = 0.40, 0.45, 0.50, and 0.55 each completed 30 training epochs, the validation-mIoU-selected checkpoints occurred at epochs 14, 17, 20, and 18. Thus, the comparison uses the validation-mIoU-selected checkpoint for each run rather than the final training epoch. This trend should not be interpreted as evidence that the broader 0.40 masks are biologically more accurate.

Although τ = 0.40 maximized agreement with the corresponding pseudo-label validation distribution, τ = 0.45 was retained as an exploratory conservative operating point for subsequent relative-burden analysis. The lower threshold generated visually broader activation and was expected to include more low-confidence non-lesion responses. Threshold 0.45 therefore provided a pragmatic compromise between spatial coverage and background suppression within the weak-supervision pipeline. At this operating point, Light-UNet achieved pseudo-label agreement values of lesion IoU 0.5136, Dice 0.6739, mIoU 0.6490, and pixel accuracy 0.8306 at epoch 17. These values are not independent lesion-segmentation accuracy estimates.

This selection separates numerical pseudo-label agreement from descriptive operating-point choice. Because both training and validation masks are CAM-derived, a permissive CAM threshold can increase agreement with its own broader weak-label distribution. A slightly higher threshold was therefore used to reduce the influence of low-confidence activation in downstream burden profiling. Accordingly, 0.40 is reported as the numerical optimum for pseudo-label agreement, whereas 0.45 is treated only as a conservative operating point and not as a biologically validated optimum.

Representative pseudo-label construction examples at the selected CAM threshold are shown in Fig. 4. For correctly classified samples, the fused-feature CAM generally produced compact activation that could be

converted into spatially coherent candidate masks after thresholding and refinement. The masks were still coarse and occasionally covered healthy leaf, spike, or background structures. The low-confidence misclassified sample generated a particularly large and spatially biased pseudo-mask, illustrating how classification errors can propagate into weak pixel supervision. These examples support the use of conservative threshold selection, morphology, leaf-region constraints, connected-component filtering, and sample-level quality control.

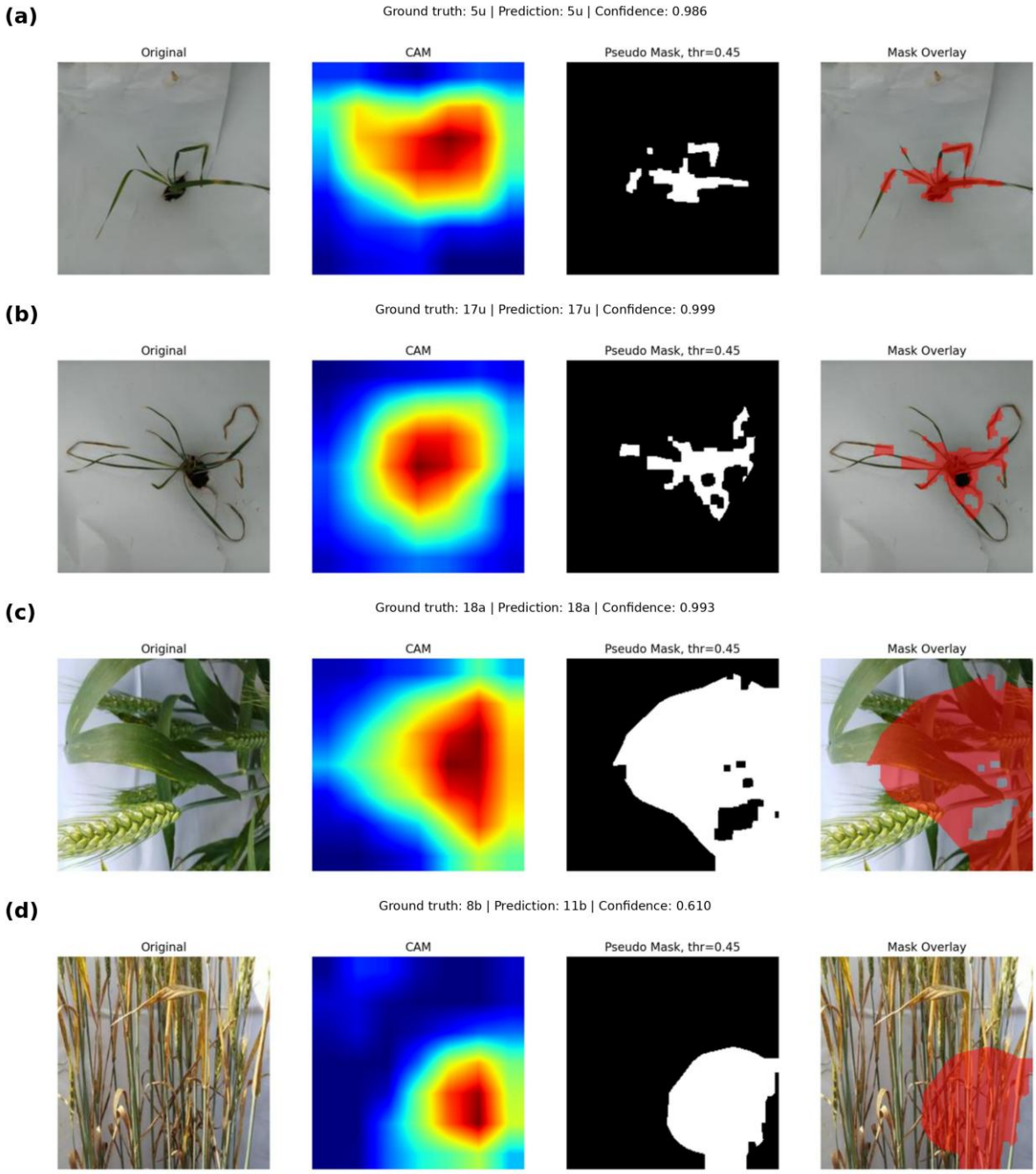


Fig. 4. Representative CAM-derived pseudo-labels at the selected operating threshold ($\tau = 0.45$). Each row shows the original RGB image, the normalized fused-feature Grad-CAM, the thresholded and refined pseudo-mask, and the corresponding mask overlay. Rows (a)-(c) are correctly classified examples under different visual conditions; row (d) is a low-confidence misclassified case illustrating pseudo-label noise and spatial uncertainty.

4.3 Probability-threshold sensitivity and descriptive lesion-burden profiling

The segmentation probability threshold used for relative-burden profiling was examined separately from the CAM threshold. In a 300-image subset drawn from the 3,430-image profiling split, a probability threshold of 0.50 produced a mean predicted lesion-to-leaf ratio of 55.02% and a median of 52.90%; 98.33% of samples exceeded 30% predicted burden. Raising the probability threshold to 0.80 reduced the mean to 29.63% and the

median to 31.39%. The proportion above 30% fell to 53.00%, and no sample exceeded 80%. These changes demonstrate strong sensitivity of the derived lesion-to-leaf ratio to probability thresholding; without independent lesion or severity ground truth, they do not establish that 0.80 is more biologically accurate than 0.50. Because the exploratory operating point was informed by this subset, the subsequent 3,430-image analysis is descriptive profiling rather than held-out severity testing. Supplementary Table S5 reports the descriptive endpoint summary.

The 3,430-image profiling split was processed using CAM threshold 0.45 for pseudo-label training and segmentation probability threshold 0.80 for descriptive inference. The empirical quartiles of predicted lesion burden were 15.7567%, 27.7772%, and 37.6485%. These boundaries produced four nearly equal relative-burden groups: 858 low, 857 medium, 857 high, and 858 very-high images. Equal frequencies are expected because the groups are defined by quartiles. These results should therefore be read only as a stratified ranking of model-predicted burden within this dataset, not as a held-out estimate of agronomic severity performance, evidence that the population naturally contains equal disease classes, or evidence that the quartile boundaries correspond to biological severity grades. The quartile groups are summarized in Table 2.

Table 2. Quartile-based relative lesion-burden groups for the 3,430-image profiling split.

Relative level	Definition using predicted S	Images	Proportion
Low	$S \leq 15.76\%$	858	25.01%
Medium	$15.76\% < S \leq 27.78\%$	857	24.99%
High	$27.78\% < S \leq 37.65\%$	857	24.99%
Very high	$S > 37.65\%$	858	25.01%

The continuous distribution of predicted lesion burden is shown in Fig. 5. The distribution was unimodal, with the highest image frequency at approximately 25%-40%, followed by a progressively thinning right tail extending beyond 80%. This pattern is more informative than a bar chart of the four quartile groups, whose counts are predetermined by the grouping rule. Representative overlays from the four relative-burden groups are shown in Fig. 6. The examples demonstrate increasing predicted mask coverage across the dataset ranking, while also revealing substantial variation in plant architecture, background complexity, and spatial compactness within each group.

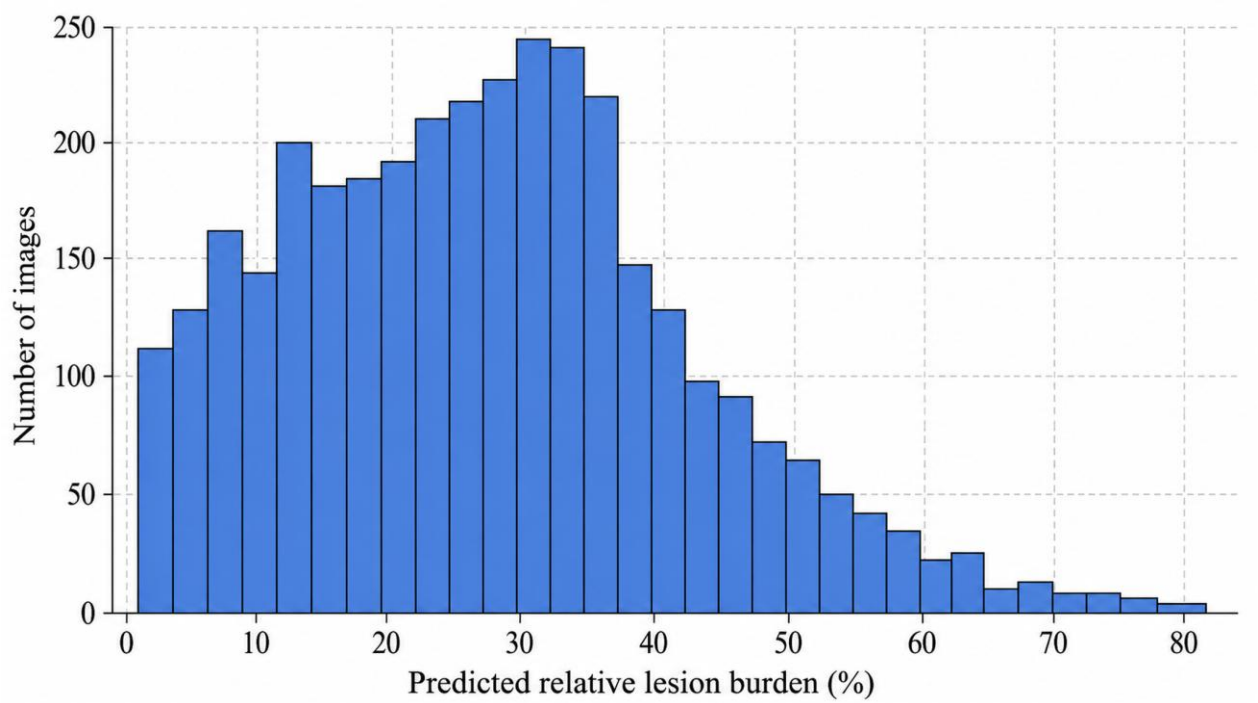


Fig. 5. Distribution of predicted lesion-to-leaf relative burden for the 3,430-image profiling split. Predictions were generated using CAM threshold 0.45 for pseudo-label training and segmentation probability threshold 0.80 for descriptive relative-burden inference. Because the probability operating point was explored on a subset of the same split, this distribution is descriptive and is not a held-out severity-performance evaluation.

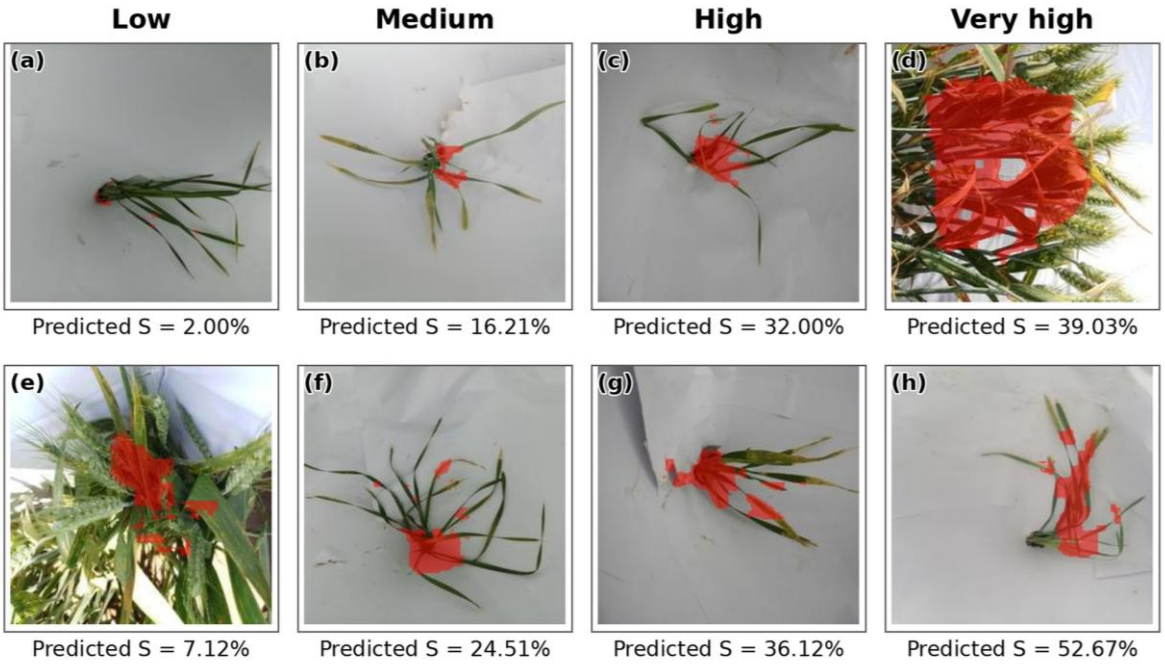


Fig. 6. Representative predicted lesion overlays from the quartile-based relative-burden groups. Columns correspond to low, medium, high, and very-high groups; two examples are shown for each group. The percentages are model-derived lesion-to-leaf ratios and should be interpreted only as dataset-relative burden indicators, not as independently validated agronomic severity scores.

4.4 Comparison with representative stripe rust methods

Direct numerical comparison with published stripe rust methods is limited by differences in task definition, annotation type, image background, and evaluation labels. Supplementary Table S6 compares supervision and output rather than ranking methods by metric alone. Fully supervised or model-assisted methods can provide stronger absolute severity validation. The present method instead emphasizes low annotation cost and a controlled analysis of CAM-derived supervision.

5. Discussion

5.1 Threshold behavior and reuse of CF2-WRD

The ablation results show that lowering the CAM threshold from 0.55 to 0.40 improved all reported overlap metrics. This pattern is consistent with the known behavior of CAM. The strongest activation is usually concentrated on highly discriminative lesion regions, while weaker responses often extend over peripheral pustules and less saturated symptoms. A high threshold therefore creates precise but incomplete seeds. When those seeds supervise Light-UNet, the model receives little information about lesion edges and low-contrast areas. A lower threshold supplies broader spatial support and allows the encoder-decoder to learn continuity across nearby lesion pixels.

The benefit of lower thresholds should not be read as proof that every additional activated pixel is a true lesion. The validation masks were produced by the same CAM threshold as the training masks. A model trained on broader pseudo-labels is therefore evaluated against broader pseudo-labels. The observed improvement indicates greater learnability and agreement within the weak-supervision system; it is not independent evidence of better biological segmentation. For this reason, the main pipeline used 0.45 only as a conservative operating point for descriptive relative-burden profiling even though 0.40 achieved the highest pseudo-label agreement mIoU.

Reusing CF2-WRD avoids training a separate localization classifier and preserves information already learned from 27,834 image-level examples. The MobileNetV2-like branch provides an efficient global representation, whereas the EfficientNet-like branch contributes MBConv and squeeze-excitation features. The cosine-similarity-based gate and ACFFM produce a fused representation whose gradients can be converted into spatial cues. This pathway is useful when image-level labels are abundant but pixel masks are scarce.

The qualitative comparison in Fig. 2 provides visual evidence for this complementarity. MobileNetV2-like responses tend to be spatially broad; EfficientNet-B0-like responses are often more localized. Their fused representation consolidates these cues, although the failure case also shows that fusion cannot by itself correct an erroneous image-level target. Classifier confidence and CAM plausibility therefore remain important quality-control signals for downstream pseudo-labeling.

The classification accuracies of 0.8776 and 0.8711 indicate that the fused representation generalizes at the image level. Still, classification errors can propagate into pseudo-label generation. Future implementations should record class confidence, branch weights, and CAM area, then either exclude uncertain images or reduce their loss contribution. Confidence-aware weighting would make the weak-supervision pathway less vulnerable to mislabeled or visually ambiguous samples. In addition, the reported CAMs were targeted to the classifier-predicted class rather than the known image-level label, so misclassification could directly redirect the spatial supervision. A ground-truth-target CAM comparison was not performed in this study.

5.2 Role of Light-UNet under weak supervision

Light-UNet provides a dedicated pixel predictor rather than relying on thresholded CAM alone. Its skip connections recover local structure, and the combined cross-entropy and Dice objective mitigates background dominance. The compact architecture may also reduce the tendency to fit high-frequency pseudo-label noise relative to a very large segmenter. The best validation mIoU of 0.6584 therefore demonstrates that the network can reproduce and spatially regularize the CAM-derived target distribution; it does not, by itself, demonstrate lesion-boundary accuracy against biological ground truth.

The pseudo-label examples in Fig. 4 clarify this role. Light-UNet functions as a regularizer, not as an independent source of lesion truth. It can smooth fragmented CAM seeds and learn local continuity, but broad or misplaced pseudo-masks still define the spatial information available during training.

Light-UNet also cannot recover lesion information that is systematically absent from all pseudo-labels. If CAM never responds to faint infection, or if the leaf mask removes true brown or yellow tissue, the segmenter receives no positive supervision for those regions. Conversely, persistent activation on veins, shadows, or leaf margins can become learned false positives. Architectural changes alone cannot resolve this identifiability problem. A small manually annotated calibration subset would be the most direct improvement.

5.3 Interpreting relative lesion burden

The lesion-to-leaf ratio is an interpretable image-derived quantity and is conceptually related to established area-based severity approaches. In the present study, however, it remains a model-derived relative-burden indicator because neither lesion masks nor expert severity scores were available for calibration. The probability-threshold sensitivity experiment further shows that the numerical ratio depends strongly on output thresholding. A threshold of 0.50 produced more spatially expansive predictions, whereas 0.80 produced a more conservative distribution. This difference reflects threshold sensitivity and should not be interpreted as evidence that either threshold is biologically correct.

The histogram in Fig. 5 shows that the continuous predictions were not uniformly distributed, even though the quartile groups contained nearly equal numbers of images. This distinction matters. Quartile labels preserve rank but discard distributional shape. The overlays in Fig. 6 further show that similar numerical ratios can arise from different spatial patterns, including compact central activation, multiple disconnected regions, and broad masks in crowded canopies. Downstream use should therefore retain the continuous ratio and mask visualization whenever possible, rather than relying only on the four group labels.

The quartile boundaries are useful for selecting representative low- and high-burden images, stratifying downstream experiments, and prioritizing manual review. They are not equivalent to agronomic severity standards, because quartile classes are distribution-dependent and no expert ground truth was used to calibrate the area estimates. Until independent validation is available, S should therefore be treated as an internal ranking variable rather than as an absolute disease-severity measurement. External validation should compare predicted area ratios with manually delineated lesion and leaf masks and independent expert scores, report R^2, MAE, RMSE, and inter-rater reliability where appropriate, and test whether fixed biological thresholds transfer across growth stages and acquisition conditions.

5.4 Practical implications, limitations, and future work

The workflow offers a cost-efficient route for research groups that already possess an image classifier and large image-level datasets. It produces masks that are suitable for visualization, approximate lesion-burden screening, dataset stratification, and human-in-the-loop annotation. The framework is also modular: Grad-CAM can be replaced by Score-CAM, Grad-CAM++, or full-resolution CAM, and Light-UNet can be replaced by another compact decoder. The threshold ablation provides a general template for selecting the weak-label operating point rather than adopting a single arbitrary value.

In the current experiments, each CAM-threshold condition was evaluated as a single training run; repeated experiments with controlled random seeds would therefore be useful for quantifying run-to-run variability. Because Grad-CAM was generated for the classifier-predicted category, improvements in classification robustness are also expected to benefit downstream pseudo-label quality. In addition, the HSV-based leaf constraint provides a simple and computationally inexpensive means of suppressing background activations, although more adaptive leaf-segmentation methods may further improve robustness under complex illumination, senescence, overlapping foliage, and field backgrounds.

Future work will focus on independent validation using a stratified manually annotated subset covering different growth stages, background conditions, and lesion-burden levels. Such a reference subset could include both lesion and leaf masks, allowing direct evaluation of segmentation overlap and lesion-to-leaf area estimation. Expert disease assessments could further support correlation and calibration analyses between image-derived lesion burden and agronomic severity. Methodologically, future studies may compare fused-feature Grad-CAM with branch-specific CAM, Grad-CAM++, Score-CAM, and F-CAM; evaluate repeated training runs; investigate iterative pseudo-label refinement and self-training; and explore more illumination- and color-robust leaf extraction. These extensions would further strengthen the transition from weakly supervised lesion localization toward practical field-oriented disease assessment.

6. Conclusion

This study developed a weakly supervised wheat stripe rust lesion-segmentation framework that reuses a dual-branch CF2-WRD classifier as a Grad-CAM generator and trains a compact Light-UNet from CAM-derived pseudo-labels. CF2-WRD achieved validation and test accuracies of 0.8776 and 0.8711. In the controlled CAM-threshold experiment, threshold 0.40 produced the highest agreement with its corresponding CAM-derived validation targets at the validation-mIoU-selected checkpoint (mIoU 0.6584, Dice 0.7101, lesion IoU 0.5555; epoch 14), while 0.45 was used as a more conservative exploratory operating point. A post hoc probability-threshold sensitivity analysis showed that the model-derived lesion-to-leaf ratio was strongly threshold dependent. Because that analysis used a subset of the same 3,430-image split subsequently summarized by empirical quartiles at 15.76%, 27.78%, and 37.65%, those burden results are presented only as descriptive profiling and not as held-out severity performance. The study therefore demonstrates a reproducible low-annotation route from image-level recognition to spatial lesion cues and dataset-relative burden ranking. Because no independent manual lesion masks or expert-calibrated severity labels were available, the current overlap metrics should be interpreted as weak-label agreement rather than validated lesion accuracy, and the burden values should not be interpreted as agronomic severity measurements. A confirmatory severity evaluation will require selecting the probability operating point on the validation split or an independent calibration subset before testing it on an untouched reference set.

Acknowledgements

The authors gratefully acknowledge the greenhouse of Gansu Academy of Agricultural Sciences for supporting the collection of wheat seedling samples, and the demonstration base for breeding new rust-resistant and high-yield winter wheat varieties and supporting cultivation techniques in Qingshui County, Gansu Agricultural University, for providing conditions for collection of adult wheat plant samples used in this study.

Author contributions

B.G. and X.L. contributed to conceptualization, data curation, and formal analysis. B.G. conducted the investigation and provided resources. B.G., X.L., and A.K.. developed the methodology. T.N.I. was responsible for project administration and supervision. X.L. and A.K. performed validation. B.G. and X.L. wrote the original manuscript draft. All authors reviewed and edited the manuscript and approved the final version.

Funding

No external funding was received for this study.

Competing interests

The authors declare no competing interests.

Data availability

Data supporting the reported quantitative analyses, including the 87-class mapping, image-level train/validation/test split manifests, CF2_WRD_CAM_LightUNet training logs, CAM-threshold ablation logs, threshold-specific pseudo-label sample counts, per-class split counts, and derived summary tables, have been provided in the associated GitHub repository: https://github.com/42328540G/CF2_WRD_CAM_LightUNet.

The wheat stripe rust image dataset generated and analyzed during the current study is not publicly deposited in an open-access repository. The dataset contains images of healthy and stripe-rust-infected wheat collected across seedling and adult growth stages under controlled and field conditions. The data are available from the corresponding author upon reasonable request for academic and non-commercial research purposes.

Code availability

The source code supporting the implementation of CF2_WRD_CAM_LightUNet is publicly available at: https://github.com/42328540G/CF2_WRD_CAM_LightUNet. The repository includes code for image preprocessing, complementary dual-branch feature extraction, cosine-similarity-based complementary gating, adaptive complementary feature fusion, Grad-CAM pseudo-label generation, Light-UNet training, evaluation, inference, and feature visualization. Configuration files and documentation required to reproduce the computational workflow are also provided in the repository.

References

[1] Chen, W., Wellings, C., Chen, X., Kang, Z. & Liu, T. Wheat stripe (yellow) rust caused by Puccinia striiformis f. sp. tritici. *Mol. Plant Pathol.* **15,** 433–446 (2014).

[2] Mohanty, S. P., Hughes, D. P. & Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **7,** 1419; 10.3389/fpls.2016.01419 (2016).

[3] Ferentinos, K. P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **145,** 311–318 (2018).

[4] Too, E. C., Li, Y., Njuki, S. & Liu, Y. A comparative study of fine-tuning deep learning models for plant disease identification. *Comput. Electron. Agric.* **161,** 272–279 (2019).

[5] Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 3431–3440 (2015).

[6] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer, 2015).

[7] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40,** 834–848 (2018).

[8] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 2921–2929 (2016).

[9] Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc. IEEE Int. Conf. Comput. Vis. 618–626 (2017).

[10] Kolesnikov, A. & Lampert, C. H. Seed, expand and constrain: three principles for weakly-supervised image segmentation. In *Computer Vision – ECCV 2016* 695–711 (Springer, 2016).

[11] Ahn, J. & Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 4981–4990 (2018).

[12] Wang, Y., Zhang, J., Kan, M., Shan, S. & Chen, X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. 12275–12284 (2020).

[13] Mi, Z., Zhang, X., Su, J., Han, D. & Su, B. Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Front. Plant Sci.* **11,** 558126; 10.3389/fpls.2020.558126 (2020).

[14] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 770–778 (2016).

[15] Schirrmann, M., Landwehr, N., Giebel, A., Garz, A. & Dammer, K.-H. Early detection of stripe rust in winter wheat using deep residual neural networks. *Front. Plant Sci.* **12,** 469689; 10.3389/fpls.2021.469689 (2021).

[16] Liu, H. et al. WSRD-Net: a convolutional neural network-based arbitrary-oriented wheat stripe rust detection method. *Front. Plant Sci.* **13,** 876069; 10.3389/fpls.2022.876069 (2022).

[17] Zang, H. et al. Constructing segmentation method for wheat powdery mildew using deep learning. *Front. Plant Sci.* **16,** 1524283; 10.3389/fpls.2025.1524283 (2025).

[18] Jiang, Q., Wang, H. & Wang, H. Two new methods for severity assessment of wheat stripe rust caused by Puccinia striiformis f. sp. tritici. *Front. Plant Sci.* **13,** 1002627; 10.3389/fpls.2022.1002627 (2022).

[19] Jiang, Q., Wang, H. & Wang, H. Severity assessment of wheat stripe rust based on machine learning. *Front. Plant Sci.* **14,** 1150855; 10.3389/fpls.2023.1150855 (2023).

[20] Liu, W. et al. StripeRust-Pocket: a mobile-based deep learning application for efficient disease severity assessment of wheat stripe rust. *Plant Phenomics* **6,** 0201; 10.34133/plantphenomics.0201 (2024).

[21] Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. IEEE Winter Conf. Appl. Comput. Vis. 839–847 (2018).

[22] Wang, H. et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. Workshops 111–119 (2020).

[23] Belharbi, S. et al. F-CAM: full resolution class activation maps via guided parametric upscaling. Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. 3727–3736 (2022).

[24] Chen, J., Guo, J., Zhang, H., Liang, Z. & Wang, S. Weakly supervised localization model for plant disease based on Siamese networks. *Front. Plant Sci.* **15,** 1418201; 10.3389/fpls.2024.1418201 (2024).

[25] Franchetti, B. & Pirri, F. Detection and localization of tip-burn on large lettuce canopies. *Front. Plant Sci.* **13,** 874035; 10.3389/fpls.2022.874035 (2022).

[26] Lee, D.-H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop: Challenges in Representation Learning (WREPL) (2013).

[27] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: inverted residuals and linear bottlenecks. Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 4510–4520 (2018).

[28] Tan, M. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. *Proc. Mach. Learn. Res.* **97,** 6105–6114 (2019).

[29] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9,** 62–66 (1979).

[30] Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proc. Mach. Learn. Res.* **37,** 448–456 (2015).

[31] Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26,** 297–302 (1945).

[32] Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Proc. 3rd Int. Conf. Learn. Represent. (2015).

[33] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Proc. 7th Int. Conf. Learn. Represent. (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf