plants

MDPI

Article

Cross-Domain Beech Leaf Disease Recognition via Orthogonal Projection and Pseudo-Label Self-Training

Weiguo Lai [1,2], Daming Huang [3], Li Wang [1,2,*], Wenjun Huo [4], Ji Li [1,2] and Hao Yu [1]

[1] School of Computer Science, Shandong Xiehe University, Jinan 250109, China
[2] Jinan Engineering Research Center of Intelligent Elderly Care and Companion Robots, Shandong Xiehe University, Jinan 250109, China
[3] Shandong Vocational College of Industry, Zibo 256414, China
[4] School of Defense Science and Technology, Xi'an Technological University, Xi'an 710021, China
* Correspondence: wangli65@sdxiehe.edu.cn

Abstract

Beech leaf disease classifiers must transfer from concentrated acquisition to public photographs with different devices, backgrounds and symptom appearance. We propose STAND (Session-aware Teacher And uNlabeled Distillation), which reuses frozen DINOv3 features, constructs within-class session-based orthogonal projections, and combines global and local classification. An initial teacher allocates pseudo-label quotas on an independent unlabeled photo pool; two student rounds retain those quotas and refit the classification heads. On 1856 photographer-disjoint, weakly labeled external photographs, five student fits achieve accuracy of 91.23 ± 0.10%, F1 of 91.85 ± 0.09%, and ROC-AUC of 97.20 ± 0.04% (mean ± sample standard deviation). Relative to confidence-threshold self-training with the same encoder, mean AUC increases by 0.47 percentage points (paired photographer-bootstrap 95% interval: 0.28–0.68), while the accuracy-difference interval includes zero. Matched component experiments show that quota allocation is most useful under imbalanced retrieval-source composition; projection gains depend on pool conditions, and the default projection and normalization choices exhibit condition-specific strengths. Adaptation updates lightweight heads while preserving the encoder. These findings support further evaluation of public-photo screening under weak supervision; the reported metrics quantify agreement with existing weak labels. Standardized expert symptom adjudication and prospective testing with locked models guide field validation.

Keywords: beech leaf disease; public monitoring; domain adaptation; frozen visual representations; multiple-instance learning; pseudo-labels; grouped evaluation

Check for updates

Received: 24 October 2026
Revised: 14 November 2026
Accepted: 17 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Image-based beech leaf disease (BLD) recognition, which aims to distinguish symptomatic from asymptomatic photographs using visible leaf symptoms, is an important image classification task for forest disease monitoring [1]. Cross-source recognition faces acquisition shifts because labeled photographs usually cover limited capture conditions, whereas camera devices, illumination, backgrounds, and symptom appearance may change together in public photographs. In other words, decision rules learned in the source domain may not generalize to new photographers, even when performance is strong on test images from the same acquisition source. Symptom information learned from labeled photographs must therefore transfer to images from new sources. When symptom training labels are unavailable at the new source, unlabeled photographs can be used to adapt the

decision rules. Existing studies primarily build cross-source recognition systems through supervised transfer of pretrained visual models or domain adaptation using source and target data [1–3].

BLD classifiers based on pretrained convolutional networks are representative of supervised transfer. Waldo et al. preprocess and augment leaf images, fine-tune pretrained networks for symptom classification, and produce image-level predictions [1]. Most learning in this pipeline takes place on the labeled source domain. El Karch et al. combine frozen foundation models with lightweight classification heads and fuse global features from multiple backbones to improve field plant disease recognition [4]. However, although these supervised pipelines improve feature extraction and combination, they do not introduce a stage that updates the decision rules using unlabeled public photographs. The subdomain adaptation of Wu et al. and the multi-source adaptation of Gao et al. already use target images for cross-domain learning, primarily through joint feature training or alignment [2,3]. Reusing fixed visual representations still requires adapting source-domain decision rules to public photographs and accounting for changes in the class composition of the unlabeled pool. Such changes affect sample selection during pseudo-label self-training. Assigning equal quotas to both classes may include less reliable predictions from the class with fewer suitable candidates.

Kumar et al. [5] show that, when pretrained features are strong and distribution shifts are substantial, training only a linear head on a frozen feature extractor can outperform full fine-tuning in out-of-distribution recognition. This finding motivates preserving existing visual representations. With a frozen DINOv3 encoder [6], we address source acquisition variation and target-photo classifier adaptation in separate stages. We define a capture session as a consecutive sequence of photographs taken with the same device, with adjacent timestamps no more than 300 s apart. In the source stage, feature differences between capture sessions within each class define projections, and a teacher combines whole-image and patch information. In the target stage, the projections are retained while the classification heads are relearned from independent unlabeled photographs. Source-session information thus informs feature processing, whereas target photographs inform the decision rules, giving the two stages distinct learning roles.

We name the proposed method STAND, Session-aware Teacher And uNlabeled Distillation. STAND uses capture-session information to construct an explicit feature transformation. Within each symptom class, it first computes the deviation of each session mean from the overall class mean, then applies singular value decomposition to the weighted deviation matrix. Separate orthogonal projections for the global and patch branches reduce the influence of the dominant variation directions on classification. During student training, the initial teacher's positive prediction proportion determines the numbers of positive and negative pseudo-labels. These numbers remain fixed over two rounds, while the photographs are reselected using the current model's scores. The initial teacher therefore determines class quotas, and the current model determines sample ranking. The student retains the source projections and refits the global and local classification heads, allowing target photographs to contribute to learning while preserving the constraints derived from source-session statistics.

On the external weak-label set, five student fits attain $91.23 \pm 0.10\%$ accuracy and $97.20 \pm 0.04\%$ AUC, compared with 88.36% and 95.09% for the fixed source teacher. Matched comparisons with confidence-threshold self-training and fixed-feature covariance alignment evaluate adaptation under the same encoder. Further experiments separate teacher and student projection, match the sample budget of equal-quota selection, and compare local normalization rules. The results support target-head adaptation, while showing that the contribution of individual design choices depends on pool composition.

The current evaluation concerns agreement with weak labels on a previously inspected test set, rather than confirmed disease diagnosis.

Field plant disease transfer studies address different sources of distribution shift. MSUN and MDFAN use source–target distribution or classifier alignment for changes in acquisition conditions [2,3]. More recent approaches target style, background, or crop identity through style distillation and logit–prototype matching, background recombination before adaptation, and separation of disease and crop representations [7–9]. Cross-domain few-shot methods additionally use labeled target support sets [10], while TBA combines unlabeled target statistics during training with support-based prototype matching at test time [11]. These supervision settings differ from fitting a classifier on an independent unlabeled photo pool while reserving labeled development data for configuration selection.

Frozen visual representations reduce task-specific training and provide both global and local information [5,6,12]. Linear probing of frozen DINOv2 has been examined for transfer from PlantVillage to field datasets, with crop identity contributing substantially to gains across the label space [13]. PlantCaFo and frozen CLIP–DINO fusion instead combine complementary foundation-model priors with lightweight adaptation or classification heads [4,14]. In the present single-species task, STAND combines whole-image and patch features from one fixed encoder; image-level supervision trains the local aggregation branch, whose response maps are used for qualitative inspection.

Unlabeled adaptation also depends on pseudo-label selection and evaluation. AdaContrast refines pseudo-labels with contrastive and neighborhood information, while FreeMatch adapts global and class-specific confidence thresholds [15,16]. L2C and DINO Teacher illustrate adaptation through additional visual branches or teacher-generated target labels [17,18]; BCA and DART explicitly account for class distributions through prior updates or prediction correction [19,20]. STAND allocates rank-based pseudo-label quotas from the teacher's initial prediction proportion and refits classification heads without updating the shared encoder. This proportion organizes the training sample counts rather than estimating disease prevalence. The evaluation separates source fitting, unlabeled-pool adaptation, labeled development selection, and photographer-grouped testing, while disclosing development-time test exposure in Section 2.8, consistent with concerns about adaptation protocols and data leakage [21,22].

Overall, the contributions of this study are threefold.

1. We present a global–local teacher that uses within-class differences between capture sessions. Deviations of session means from class means estimate the dominant variation directions, and separate projections are constructed for the two representations. This converts capture-session information into feature constraints applied before teacher classification and inherited by the student.
2. We introduce iterative self-training in which the initial teacher determines the two class quotas. Its initial positive prediction proportion allocates positive and negative pseudo-label counts, which remain fixed over two rounds while the selected photographs are updated using current scores. Separating quota allocation from sample ranking addresses the limitations of equal-class selection when source proportions vary.
3. With a shared visual encoder, we evaluate classification-head adaptation against matched alternatives over five fits and report paired uncertainty. Component and pool-composition experiments identify conditional benefits and counterexamples, separating the supported adaptation gain from claims of universal component optimality.

The code and accompanying research materials are available at https://github.com/liProj/STAND-BLD.

2. Materials and Methods

This section brings the algorithm, data definitions and evaluation design together. The following subsections specify the model and training procedure, dataset separation, matched comparators, uncertainty analysis and development-time test exposure.

2.1. Problem Definition and Overview

We study BLD classifier adaptation from concentrated image acquisition to public photographs. The objective is to learn decision rules for photographs from new photographers while keeping the visual encoder fixed and using adaptation images without symptom training labels. STAND constructs a teacher from source class labels and capture-session information, then trains a student on independent public photographs. Source learning establishes orthogonal projections and initial decision rules from session differences. Target adaptation updates the classification heads and their normalization.

Let the labeled source domain be $\mathcal{S} = \{(x_i, y_i, g_i)\}_{i=1}^{N}$, where $y_i \in \{0,1\}$ denotes asymptomatic or symptomatic status, respectively, and g_i denotes the capture session defined by device and timestamp. The unlabeled adaptation pool is $\mathcal{U} = \{u_j\}_{j=1}^{M}$. Let $\mathcal{V}$ be the labeled development set and $\mathcal{T}$ the external test set. Photographers are disjoint across $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{T}$. Labels and session identifiers in $\mathcal{S}$ train the teacher; image features from $\mathcal{U}$ generate pseudo-labels and fit the student; labels in $\mathcal{V}$ select configurations. Student fitting does not use $\mathcal{T}$. Section 2.7 and the subsequent model-selection description detail development and test usage.

The source teacher and target student share the global and patch representations produced by the frozen encoder, as illustrated in Figure 1. Teacher construction begins with deviations of session means from the overall mean within each class. Weighting and singular value decomposition then yield separate projections for the global and local branches. Two heads learn from projected source features and are fused into the teacher. For student adaptation, the initial teacher's positive prediction proportion on $\mathcal{U}$ determines the two pseudo-label counts, while current scores determine which photographs are selected. Both counts and the source projections remain fixed over two rounds; selected photographs and student heads are updated. The feature constraints derived from source sessions are thus retained while target photographs inform the decision rules.

After training, the student heads, projection bases, and normalization statistics remain fixed. Each new photograph independently passes through encoding, normalization, projection, and two-branch scoring. The fused score determines symptomatic or asymptomatic status. Prediction does not revisit the adaptation pool or other test photographs. Unlabeled distillation here takes the form of teacher-guided hard pseudo-label self-training, with supervision supplied by the class labels assigned after sample selection.

2.2. Frozen Encoder and Session-Based Orthogonal Projection

We use a DINOv3 ViT-B/16 encoder [6]. RGB images are resized to 384×384 and normalized according to the pretrained configuration. The encoder produces a $d = 768$ dimensional global feature h_i and $T = 576$ patch features $p_{it} \in \mathbb{R}^d$, excluding prefix tokens. Feature extraction uses inference mode, and all subsequent training operates on cached features.

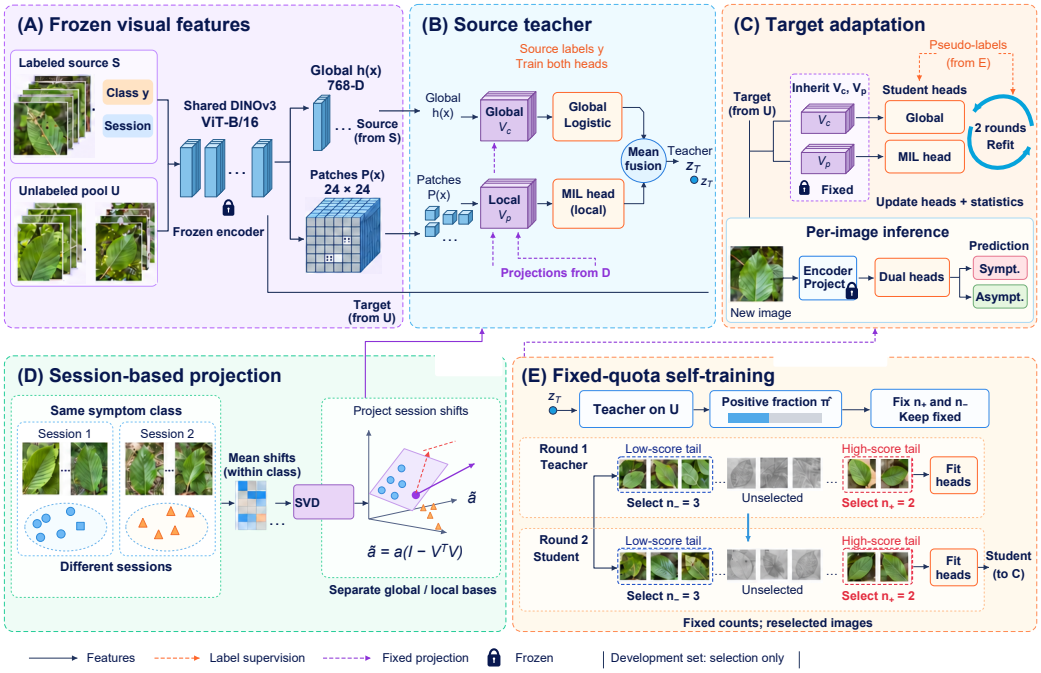


Figure 1. STAND training and inference. (**A**) A shared frozen DINOv3 extracts global and patch features. (**B**) The source teacher fuses global and local predictions after orthogonal projection. (**C**) The student inherits the fixed projections and updates the classification heads and normalization statistics on an independent unlabeled pool, then predicts new images individually. (**D**) Weighted SVD of within-class session differences estimates projection directions separately for the global and local branches; the dashed red component is removed by orthogonal projection. (**E**) The initial teacher determines class quotas, which remain fixed across two rounds; the round-1 student rescores the pool before the second selection. Low-score and high-score tails provide asymptomatic and symptomatic pseudo-labels, respectively. Matching z_T ports indicate transfer of teacher scores. Leaf thumbnails reuse the existing photo examples; feature blocks and the displayed selection counts are illustrative. The development set selects configurations. Sympt./asympt. denote symptomatic/asymptomatic predictions. External test images do not enter student fitting or statistical estimation.

We estimate dominant variation directions from differences between session means within the same symptom class and remove feature components along those directions by orthogonal projection. Estimating differences within each class avoids directly treating the positive–negative class-mean difference as a direction to remove. Global features are standardized with per-dimension source-training means and standard deviations, giving $a_i = (h_i - \mu_S^c)/(\sigma_S^c + \epsilon_c)$ with $\epsilon_c = 10^{-8}$. For class c in session g, define

$$D_{cg} = \sqrt{n_{cg}}\left(\bar{a}_{cg} - \bar{a}_c\right), \qquad n_{cg} \geq 5, \tag{1}$$

where $\bar{a}_{cg}$ and $\bar{a}_c$ are the class means within that session and across all training samples, respectively. Stacking these row vectors forms D. We apply singular value decomposition and retain the first k right singular vectors as the rows of the orthonormal matrix V_c. The projected feature is

$$\tilde{a}_i = a_i(I - V_c^\top V_c). \tag{2}$$

Patch features are standardized using the mean and standard deviation of all source-training patches ($\epsilon_p = 10^{-6}$). The mean patch representation of each image is then substituted into Equation (1) to obtain V_p independently. The two branches do not share a projection basis. The estimated V_c and V_p transform the teacher's global and patch features, respectively, and remain fixed throughout both student rounds.

The projection attenuates dominant directions associated with within-class session differences; it does not guarantee a purely non-disease subspace. Symptom severity may also vary across sessions within a class, so removed directions may contain useful information. We therefore include $k = 0$ as an important control and quantify the projection's limited contribution separately.

2.3. Global–Local Teacher

Whole-image and patch representations provide classification information at different spatial scales. We train a head on each projected representation and fuse them into a teacher that supplies a common score for target-pool selection. The global branch uses L_2-regularized logistic regression,

$$z_c(x) = w_c^\top \widetilde{a}(x) + b_c. \tag{3}$$

with class-balanced weights and regularization parameter $C = 0.1$. The local branch applies a shared linear scorer to each standardized and projected patch, $s_t(x) = w_p^\top \widetilde{p}_t(x) + b_p$, and aggregates the scores by log-mean-exp,

$$z_p(x) = \frac{1}{\tau} \log \left[\frac{1}{T} \sum_{t=1}^{T} \exp\big(\tau s_t(x)\big) \right], \qquad \tau = 1. \tag{4}$$

This aggregation learns local responses from image-level labels without specifying banding locations. The local branch uses weighted binary cross-entropy,

$$\mathcal{L}_p = -\frac{1}{n} \sum_i \big[\alpha y_i \log \sigma(z_p(x_i)) + (1 - y_i) \log(1 - \sigma(z_p(x_i))) \big], \tag{5}$$

where $\alpha = n_- / n_+$ and σ is the sigmoid function. The teacher averages the two logits with equal weights,

$$z_T(x) = \tfrac{1}{2} z_c(x) + \tfrac{1}{2} z_p(x). \tag{6}$$

avoiding an additional fusion network fitted on the limited development data. Local score maps support qualitative inspection and are not treated as validated lesion annotations.

2.4. Teacher-Proportion Selection and Student Training

Student training must determine both how many photographs to select from each class and which photographs to select in each round. STAND handles these decisions separately. Selecting equal numbers from both score tails can admit less reliable samples when one side contains fewer reliable candidates. We first use the initial teacher's positive prediction proportion on $\mathcal{U}$ to allocate the class counts,

$$\widehat{\pi} = \mathrm{clip}\left(\frac{1}{M} \sum_{j=1}^{M} \mathbf{1}[z_T(u_j) > 0],\, 0.05,\, 0.95 \right). \tag{7}$$

where $\widehat{\pi}$ determines pseudo-label counts and is not a calibrated prevalence estimate. With $\rho = 0.6$, the positive and negative counts are

$$n_+ = \max\{10, \lfloor \rho \widehat{\pi} M \rfloor\}, \qquad n_- = \max\{10, \lfloor \rho (1 - \widehat{\pi}) M \rfloor\}. \tag{8}$$

The initial teacher's $\widehat{\pi}$, n_+, and n_- remain fixed over both rounds. Only the ranking and resulting training subset change. In round r, the n_+ highest-scoring photographs receive label 1 and the n_- lowest-scoring photographs receive label 0, forming $\mathcal{Q}_r$. The teacher selects the first round; the first-round student rescores all of $\mathcal{U}$ for the second

round. Class counts are therefore not re-estimated from the student's positive prediction proportion, while the selected photographs can change with the classifier. The smallest pool tested contains 100 photographs, for which the selected tails do not overlap. We make no performance claim for smaller pools.

In each round, the student inherits the source-estimated V_c and V_p and refits both heads under hard pseudo-label supervision from $\mathcal{Q}_r$, without mixing in labeled source samples by default. The orthogonal projections remain fixed, while target photographs determine classification weights and normalization, applying source-derived feature constraints to target-classifier learning. The global head fits $\mathcal{Q}_r$ using means and standard deviations from the full $\mathcal{U}$. The local scorer is first trained using patch statistics from $\mathcal{Q}_r$; at scoring time, these statistics are replaced with those from the full $\mathcal{U}$, without re-optimizing the scorer. Because local training and scoring use different statistics, its performance changes reflect both classification weights and normalization. After two rounds, fusion follows Equation (6) to produce the final student.

For the additional normalization comparison, $Q \to U$ denotes the default local rule: fit patch statistics on the selected subset Q during training and replace them with full-pool U statistics during scoring. The alternatives use $Q \to Q$ or $U \to U$. The global branch is unchanged. Each alternative executes both rounds, including reselection, so these experiments compare complete normalization-dependent training trajectories rather than an isolated final scoring substitution.

2.5. Training Settings and Decision Thresholds

Projection dimensions are selected from $\{0, 2, 4, 8, 12, 16\}$ by internal session-grouped cross-validation, using pooled out-of-fold AUC. Ties after rounding AUC to four decimal places favor the smaller dimension. The teacher used in external experiments has $k = 8$. Logistic regression allows at most 5000 iterations. The patch head is initialized to zero and optimized with AdamW for 400 steps at a learning rate of 3×10^{-3} and weight decay of 10^{-2}. Each step samples 128 images with replacement. The main configuration trains two student rounds; seed, round-count, and quota-coefficient sensitivity are reported separately.

The default decision rule is $z > 0$, equivalent to a sigmoid probability threshold of 0.5. To diagnose ranking–threshold mismatch, we also fit a two-component one-dimensional Gaussian mixture to the scores of the test batch using 200 EM iterations. Among 2001 grid points between the component means, the point with the smallest difference between weighted densities becomes the threshold. This procedure reads no labels but does access the test batch, so it is transductive threshold adjustment. It is used only in explicitly identified diagnostic results and is excluded from the external main metrics. Bimodality and the presence of both classes are not assured, so the rule has no general reliability guarantee.

2.6. Data, Labels, and Separation Units

The labeled source comprises publicly available leaf photographs collected by Waldo et al. in Maryland, USA, in 2025 [1]. The 1934 accessible images contain 787 symptomatic and 1147 asymptomatic photographs. Images are sorted by device and timestamp, and consecutive photographs separated by no more than 300 s form a capture session. This produces 23 sessions over 6 capture days. A 10 s interval instead yields 717 burst sequences. Sessions represent acquisition dependence, not tree or leaf identity. This organization provides an explicit basis for teacher training and evaluation across acquisition conditions.

External data come from openly licensed iNaturalist photographs. A common content filter reduces an initial set of 2999 images to 2383 leaf photographs from 1436 photographers. SigLIP 2 [23] compares prompts describing leaves, trunks, whole trees, fruit, and other content; only images for which the leaf prompt scores highest are retained. The same

filter is applied to both source categories. We retain one photograph per observation and at most one observation per photographer per day, with capture months restricted to May–September. This process retains leaf photographs and reduces repeated observations. Symptom labels follow the separate rules below.

External positives are research-grade observations assigned to the disease taxon. Negatives are research-grade American beech observations selected using a geographic rule, defined at collection, for regions without reported BLD. These regions are Georgia, Alabama, South Carolina, Mississippi, Louisiana, Arkansas, Texas, Florida, and Missouri. Platform identification and geographic priors jointly provide weak labels, and all external metrics are interpreted under this definition. A photographer-based split yields 527 development images and 1856 test images. A separate adaptation pool contains 2976 photographs from 2155 photographers who appear in neither development nor testing. Of these, 1381 come from the disease-taxon source and 1595 from the beech-taxon source. Retrieval source is used in source-proportion experiments, whereas student training uses teacher pseudo-labels. This split tests whether unlabeled adaptation transfers to photographs from new photographers.

Adaptation on an independent target pool must be evaluated on photographs from photographers absent from training. If the student directly fits the test photographs, evaluation also measures adaptation to that particular batch, obscuring transfer to later images. We therefore separate teacher training, student adaptation, configuration selection, and external testing, with dataset sizes in Table 1. Disjoint photographers in the adaptation and test sets require the student to transfer rules learned from one group of public photographs to a new group. Keeping the trained model fixed and predicting images individually makes external evaluation consistent with photographs arriving sequentially in monitoring.

Table 1. Data splits and their roles. External classes follow the platform- and geography-based weak-label rules. The adaptation pool supplies pseudo-label training images. Dashes indicate unavailable training ground truth or photographer identifiers.

Data	Images	Positive	Negative	Photographers	Role
Source $\mathcal{S}$	1934	787	1147	—	Teacher training; session evaluation
Development $\mathcal{V}$	527	294	233	343	Configuration selection
Test $\mathcal{T}$	1856	1019	837	1093	External evaluation
Adaptation $\mathcal{U}$	2976	—	—	2155	Pseudo-label student training

Symptoms in public photographs do not always appear at the same scale. A close-up leaf may fill most of the image, whereas the same cues may occupy only small regions in a multi-leaf scene with different backgrounds and framing. Source and public-photo examples illustrate these differences in Figure 2. Whole-image features preserve leaf structure and context, while local aggregation allows small high-response regions to contribute to the decision. Scale variation motivates combining both forms of information; acquisition-source variation motivates adapting the decision rules after feature extraction.

Figure 2 retains the original source class–device and public-retrieval groups while pairing six existing photographs with enlarged regions. The annotations distinguish visible dark/light variation, vein contrast, openings, and missing marginal tissue without assigning their causes. The source labels are inherited from the original dataset, and the public labels remain weak labels based on platform and geographic collection rules. Consequently, an asymptomatic dataset label does not imply that a leaf is free of all visible damage, and a positive public label is not an expert-confirmed diagnosis. These examples clarify the visual task and its labeling limits; they are not a diagnostic atlas.

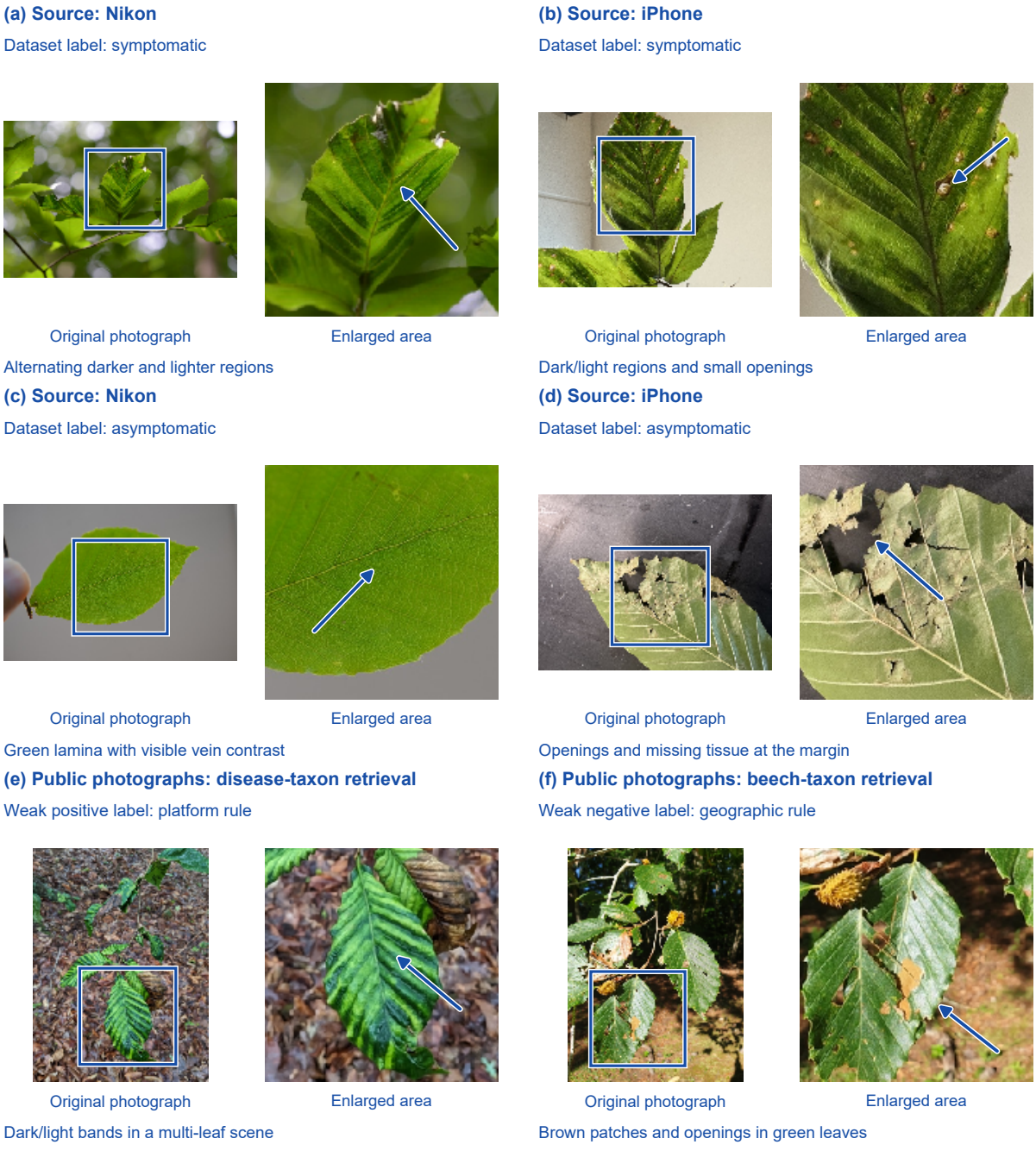


Figure 2. Source-domain and public-photo examples with visible-appearance annotations. Each panel pairs the complete photograph with an enlarged region indicated by a box; arrows highlight visible appearance only. (**a**,**b**) Source photographs carrying the original symptomatic dataset label, acquired with Nikon and iPhone devices, show alternating darker and lighter regions; small openings are also visible in (**b**). (**c**,**d**) Source photographs carrying the original asymptomatic label show a green lamina with visible vein contrast and a leaf with openings and missing marginal tissue, respectively. (**e**) A disease-taxon retrieval example shows dark/light bands in a multi-leaf scene. (**f**) A beech-taxon retrieval example shows brown patches and openings. Public examples retain weak labels assigned by the platform and geographic collection rules, not expert-confirmed disease status. The annotations describe photograph appearance and do not diagnose BLD, insect injury, or another cause. All six photographs were already included in the original example montage; enlargement uses raw image crops without color or contrast alteration. Source identifiers and available license information are supplied in Supplementary Table S26.

Label Verification Status

Platform identification is not equivalent to expert confirmation of visible BLD symptoms, and a geographic negative rule cannot establish absence of infection in every photograph. A stratified random audit sample of 300 legacy-test images (150 per weak-label class) has been prepared for two independent experts. The blinded forms conceal weak labels, retrieval sources and model predictions; the sampling manifest records stratum weights for weighted agreement estimation. Current metrics measure weak-label agreement; the benchmark and analysis program support standardized expert symptom adjudication.

2.7. Comparators and Evaluation Protocols

Comparators include the public EfficientNetV2-S checkpoint [1] and the source-fine-tuned models EfficientNetV2-S [24], Inception-v3 [25], MobileNetV3-Large [26], and ResNet-50 [27]. All four CNNs start from ImageNet weights, with the encoder frozen for the first three epochs. Encoder and head learning rates are then 10^{-4} and 10^{-3}, respectively. The batch size is 16, with at most 100 epochs. Training stops after 5 epochs without validation-loss improvement, restoring the weights with the lowest validation loss. Augmentation includes random cropping, flipping, affine transformations, grayscale conversion, and color perturbation. External comparators retain the model names used in their original papers; table notes specify fine-tuning. The public checkpoint cites Waldo et al.'s BLD application paper. Internal configurations are frozen global classification (L1), global classification with orthogonal projection (L2), and the STAND teacher (L3), representing three stages of teacher construction. L1–L3 are experiment identifiers defined in this study, all using DINOv3 [6]. Component analysis compares them with full STAND.

Recent representation baselines use DINOv2-B/14 [12] and SigLIP 2-B/16 [23], representing self-supervised visual and vision–language pretraining. Both freeze the encoder and fit class-balanced logistic regression on cached whole-image features, without orthogonal projection, a local branch, or target-pool adaptation. Input sizes are 392 and 224, respectively. Per-dimension standardization statistics use only source fitting samples. The regularization parameter C is selected from $\{0.001, 0.01, 0.1, 1, 10\}$ using pooled out-of-fold AUC from five-fold source session-grouped validation; ties at four decimal places favor the smaller C. DINOv2 and SigLIP 2 select 10 and 1, respectively. Heads are then fitted on all 1934 source images and evaluated on the same external test set at a zero-score threshold. Both baselines follow the predefined candidate ranges without external labels for parameter selection. These are source-only linear-probing experiments added in this study. The DINOv2 configuration in Supplementary Table S6 uses orthogonal projection and different head settings to evaluate projected representations.

Internal evaluation uses five-fold stratified group cross-validation with entire capture sessions held out. Each fold serves once as the test fold. CNNs use the next fold for validation and the remaining three for parameter updates. Frozen heads fit all non-test folds and select projection dimensions through nested session-grouped validation within those folds. These different supervised fitting budgets are reported explicitly. We also perform leave-one-group-out evaluation on 7 groups defined by capture day $\times$ device. Session and day–device grouping test changes in local consecutive acquisition and larger acquisition units, respectively, providing complementary evidence for the global–local teacher across capture conditions.

All external evaluations use the same public-photo set in Table 1. The main baseline comparison shares frozen DINOv3 representations and the adaptation pool across STAND, confidence-threshold self-training and fixed-feature CORAL. L3 is the unadapted same-encoder reference; the public EfficientNetV2-S checkpoint is an application reference with a different architecture. The original broader complete-system comparison is retained in

Supplementary Table S24, with its distinct source-fitting budgets and three-seed CNN means identified. Those cross-backbone results describe system performance and cannot isolate the proposed adaptation method.

2.7.1. Matched Adaptation Baselines

Confidence-threshold pseudo-labeling retains STAND's source teacher, rank-8 teacher/student projection, global and local heads, optimization settings, equal-weight logit fusion, and two adaptation rounds. At each round, the fused logit is converted once to $p = \text{sigmoid}(z)$; images with $p \geq \tau$ receive a positive pseudo-label and those with $p \leq 1 - \tau$ receive a negative pseudo-label. The second round uses scores from the first student. The selected-image count is allowed to vary with confidence, while the global branch keeps U statistics for training and scoring and the local branch uses Q statistics for training and U statistics for scoring. A class count below ten would retain the previous model; this fallback was not triggered in the reported runs. Validation AUC selects $\tau = 0.9$ from $\{0.7, 0.8, 0.9\}$ at seed 2026 before the five formal fits at seeds 0–4.

Fixed-feature CORAL [28] separately aligns source global CLS features and source patch tokens to the unlabeled target pool U. Each branch first expresses source and target features in the same coordinate system standardized by the target-pool mean and standard deviation. Let m_S, m_U and C_S, C_U denote the means and covariance matrices in this common coordinate system, and define $\widetilde{C}_D = C_D + \lambda \ \text{tr}(C_D) I/d$ for $D \in \{S, U\}$. Source row vectors are transformed as

$$x'_S = (x_S - m_S)\widetilde{C}_S^{-1/2}\widetilde{C}_U^{1/2} + m_U.$$

The ridge term regularizes the covariance inverse square root; the patch covariances use all source and pool tokens rather than image-level averages. Source labels train a balanced logistic global head ($C = 0.1$, maximum 5000 iterations) and the same local log-sum-exp head (temperature 1, 400 AdamW steps, learning rate 0.003, weight decay 0.01, batch size 128). Target features remain in the common U coordinate system for scoring, and the two logits are fused with weights 0.5/0.5. This comparator uses neither nuisance projection nor pseudo-label adaptation. Validation AUC selects $\lambda = 0.1$ from $\{0.001, 0.01, 0.1\}$ at seed 2026; formal results use seeds 0–4. The grids and observed validation scores are reported in Table S20.

2.7.2. Repeated Fits and Statistical Comparisons

Metrics are accuracy, precision, recall, F1 and ROC-AUC, with weak-positive photographs as the positive class. AUC is multiplied by 100 in tables; differences in percentage metrics are reported in percentage points. Five additional fits use training seeds 0–4, with mean and sample standard deviation computed from model-wise metrics. These seeds vary student or local-head training; source teachers remain fixed at seed 0. They do not represent five independent teacher retrainings. The original seed-0 STAND predictions are exactly reproduced.

Full-pool method differences use 2000 paired photographer-cluster bootstrap resamples. Each draw resamples photographers with replacement, retains all their images, and applies the identical draw to both methods. Percentile intervals condition on the fitted models. Matched-subpool differences instead use five paired runs and approximate Student-t intervals with four degrees of freedom; each pair shares the same sampled pool and training seed, while different pairs vary both. All intervals are nominal 95% intervals without multiplicity correction. Original fixed-model diagnostic figures retain their 1000-draw descriptive intervals; the original pool grid retains its two-run ranges. These uncertainty summaries address different sources of variation and are identified separately.

2.8. Model Selection and Test Exposure

A labeled development set selects pseudo-label rules, round counts, projection settings, source-sample mixing and normalization. Here, "unlabeled" refers to the adaptation pool used for student fitting. External test performance for 15 earlier configurations was inspected during development, and pool-composition analysis informed quota design. The additional confidence-threshold and CORAL parameters are selected by development-set AUC; their trial runner also computes and saves legacy-test outputs. The 35 subsequent projection experiments were added after inspecting the main comparisons. Consequently, neither the original nor the additional results on this set constitute an untouched confirmatory test. Photographer separation prevents overlap in acquisition identity but does not remove development-time test exposure.

Locked Independent-Test Protocol

A prospective protocol reserves newly collected, nonoverlapping photographs for one evaluation after model and comparison choices are fixed. Existing image identifiers and photographer records provide an exclusion index; model files and the primary comparison of STAND against the development-selected confidence-threshold baseline are specified before new evaluation. The primary endpoint is the five-seed mean paired AUC difference between STAND and the strongest additional baseline selected on development set V, with uncertainty estimated by paired photographer-level resampling. The package retains an exclusion index of 7293 records and model states with hashes for the primary comparison. The locked protocol specifies independence checks, blinded prediction and single-pass evaluation of new seasonal photographs under prespecified rules.

3. Results

Using full STAND as the reference, we examine overall performance on the common external test set, component progression and student ablations, target-pool conditions, score behavior, and individual cases. Component experiments retain the same test photographs and specify the change relative to the full method. Per-fold teacher results and representation and aggregation choices appear in Supplementary Section S2 to explain teacher initialization.

3.1. Cross-Source Recognition of Public Photographs

Table 2 compares STAND with two adaptation alternatives on the same frozen DINOv3 features, the unadapted L3 teacher, and the public-model reference. The five-fit STAND means are 91.23 ± 0.10% accuracy, 91.85 ± 0.09% F1 and 97.20 ± 0.04% AUC. Confidence-threshold self-training reaches 90.88 ± 0.13% accuracy and 96.73 ± 0.10% AUC; fixed-feature CORAL reaches 81.22 ± 0.26% and 89.59 ± 0.21%. The shared encoder controls representation choice, while each comparator retains its stated adaptation rule.

The paired STAND-minus-threshold AUC difference is 0.47 percentage points (95% interval: 0.28–0.68); the accuracy difference is 0.34 points (interval: −0.36 to 1.07). Thus, the observed ranking advantage is better supported than an accuracy advantage at the fixed threshold. Against L3, mean AUC increases by 2.11 points (1.58–2.71), and accuracy by 2.87 points (1.74–3.99). Against fixed-feature CORAL, the AUC difference is 7.61 points (6.45–8.84). Complete paired comparisons are given in Supplementary Table S14. These results support head adaptation under the evaluated protocol, without establishing superiority to every form of domain adaptation.

The original seed-0 models are retained for the diagnostic ROC curves in Figure 3; their values are not five-fit means. Student AUC increases from the fixed teacher's 95.09 to 97.25 for that fit. A threshold change on unchanged teacher scores cannot produce this

AUC difference, supporting a change in ranking through target-head fitting and statistical processing. The original complete-system comparison, including DINOv2 and SigLIP 2 source-only probes, remains available in Supplementary Table S24.

Table 2. Additional adaptation comparisons on the legacy external test set.

Method	Fits	Accuracy (%)	F1 (%)	AUC (%)
STAND (full)	5	91.23 ± 0.10	91.85 ± 0.09	97.20 ± 0.04
Confidence-threshold pseudo-labeling	5	90.88 ± 0.13	91.45 ± 0.12	96.73 ± 0.10
Fixed-feature CORAL	5	81.22 ± 0.26	82.64 ± 0.29	89.59 ± 0.21
Global–local source teacher (L3)	Fixed	88.36	88.82	95.09
Public EfficientNetV2-S	Fixed	81.30	83.50	89.64

Accuracy, F1 and AUC are reported on a 0–100 scale. For fitted adaptation methods, values are means ± sample standard deviations over training seeds 0–4; source teachers, when used, are fixed at seed 0. The legacy external test set (T_{old}) contains 1856 images from 1093 photographers and uses the original weak labels. Analyses on this development-exposed set are exploratory. The first four methods share the same frozen DINOv3 features. The public EfficientNetV2-S model is an external architecture reference, not a matched-backbone comparator. The selected probability threshold is 0.9 and the CORAL covariance ridge coefficient is 0.1.

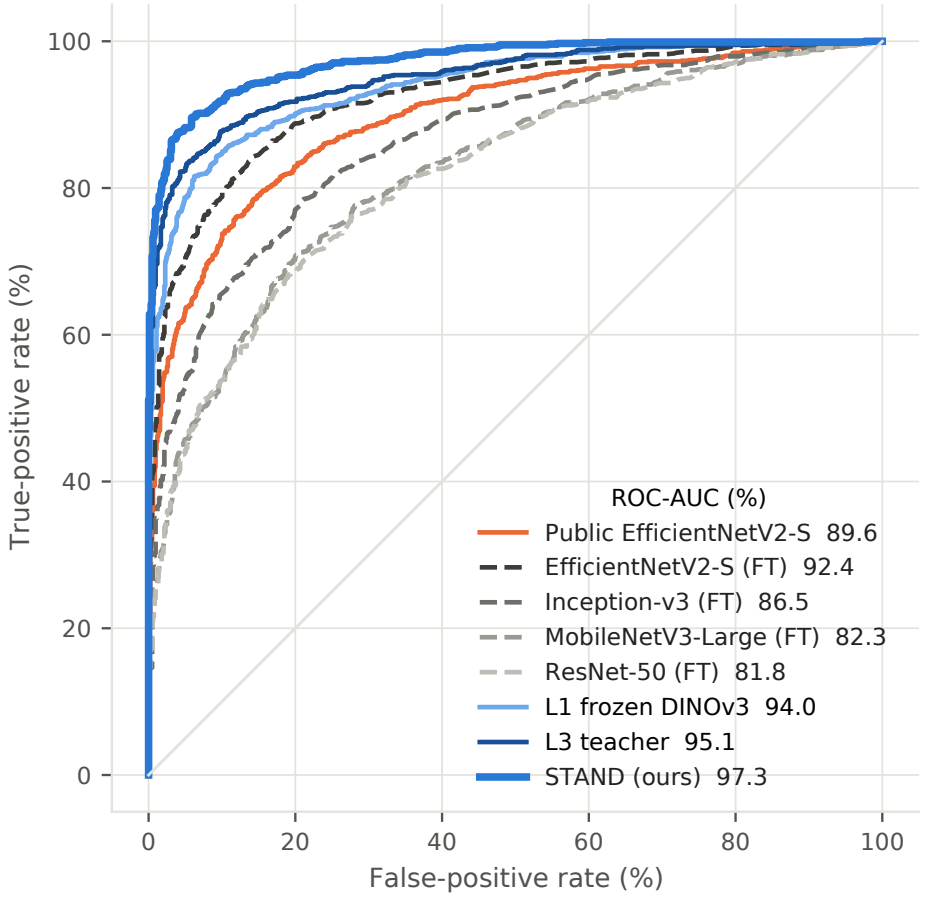


Figure 3. Original fixed-model ROC curves on the external weak-label set. STAND uses seed 0 (AUC 97.25), not the five-fit mean in Table 2. CNN curves also use seed 0; FT denotes source fine-tuning. L1 and L3 are the fixed configurations in Table 3.

The fixed-model intervals in Figure 4 describe sampling variation among the observed photographers using the original 1000 bootstrap draws. For the original STAND fit, the accuracy interval is 90.07%–92.78% and the AUC interval is 96.59–97.87. These descriptive intervals differ from the paired five-fit comparisons above and do not include development-time selection bias.

Table 3. Component progression from internal configurations to full STAND (%). All share DINOv3 and the external test set. L1 uses a global head; L2 adds orthogonal projection; L3 adds a local branch to form the teacher; full STAND adapts the student on the target pool. L1–L3 are configurations defined here, not independently published methods. Column-best metrics are bold.

Configuration	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
Frozen global head (L1)	80.60	97.41	66.44	79.00	94.02
Global head + projection (L2)	83.41	**97.98**	71.25	82.50	94.42
STAND teacher (L3)	88.36	93.98	84.20	88.82	95.09
STAND (full method)	**91.38**	93.87	**90.19**	**91.99**	**97.25**

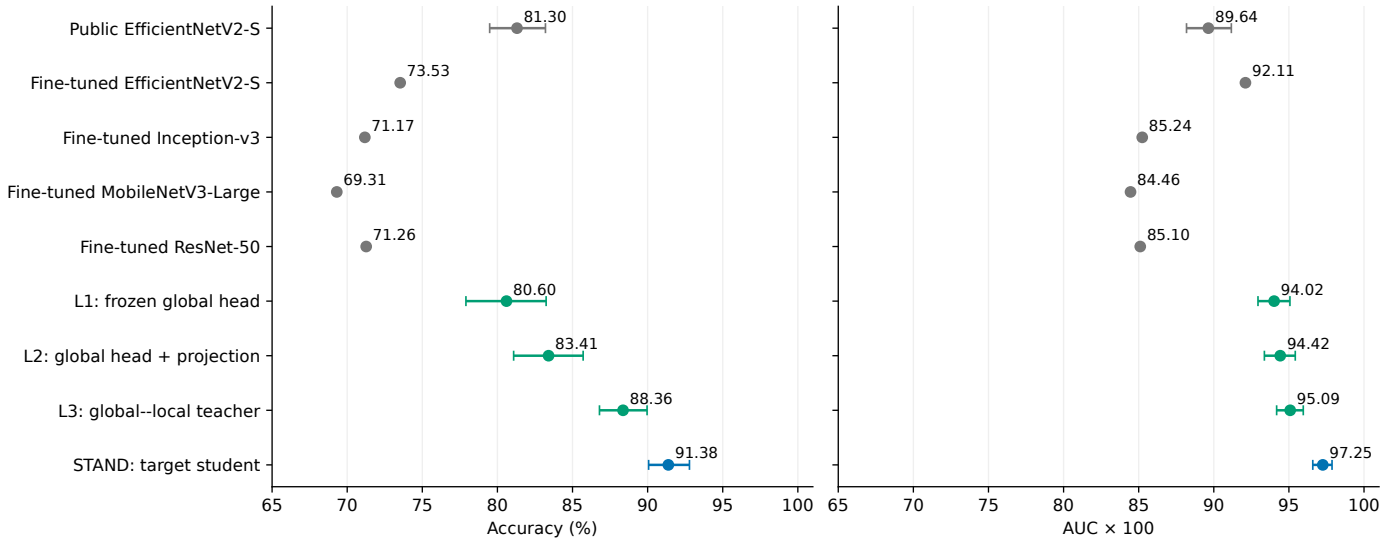


Figure 4. Original fixed-model performance diagnostics. Points show individual models, including the seed-0 STAND fit; error bars are 95% intervals from 1000 photographer-cluster bootstrap draws. CNN bars are three-seed means without intervals. These intervals differ from the paired comparisons in Supplementary Table S14.

3.2. Component Progression from Source Classifier to STAND

The original fixed-model progression in Table 3 retains L1 (global classification), L2 (added projection), L3 (added local branch), and the seed-0 STAND student. It documents the initial system progression; the STAND values are single-fit results, whereas Table 2 reports five-fit means. The new factorial projection comparison below evaluates component dependence more directly.

A whole-image representation compresses the image into one vector, whereas the local branch retains regional contributions to the classification score. Adding a local head can move some photographs across the decision threshold without substantially reordering all samples. From L2 to L3, recall rises from 71.25% to 84.20%, and the accuracy improvement exceeds the AUC change (Table 3). Local information therefore contributes mainly by increasing positive detection. Student adaptation then improves both ranking and default-threshold accuracy, supporting a teacher that combines whole-image and regional information before learning target decision rules from public photographs.

3.3. Ablation Analysis Relative to Full STAND

The repeated component analysis compares four teacher/student projection settings with ranks 0 or 8, equal quotas with exactly the same selected-sample budget, and three local normalization rules (Table 4). All full-pool configurations use five fits; the source

teacher for each setting remains fixed. The original single-fit ablations are retained in Supplementary Table S25. In those diagnostics, removing the local branch lowers AUC from 97.25 to 96.02 and accuracy from 91.38% to 87.98%, supporting regional information in that particular fit.

Table 4. Projection, quota and local-normalization comparisons with the full unlabeled pool.

Configuration	k_{teacher}	k_{student}	Local Stats.	Accuracy (%)	AUC (%)
STAND (full)	8	8	$Q \to U$	91.23 ± 0.10	97.20 ± 0.04
Teacher projection only	8	0	$Q \to U$	91.05 ± 0.16	97.07 ± 0.07
Student projection only	0	8	$Q \to U$	91.41 ± 0.03	97.29 ± 0.05
No projection	0	0	$Q \to U$	91.34 ± 0.25	97.11 ± 0.08
Equal-class quotas	8	8	$Q \to U$	90.65 ± 0.17	97.13 ± 0.04
$Q \to Q$ normalization	8	8	$Q \to Q$	91.23 ± 0.08	97.21 ± 0.05
$U \to U$ normalization	8	8	$U \to U$	90.60 ± 0.22	96.98 ± 0.09

All rows use the same 2976-image pool and two self-training rounds (five student seeds, fixed source teacher seed 0). $k = 0$ disables projection; $k = 8$ retains the original projection rank. Teacher-proportional quotas are used except in the equal-quota row, which matches the actual total number selected in each corresponding STAND run. Q denotes the selected pseudo-labeled subset and U the unlabeled pool. The global branch always uses U statistics. Each normalization variant reruns both rounds, so the comparison includes changes in second-round selection. Single-sided projection runs were added after the main experiment.

3.3.1. Projection Depends on the Configuration

With the complete pool, default projection in both stages gives an AUC of $97.20 \pm 0.04\%$, compared with $97.11 \pm 0.08\%$ without either projection. The paired AUC difference is 0.084 points (-0.044 to 0.208), and the accuracy difference is -0.108 points (-0.608 to 0.443). Student-only projection attains $97.29 \pm 0.05\%$ AUC, exceeding the default by 0.090 points (0.010–0.176). The four configurations thus achieve similarly high performance on the full pool. Matched-subpool analysis finds an AUC gain of 0.679 points (0.349–1.008) for 200 images and 0.912 points (0.642–1.182) for the 90% disease-source pool, whereas the 800-image interval crosses zero (Supplementary Tables S16 and S17). These comparisons characterize empirical projection gains across pool conditions; nuisance removal and projection–quota interactions are separate mechanistic questions.

3.3.2. Normalization Changes Are Small but Measurable

The default $Q \to U$ and the consistent $Q \to Q$ rules both average 91.23% accuracy. Relative to the default, $Q \to Q$ changes 0.259% of class predictions on average and has slightly higher AUC (97.21 versus 97.20) and lower Brier score (0.06603 versus 0.06631). The $U \to U$ alternative reaches 90.60% accuracy, 96.98% AUC and a Brier score of 0.07007, with a 1.056% mean prediction-flip rate. These complete two-round comparisons show output sensitivity to normalization, including any induced reselection changes. They neither establish equivalence nor isolate a calibration mechanism (Supplementary Table S18).

3.4. Adaptation to Photo-Pool Size and Composition

The original pool-size grid uses two runs per size (Table 5 and Figure 5). With 100 photographs, its mean AUC is 96.84; this descriptive result suggests that small pools can support adaptation in this dataset, without establishing stable performance across new pools. We therefore add five matched pairs for 200 and 800 images at natural source composition, and for 1000-image pools with 10%, 50% and 90% disease-source recruitment. Pool seeds 100–104 are paired with training seeds 0–4. Within each pair, both methods receive the same photographs and training seed; across pairs, both sampling and training vary.

Table 5. Sensitivity of STAND and fixed-quota selection to unlabeled-pool size and composition (%). Ranges span two runs; the full pool uses the single main run. Composition is defined by disease-taxon retrieval source. The highest mean in each performance column within each experimental group is bold.

Pool Condition	Runs	STAND AUC	Range	STAND Accuracy	Fixed-Quota AUC
Pool size					
100 images	2	96.84	96.18–97.51	90.17	—
200 images	2	96.34	96.34–96.35	88.50	—
400 images	2	96.97	96.54–97.40	90.38	—
800 images	2	97.08	96.63–97.53	90.76	—
1600 images	2	97.04	97.00–97.08	90.73	—
2976 images (full pool)	1	**97.25**	—	**91.38**	**97.07**
Composition: 1000 images per pool; disease-taxon source fraction					
5%	2	95.37	95.25–95.50	87.90	87.35
10%	2	95.82	95.43–96.21	88.44	89.22
20%	2	96.95	96.74–97.15	90.03	93.52
35%	2	96.96	96.72–97.20	90.41	96.53
50%	2	**97.35**	96.92–97.78	**91.46**	**97.23**
65%	2	97.05	96.82–97.29	90.52	96.33
80%	2	95.81	95.59–96.03	88.28	94.56
90%	2	93.58	92.81–94.35	85.61	93.86
95%	2	92.15	91.71–92.58	84.05	93.14

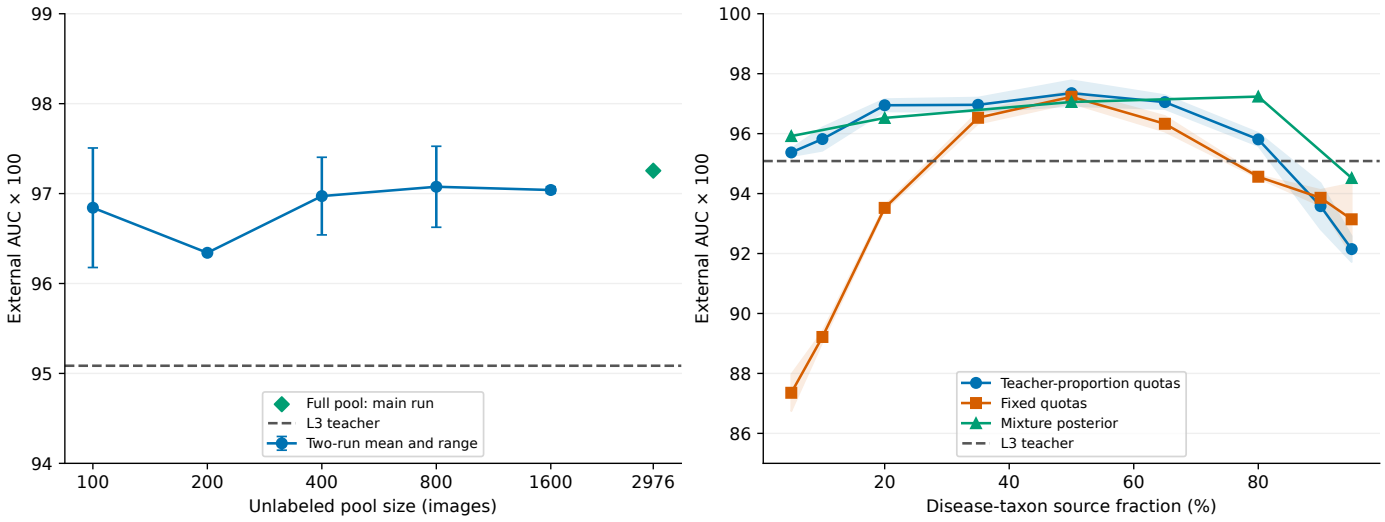


Figure 5. Sensitivity to unlabeled-pool size and composition. Teacher-proportion selection is STAND; fixed quotas and mixture-posterior selection are variants. Left-panel error bars and right-panel bands span the minimum and maximum of two runs. The full-pool and mixture-posterior controls are single runs. Dashed lines indicate teacher AUC. The panels use different vertical ranges. Right-panel proportions refer to retrieval sources.

Equal-quota selection now matches the default method's actual number of selected samples after integer rounding and divides it as evenly as possible between classes. On the full pool, teacher-proportion selection improves mean accuracy by 0.582 points (−0.042 to 1.276) and AUC by 0.073 points (0.003–0.143). In the five-pair subpool analysis, its largest observed gains occur at 10% disease-source recruitment: accuracy improves by 11.42 points (10.20–12.65) and AUC by 6.00 points (5.07–6.94). At 90%, the accuracy gain is 7.97 points (5.93–10.02), but the AUC interval includes zero. Several more balanced or

small-pool intervals also include zero. Thus, the quota advantage is condition-dependent (Supplementary Table S15). Retrieval-source proportions describe the sampling design, not expert-verified disease prevalence.

The original two-run grid also shows declining AUC as recruitment approaches a single source: the default reaches 93.58 at 90% and 92.15 at 95% disease-source recruitment. These historical grid values and the new five-pair estimates use different sampled pools and should not be pooled as repeated measurements of one experiment. Both analyses indicate that quota adjustment cannot compensate for all limitations in source coverage; they do not establish the underlying biological class composition.

3.5. Ranking, Decision Thresholds, and Probability Calibration

The score, threshold and calibration diagnostics in this subsection use the original fixed teacher and seed-0 student. They provide context for the repeated comparisons and are not five-fit estimates. Good ranking must translate into appropriate class decisions for image-by-image screening. A fixed deployment threshold particularly requires adapted scores to match the target distribution. We therefore inspect positive and negative scores relative to both the default and mixture-derived thresholds. STAND's default accuracy is close to, and slightly higher than, its accuracy after unlabeled threshold adjustment (Figure 6). Target-pool training has therefore produced a directly usable decision threshold in this setting. Together with the AUC gain, this indicates that student adaptation improves both ranking and fixed-threshold classification, allowing later photographs to be processed individually.

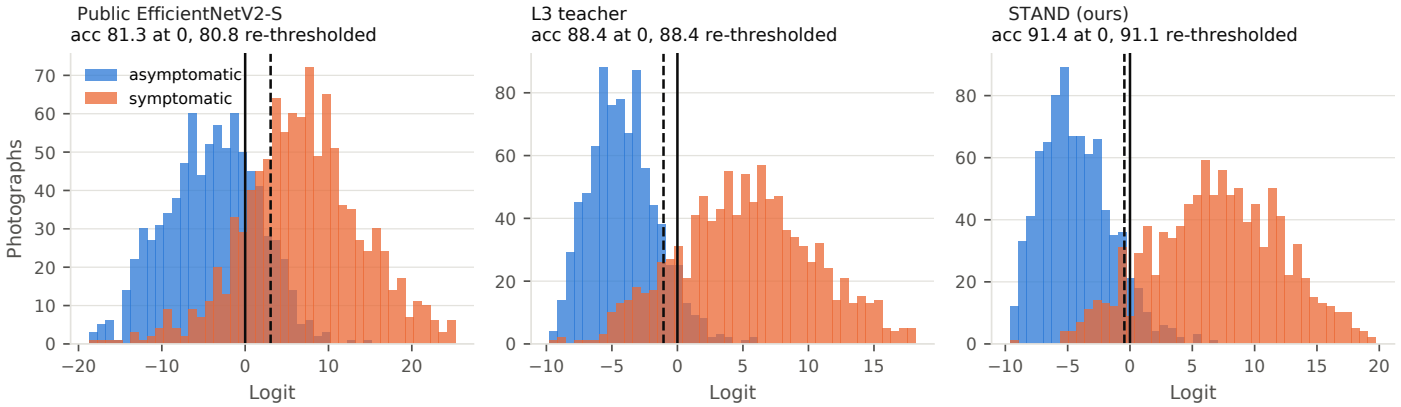


Figure 6. Original fixed-model diagnostic (seed-0 STAND). External positive- and negative-class score histograms for STAND and source classifiers. Solid lines mark the default logit threshold of 0; dashed lines mark the unlabeled mixture threshold. Class names follow the external weak-label definition. Each vertical axis counts photographs.

To assess whether the student advantage depends on the default threshold, we apply the same unlabeled threshold-adjustment rule to teacher and student. The student outperforms the teacher with both default and adjusted thresholds (Table 6). Correcting the teacher's decision threshold alone therefore does not close the gap. Together with the ROC change in Figure 3, this supports a discrimination gain from refitting the heads beyond threshold correction, validating target-photo learning of global and local decision rules.

Agreement between predicted probabilities and target-label frequencies is another relevant property. AUC evaluates only relative ranking, whereas reliability curves compare the observed positive fraction at similar predicted probabilities. Without fitting an additional calibrator, we inspect the teacher and student outputs directly. Empirical calibration error decreases from approximately 5.6% to 3.5% (Figure 7), alongside accuracy and AUC improvements. Target-pool learning therefore strengthens discrimination while making the score distribution more consistent with current weak-label frequencies. This calibration

change accompanies both target-sample fitting and statistical processing, and its probability interpretation remains specific to the current labels.

Table 6. External sampling intervals and threshold comparisons (%). Intervals are photographer-cluster bootstrap 95% intervals for single models. The four fine-tuned CNNs report three-seed metric means. Column-best point estimates are bold. Adj. denotes unlabeled threshold adjustment.

Method	Accuracy	Accuracy Interval	AUC	AUC Interval	Adjusted Accuracy
Public EfficientNetV2-S [1,24]	81.30	79.48–83.20	89.64	88.18–91.17	80.77
Fine-tuned EfficientNetV2-S [24]	73.53	—	92.11	—	80.44
Fine-tuned Inception-v3 [25]	71.17	—	85.24	—	73.76
Fine-tuned MobileNetV3-Large [26]	69.31	—	84.46	—	63.81
Fine-tuned ResNet-50 [27]	71.26	—	85.10	—	72.97
L1 frozen global head [6]	80.60	77.90–83.24	94.02	92.94–95.07	86.75
L2 global head + projection	83.41	81.08–85.71	94.42	93.37–95.42	87.18
L3 global–local teacher	88.36	86.79–89.96	95.09	94.18–95.96	88.42
STAND (ours)	**91.38**	90.07–92.78	**97.25**	96.59–97.87	**91.06**

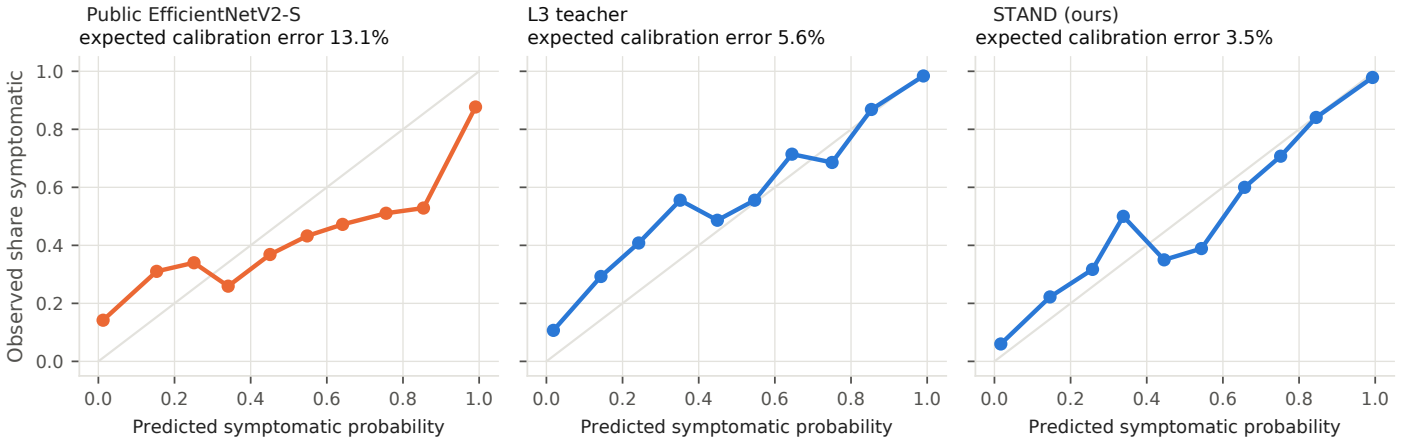


Figure 7. Original fixed-model diagnostic (seed-0 STAND). External reliability diagrams for STAND and source classifiers. Probabilities are divided into 10 equal-width bins; lines show bins with at least 15 samples. ECE weights all nonempty bins by their sample proportions. The diagonal represents ideal empirical calibration.

3.6. Stratified Performance, Local Responses, and Failure Cases

Aggregate results can depend on the numbers of photographs from different years and sources, so we stratify external performance by capture year and location fields. Yearly subsets test whether gains concentrate in one period; geographic subsets examine positive detection and negative exclusion separately. STAND outperforms the public model in each year from 2021–2026 (Figure 8). This consistency supports improvement across several acquisition conditions rather than a gain driven by one year. Because positive and negative classes follow different source rules, geographic results are descriptive and interpreted with their sample sizes, refining the scope of the overall adaptation evidence.

Changes in predictions for individual photographs help explain the contribution of local information. Figure 9 identifies the data label and both models' decisions, then juxtaposes the original photograph, the public model's Grad-CAM [29], and STAND patch scores. Group A contains public-model false negatives corrected by STAND; group B contains positives correctly detected by both; group C contains public-model false positives corrected by STAND. Corrections occur in both close-up and multi-leaf scenes and cover positive detection as well as negative exclusion. Together with the local-branch ablation in Supplementary Table S25, these examples provide qualitative support for complementary

global and local information. Blue and orange in STAND maps indicate local scores favoring asymptomatic and symptomatic classes, respectively. Bright regions in the public-model maps indicate stronger relative responses for the symptomatic class within that image. Because the maps have different definitions, regional responses should be read alongside final predictions; colors do not directly identify lesion locations.

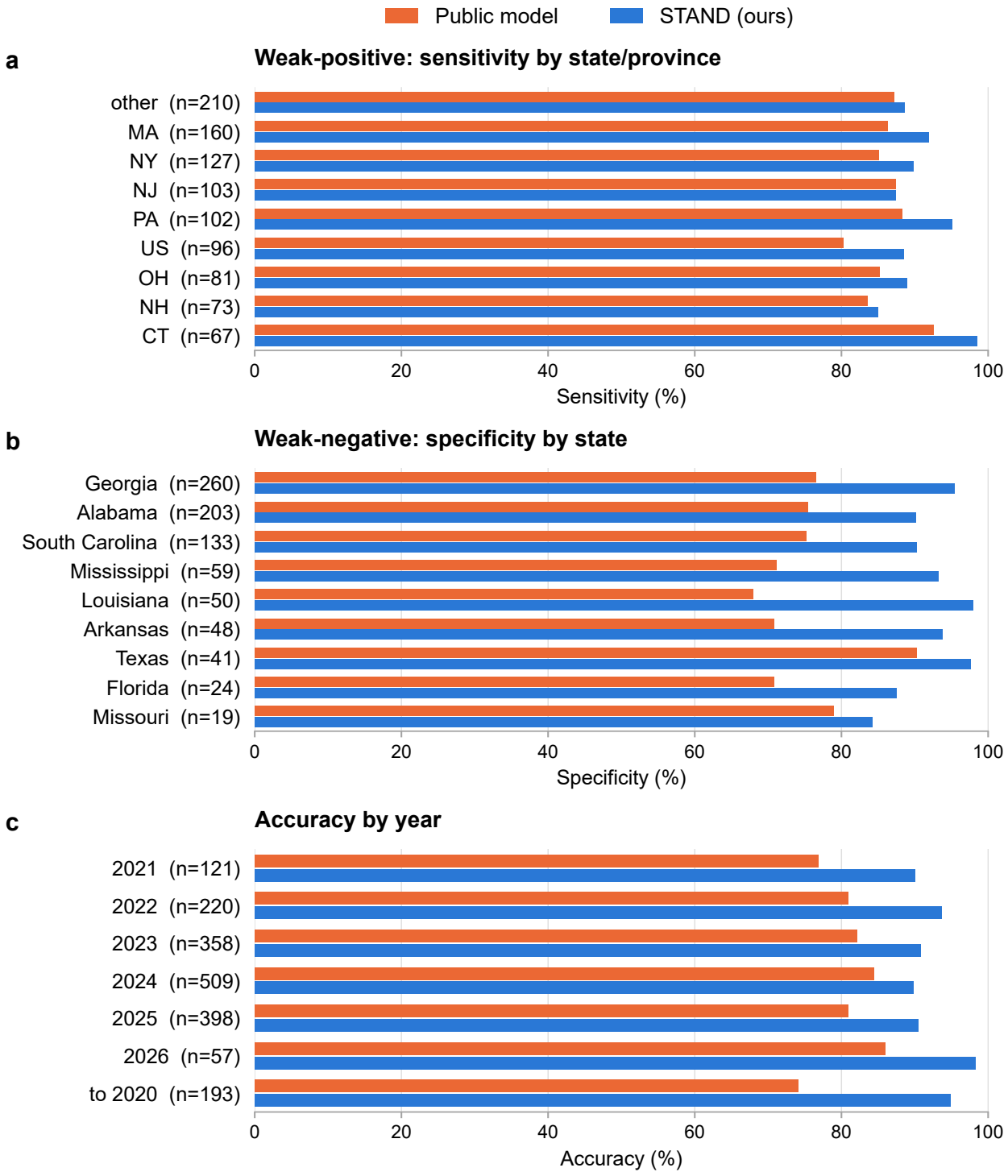


Figure 8. External performance of STAND and the public model by location field and year. (**a**) Sensitivity within the weak-positive subset by state/province. (**b**) Specificity within the weak-negative subset by state. (**c**) Accuracy by year across the full external test set. Sample counts (*n*) appear beside each group label. US and other retain their group meanings from the original location strings. All groups show descriptive point estimates for the original fixed models.

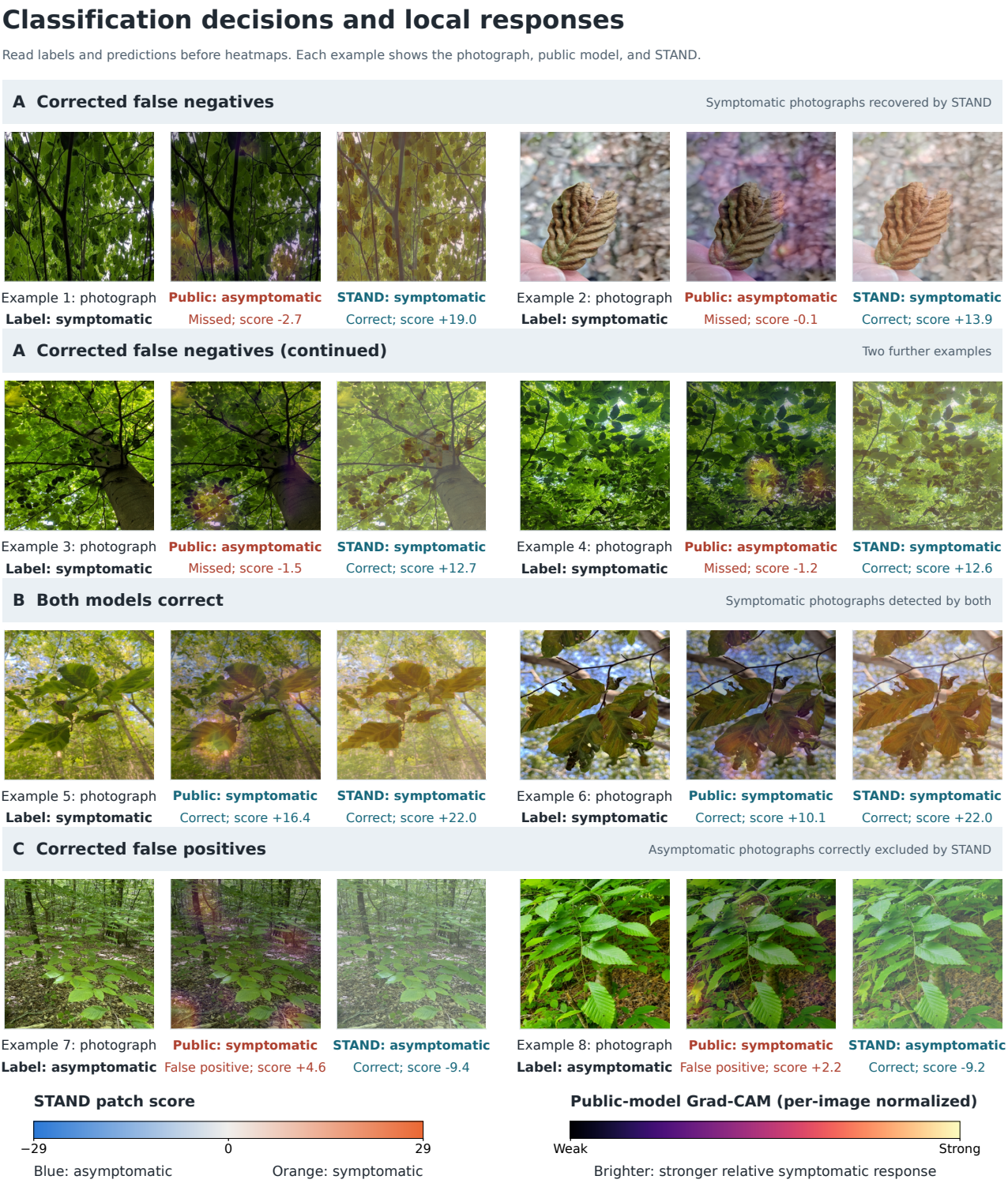


Figure 9. Classification decisions and local responses of STAND and the public model. Each example shows the original photograph, public EfficientNetV2-S Grad-CAM, and STAND map from left to right. Groups A, B, and C show corrected false negatives, positives correct for both, and corrected false positives. Labels, predictions, and outcomes are annotated. Blue and orange in STAND maps support asymptomatic and symptomatic classes; bright public-model regions indicate stronger relative symptomatic responses. Correctness follows weak labels, and response maps are not lesion annotations. Photo sources and licenses appear in Supplementary Table S13.

The original seed-0 model makes 160 errors against the weak labels (100 false negatives and 60 false positives), all retained with branch scores. Inspection of eight high-confidence errors selected by absolute fused score (four false negatives and four false positives) provides a more specific view of the failure modes. The false-negative examples include small leaves in a backlit canopy, cluttered vegetation and ground backgrounds, and overlapping green leaves with limited visible detail. A close-up false negative shows that large foreground leaves do not by themselves prevent failure. Both branches produced negative scores in all four displayed false negatives, and their maximum patch scores were also below zero; thus, these examples do not support an explanation in which a correctly detected positive patch was merely diluted by global fusion. False positives include yellow-brown curled leaves, irregular yellow-green patches, and strong light/dark variation in multi-leaf scenes. For observation 289951406, the local score was 9.03 compared with a global score of 2.09, illustrating how a strong local response can contribute to an incorrect fused prediction. These are observations about classifier behavior under the existing weak labels, not lesion localization or diagnoses of the causes of discoloration. The eight score-selected cases do not estimate the prevalence of individual error mechanisms. The eight photographs and their branch scores are provided in Supplementary Figure S12; Figure 10 retains the original illustrated failures.

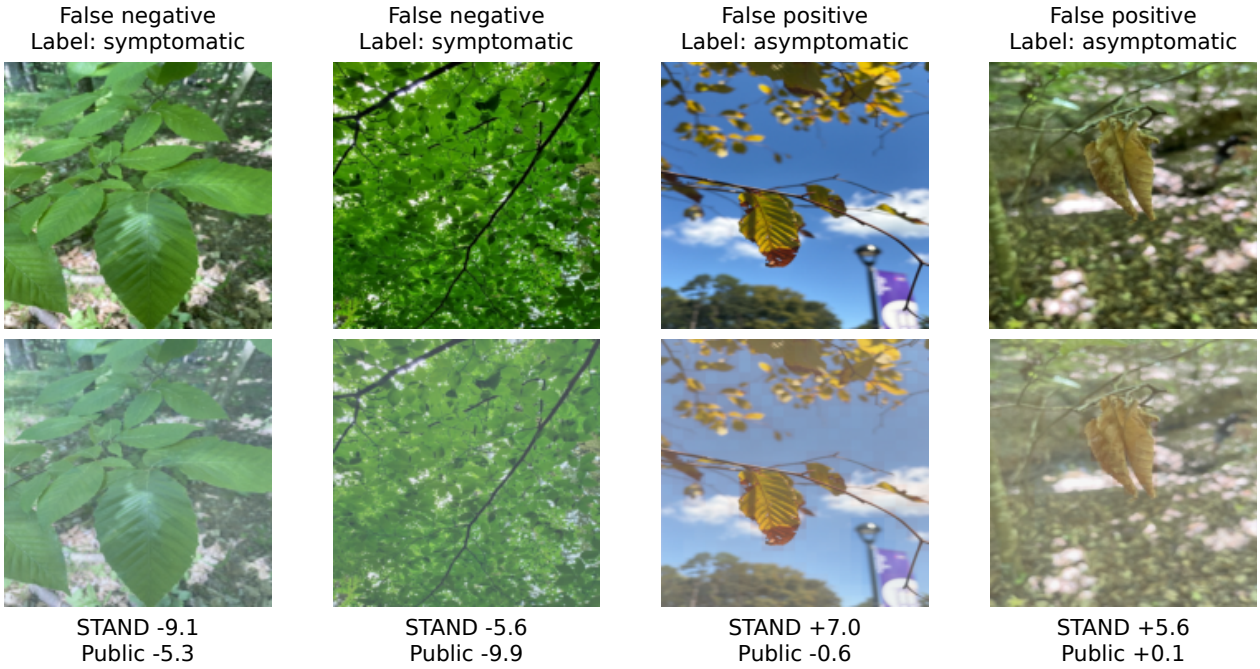


Figure 10. STAND false negatives and false positives. The first row shows photographs and the second shows student patch scores. Labels follow the current weak-label definition; numbers are student and public-model logits. Heatmaps show local classification responses.

3.7. Computational Measurements

On the measured GB10 system, the frozen encoder contains 85,641,216 parameters and each global–local head set contains 1538 trainable parameters. Over 100 single-image measurements, feature extraction has median latency 24.05 ms (95th percentile: 32.73 ms). End-to-end prediction, including JPEG loading, preprocessing, encoding and heads, has median latency 57.02 ms (95th percentile: 128.99 ms), following ten warm-up calls. Three repetitions using cached features give source-teacher preparation of 14.43 ± 0.72 s and two-round adaptation of 34.81 ± 1.55 s; adaptation timing includes evaluation and output

writing. Peak allocated GPU memory is 10,299 MiB and 14,380 MiB, respectively, including resident source, adaptation and evaluation feature caches. These values describe the measured shared-device implementation, not minimum deployment memory or verified field throughput. Complete timing boundaries are reported in Supplementary Table S19.

4. Discussion

4.1. Complementarity of Source Representations and Target Decisions

The results support complementary roles for global–local discrimination and target-classifier adaptation. Local fusion mainly improves positive detection, and the global-only student ablation shows that regional information remains useful after adaptation. The student further improves AUC and accuracy, demonstrating that target-photo learning produces better ranking and decisions from the same frozen representation. The AUC change cannot result from moving the teacher threshold, supporting relearning of the target classifier. Together, these findings support retaining the encoder, combining information at different spatial scales, and updating decision rules with an independent target pool.

STAND's distinctive implementation combines source-session projection with teacher-proportion sample counts and repeated ranking on an independent target pool. The additional experiments give different levels of support to these choices. Head adaptation improves AUC over the fixed teacher and confidence-threshold baseline; budget-matched quota selection is particularly useful in the 10% disease-source condition. Projection is beneficial in several subpools, while the full-pool configurations achieve high AUC values of 97.07%–97.29%. Both normalization rules trained with Q, namely $Q \rightarrow U$ and $Q \rightarrow Q$, average 91.23% accuracy and yield closely agreeing predictions. These comparisons connect the framework's design choices to their observed gains under specified adaptation conditions.

4.2. Data Organization for Public Monitoring

Acquisition diagnostics and learning curves show that image count and capture coverage affect teacher learning differently. Time and device proxies reveal label associations in consecutive photographs, while broader session coverage improves cross-source discrimination. Expanding labeled data should therefore increase variation in devices, photographers, locations, and symptom appearance, rather than merely adding photographs under the same conditions. Source labels establish symptom discrimination, and the target pool supplies the application-source feature distribution. Both support STAND's cross-source adaptation.

The original fixed-model progression in Table 3 shows that target-pool learning increases positive detection while retaining reliable positive predictions, supporting the use of existing labeled knowledge for public-photo screening. Local responses allow reviewers to inspect regions contributing to decisions, while difficult cases show why responses must be interpreted with the original image and label provenance. The system provides image-level classification and regional responses for review; deployment thresholds can be chosen according to the relative costs of missed detections and false positives.

4.3. Validation Scope and Limitations

Three considerations define the scope of field validation. First, external classes follow platform identification and geographic sampling rules; the reported metrics therefore quantify agreement with weak labels. The 300-image benchmark supplies anonymized materials and weighted analysis for expert assessment of visible foliar symptoms. Photograph-level symptom assessment and specimen-based pathogen confirmation address distinct evidence levels. Second, the evaluation history in Section 2.8 motivates prospective testing with locked models, exclusion checks and single-pass assessment of new seasonal photographs.

Repeated fits and photographer-level bootstrap characterize variability within the current dataset. Third, source composition, normalization and projection effects depend on adaptation conditions, and computational measurements describe the tested hardware and implementation. The component comparisons guide screening. Field validation combines expert-adjudicated symptom labels, new photographers and fixed decisions to assess transfer across regions and acquisition conditions.

5. Conclusions

We propose STAND for classifier adaptation on frozen visual representations, combining global–local decisions, source-session projections and quota-based pseudo-label self-training. Five student fits achieve $91.23 \pm 0.10\%$ accuracy and $97.20 \pm 0.04\%$ AUC on the existing photographer-disjoint weak-label test set. Matched comparisons support an AUC gain over confidence-threshold self-training and show substantial quota benefits under selected imbalanced source conditions. Projection and normalization gains depend on their configurations. These results support public-photo screening with weak labels, while standardized expert symptom assessment and a locked prospective protocol guide validation across new photographers and growing seasons.

Data and Research Materials

Source images and the public BLD model originate from Waldo et al. [1]. External photographs come from openly licensed iNaturalist observations. The accompanying project retains image manifests, photographer splits, source-session splits, training configurations, per-image predictions, and summary tables for checking the reported results. Rights to public photographs remain with their original contributors; observation identifiers and license information should be retained when the photographs are used.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/plants15233736/s1. The accompanying Supplementary Material contains Sections S1–S7, Figures S1–S12, and Tables S1–S26.

Author Contributions: Conceptualization, W.L. and L.W.; methodology, W.L., L.W. and W.H.; software, W.L. and H.Y.; validation, D.H., J.L. and H.Y.; formal analysis, W.L., D.H. and W.H.; investigation, W.L., D.H. and J.L.; data curation, D.H., J.L. and H.Y.; writing—original draft preparation, W.L.; writing—review and editing, L.W., W.H. and J.L.; visualization, W.L. and H.Y.; supervision, L.W. and W.H.; project administration, L.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province, grant number ZR2024MF102 (Li Wang).

Data Availability Statement: The code, fixed data manifests and splits, training configurations, exported classifier heads, per-image reference predictions, and summary results are available at https://github.com/liProj/STAND-BLD. Original source images and the public BLD model are provided by Waldo et al. [1]. External photographs are identified by iNaturalist observation records in the repository; their original contributors retain the rights and licenses.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Waldo, B.D.; Vieira, P.; Borden, M.A.; Li, S. Beech leaf disease symptom detection using deep learning and computer vision tools. *J. Nematol.* **2026**, *58*, 66–77. https://doi.org/10.2478/jofnem-2026-0011.
2. Wu, X.; Fan, X.; Luo, P.; Choudhury, S.D.; Tjahjadi, T.; Hu, C. From Laboratory to Field: Unsupervised Domain Adaptation for Plant Disease Recognition in the Wild. *Plant Phenomics* **2023**, *5*, 0038. https://doi.org/10.34133/plantphenomics.0038.
3. Gao, X.; Feng, Q.; Wang, S.; Zhang, J.; Yang, S. A multi-source domain feature adaptation network for potato disease recognition in field environment. *Front. Plant Sci.* **2024**, *15*, 1471085. https://doi.org/10.3389/fpls.2024.1471085.

4. El Karch, H.; Natij, Y.; Benaly, M.; El Gouri, R.; Mezouari, A. Backbone diversity beats text supervision: a systematic study of frozen multi-foundation model fusion for in-the-wild plant disease recognition. *Front. Plant Sci.* **2026**, *17*, 1860665. https://doi.org/10.3389/fpls.2026.1860665.
5. Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; Liang, P. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
6. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.E.; Ramamonjisoa, M.; et al. DINOv3. *Trans. Mach. Learn. Res.* **2026**.
7. Zhan, K.; Peng, Y.; Liao, M.; Wang, Y. Domain generalization plant leaf disease recognition: Toward from laboratory to field. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111168. https://doi.org/10.1016/j.engappai.2025.111168.
8. Jeon, W.; Kim, T.; Choi, S.; Yang, K.; Kim, S.Y.; Song, M. Bridging the Lab-to-Field gap in plant disease diagnosis through unsupervised domain adaptation enhanced by background recomposition. *Ecol. Inform.* **2026**, *93*, 103579. https://doi.org/10.1016/j.ecoinf.2025.103579.
9. Wu, Z.; Guo, J.; Hou, W.; Zhou, K.; Cao, K.; Jung, H. Unseen-Crop Plant Disease Classification via Disentangled Representation Learning. *Electronics* **2026**, *15*, 1553. https://doi.org/10.3390/electronics15081553.
10. Yang, S.; Feng, Q.; Zhang, J.; Yang, W.; Zhou, W.; Yan, W. From laboratory to field: cross-domain few-shot learning for crop disease identification in the field. *Front. Plant Sci.* **2024**, *15*, 1434222. https://doi.org/10.3389/fpls.2024.1434222.
11. Zhao, C.; Xu, T.; Zhang, Z.; Geng, X. Lightweight Cross-Domain Few-Shot Plant Disease Recognition Through Target-Domain Statistical Calibration. *Sensors* **2026**, *26*, 3632. https://doi.org/10.3390/s26123632.
12. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193.
13. Nguyen, T.A.; Nguyen, D.S.; Dang, Q.M.; Phan, B.N.L.; Nguyen, L.H. From laboratory to field: Frozen foundation-model features toward robust plant disease recognition. *Smart Agric. Technol.* **2026**, *15*, 102531. https://doi.org/10.1016/j.atech.2026.102531.
14. Jiang, X.; Wang, J.; Xie, K.; Cui, C.; Du, A.; Shi, X.; Yang, W.; Zhai, R. PlantCaFo: An efficient few-shot plant disease recognition method based on foundation models. *Plant Phenomics* **2025**, *7*, 100024. https://doi.org/10.1016/j.plaphe.2025.100024.
15. Chen, D.; Wang, D.; Darrell, T.; Ebrahimi, S. Contrastive Test-Time Adaptation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 295–305. https://doi.org/10.1109/CVPR52688.2022.00039.
16. Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
17. Chi, Z.; Gu, L.; Liu, H.; Wang, Z.; Wu, Y.; Wang, Y.; Plataniotis, K. Learning to Adapt Frozen CLIP for Few-Shot Test-Time Domain Adaptation. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R., Eds.; 2025; Volume 2025, pp. 66359–66380.
18. Lavoie, M.A.; Mahmoud, A.; Waslander, S.L. Large Self-Supervised Models Bridge the Gap in Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 4692–4702.
19. Zhou, L.; Ye, M.; Li, S.; Li, N.; Zhu, X.; Deng, L.; Liu, H.; Lei, Z. Bayesian Test-Time Adaptation for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 29999–30009.
20. Jang, M.; Chung, H. Label Distribution Shift-Aware Prediction Refinement for Test-Time Adaptation. *arXiv* **2025**, arXiv:2411.15204.
21. Zhao, H.; Liu, Y.; Alahi, A.; Lin, T. On Pitfalls of Test-Time Adaptation. In *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 42058–42080.
22. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
23. Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv* **2025**, arXiv:2502.14786.
24. Tan, M.; Le, Q. EfficientNetV2: Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 10096–10106.
25. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 2818–2826.
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1314–1324.

27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
28. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. *Proc. AAAI Conf. Artif. Intell.* **2016**, *30*, 2058–2065. https://doi.org/10.1609/aaai.v30i1.10306.
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 618–626.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.