Machine Vision and Applications (2026) 37:148
https://doi.org/10.1007/s00138-026-01910-0

RESEARCH

Abnormality detection and quantification in maize fields using selective unsupervised domain adaptation

Aminul Huq[1] · **Dimitris Zermas**[2] · **George Bebis**[1]

Received: 28 January 2026 / Revised: 20 July 2026 / Accepted: 17 August 2026

Abstract

Timely detection and quantification of plant abnormalities in agricultural fields is critical, as delays can significantly impact crop health and yield. Automating this process can offer substantial economic and environmental benefits. The focus of this work is on detecting and quantifying abnormalities in maize fields by analyzing high-resolution RGB images captured by Unmanned Aerial Vehicles (UAVs). Common approaches rely on limited training data that are also tuned to specific field characteristics. As a result, models trained on one field do not perform well on data from another field due to the *domain shift*. To address this challenge, we present a deep learning framework for robust and accurate detection and quantification of abnormalities in maize fields, independently of growth stage, abnormality severity, or soil type. Central to the proposed approach is the idea of standardizing data collected under diverse field conditions by mapping them to a *shared domain* using unsupervised domain adaptation (UDA). UDA preserves critical abnormality-related information while filtering out irrelevant, field-specific variations. To further enhance data quality, we filter out poorly adapted images that introduce artifacts using a methodology we refer to as *selective* UDA. We leverage the shared domain to enhance training by performing data augmentation and improve domain generalization by bridging the domain shift between training and test images. During training, data collected under various field conditions are effectively aggregated by mapping them into a shared domain, enriching the training data with domain-adapted images rather than synthetic data alone. During testing, incoming images are mapped into the same shared domain. Our framework employs Vision Transformers for detection and quantification, along with an ensemble of CycleGANs for domain adaptation. We demonstrate the effectiveness of the proposed approach on a publicly available dataset containing UAV-collected RGB images of healthy and unhealthy maize plants of various growth stages from two different maize fields exhibiting variable conditions.

Keywords Maize plants · Abnormality detection and quantification · Data augmentation · Domain adaptation

1 Introduction

Plant abnormalities can arise from various factors, including nutrient deficiencies, drought, plant diseases, and pests. According to the Food and Agriculture Organization of the United Nations, plant diseases and pests account for a 40% loss in global crop yields annually, leading to an estimated $220 billion in economic damages [1]. Early detection of plant abnormalities enables timely and targeted interventions, preventing widespread crop damage while preserving yield quality and productivity. In addition, it reduces the financial burden on farmers by minimizing costly treatments and crop losses, ultimately promoting sustainable agriculture and enhancing global food security. Abnormality detection in agriculture has traditionally relied on manual labor, requiring farmers to inspect individual plant rows and leaves. This process is not only labor-intensive but also impractical for large-scale operations, as it must be repeated multiple times throughout the plants' life-cycle. Additionally, it is highly error-prone, as accurately identifying various abnormalities demands significant expertise and

✉ George Bebis
bebis@unr.edu

Aminul Huq
ahuq@unr.edu

Dimitris Zermas
zermas@sentera.com

[1] Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV 89512, USA

[2] Sentera, Saint Paul, MN 55114, USA

Published online: 02 September 2026

experience. These challenges can result in the improper or excessive application of fertilizers and pesticides, contributing to over-treatment and environmental problems.

In this study, we focus on detecting and quantifying abnormalities in maize plants, one of the most critical staple crops in the US, with an estimated 15.2 billion bushels harvested in 2023 [2]. UAVs and robotic platforms, combined with Deep Learning (DL) methods, provide a promising solution for efficient data collection and analysis in various crops, including maize [3–5]. Equipped with advanced sensors, UAVs can capture high-resolution imagery across multiple spectra, such as visible light, infrared, near-infrared, and hyper-spectral. These images can then be processed using DL algorithms to detect abnormalities at early stages, enhancing targeted treatments and optimizing crop management. While UAVs have significantly improved large-scale, high-resolution data collection, there exist several challenges in successfully deploying DL approaches in agriculture. First, DL models require extensive training data, which is not always readily available. Transfer learning with fine-tuning [6] and data augmentation [7] strategies have been used to address this issue. Second, DL models trained on data from a particular field under certain conditions often experience performance degradation when applied to the same or different fields due to variations in soil types, crop growth stages, lighting conditions, and camera type. This phenomenon, known as the *domain shift* problem, stems from differences in data distribution across field environments and remains a major obstacle to the widespread adoption of DL-based abnormality detection in agriculture.

The standard approach to mitigating domain shift is *supervised domain adaptation* (SDA), which involves retraining DL models to achieve satisfactory performance in a new domain. However, this method relies on annotated data from the target domain, which is often unavailable and requires significant time and effort to obtain. Consequently, a purely supervised approach is not scalable, as it necessitates continuous labeling efforts each time domain conditions change. An alternative and more scalable solution is *unsupervised domain adaptation* (UDA) [8], which eliminates the need for annotated data in the target domain. The key idea behind UDA is to develop algorithms that transfer knowledge learned from a labeled dataset to an unlabeled dataset. Throughout this study, we define the *source domain* as the training domain and the *target domain* as the testing domain. UDA has been applied in agriculture in two main ways: Synthetic-to-Real Unsupervised Domain Adaptation (SR-UDA) and Real-to-Real Unsupervised Domain Adaptation (RR-UDA). SR-UDA addresses the *domain gap* problem when synthetic data is used for data augmentation. In this case, UDA is used to adapt synthetic plant images to match real field conditions while leveraging labels from the source domain [9, 10]. RR-UDA addresses the *domain shift* problem by transferring images from the target domain to the source domain using labels from the source domain [11–15]. We review representative approaches from each category in Sect. 2.

The focus of this work is on developing an effective framework for detecting and quantifying abnormalities in maize fields using UAV-captured high-resolution RGB imagery. Our goal is to create a generic abnormality detection and quantification system that remains robust across different growth stages, soil types, lighting conditions, and other field variations. More broadly, this work aims to demonstrate how domain adaptation can be used both as a data augmentation strategy during training and as a generalization mechanism during deployment. Therefore, the proposed framework is not restricted to a specific domain adaptation model or downstream prediction model. In this study, we employ CycleGANs [16] to transform data from different field conditions into a shared domain and use a Vision Transformer (ViT) regression model [17] for abnormality quantification. However, other domain adaptation models or task specific downstream model could also be incorporated within the same framework. Our objective is not to claim that this particular combination is optimal, but rather to validate the feasibility and effectiveness of the overall domain adaptation-driven framework for data augmentation and generalization which can ensure robust abnormality detection and quantification. We use pixel-level UDA method CycleGAN in contrast to feature-level UDA methods because it adapts the images on pixel space which in turn provides better interpretability. Also pixel-level UDA is inherently model-agnostic, providing greater flexibility to integrate the adapted images with a wide range of downstream task models without requiring modifications to their feature extraction pipelines.

By mapping data from different field conditions into a shared domain, the proposed approach enables a more consistent representation across diverse environments. Unlike traditional data augmentation methods, this strategy does not rely solely on synthetic data, although synthetic data can be easily incorporated into the framework. While synthetic data can provide readily available ground truth information, generating realistic synthetic representations for different crops, growth stages, and field conditions remains challenging. In contrast, mapping real data into a common domain preserves a higher degree of realism, although it requires annotating at least a subset of the data [18]. During training, domain-adapted images from different field conditions can be aggregated within the shared domain to enrich the training set and reduce the impact of field-specific variations. During deployment, images from a new domain can first be

mapped to the same shared domain before being analyzed by the downstream abnormality quantification model.

To prevent contamination of the training and test data with low-quality translated samples, we apply a filtering strategy referred to as *selective* UDA, which removes poorly adapted images. This step is important because domain adaptation may introduce artifacts, alter visual characteristics related to plant abnormalities, or remove abnormality-related information in some cases. By filtering out such samples, the proposed framework aims to retain the benefits of domain adaptation while reducing the negative impact of low-quality translations. For abnormality quantification, we employ a ViT regression model that estimates the level of abnormality in image patches. By applying a threshold to the output of the ViT regressor, the same model can also be used for abnormality detection. Our method can provide critical insights to farmers by identifying areas of the field that require closer attention. Additionally, it can facilitate data annotation by directing annotators' focus toward specific regions instead of the entire field, thereby improving both the accuracy and efficiency of image labeling [18]. An earlier version of this work has appeared in [19].

The key contributions of this work can be summarized as follows:

- We propose a domain adaptation-driven framework that uses domain adaptation model for data augmentation and generalization for abnormality detection and quantification in maize fields using UAV-captured high-resolution RGB imagery. We demonstrate the feasibility of the proposed framework using CycleGAN-based domain adaptation and a ViT-based regression model for patch-level and whole image-level abnormality detection and quantification.
- We introduce a selective UDA strategy to identify and remove poorly adapted images that may contain artifacts or lose abnormality-related visual information during domain translation.

The remainder of the paper is structured as follows: Sect. 2 provides a review of related work. Section 3 outlines the proposed methodology and key contributions while Sect. 4 describes the datasets used in our experimental analysis. Section 5 presents our experiments, results, and comparative evaluations. Finally, Sect. 6 summarizes our findings and discusses directions for future research.

2 Background

Detecting nutrient deficiencies and other abnormalities, such as those caused by drought, diseases, and pest infestations, remains a significant challenge in the agriculture industry. Traditional methods are invasive, requiring manual leaf, foliar, or soil sample collection at regular intervals, followed by chemical analysis. Although these methods can detect abnormalities early, often before visual symptoms appear, they are expensive and can take days or even weeks to process. Noninvasive methods generally rely on non-imaging spectroscopy, such as chlorophyll meters and canopy reflectance sensors [20, 21], as well as imaging spectroscopy, including RGB, multispectral, and hyperspectral cameras [22, 23]. While chlorophyll meters and canopy reflectance sensors measure variations in leaf spectral properties, additional processing is needed to correlate these changes with specific abnormalities accurately. In [24], a handheld meter was used to measure chlorophyll concentration in rice plant leaves precisely. A linear regression model was then developed to correlate color features with variations in chlorophyll concentration resulting from nitrogen deficiency. In imaging spectroscopy, image analysis techniques are commonly used to identify visual indicators of plant abnormalities [25]. Features such as color, texture, and shape can be extracted from RGB images of leaf samples and used to train traditional classifiers, including Multi-Layer Perceptrons (MLPs) and Support Vector Machines (SVMs), for abnormality detection. These approaches have been applied to a variety of crops, including chili [26], coffee [27], corn [27], and tomato [28]. Alternatively, DL methods can automatically extract more powerful, though less interpretable, features from RGB images, often surpassing traditional approaches in performance [5, 28–30]. Overall, the effectiveness and robustness of these methods can be further enhanced by incorporating a richer set of features derived from multispectral or hyperspectral imagery, although this typically comes with increased cost and complexity [31, 32]. In this work, however, we focus our review on approaches that utilize high-resolution RGB cameras for data collection, as they offer a cost-effective solution while still capturing a substantial amount of visual information.

Estimating the severity of abnormalities for an entire field using only a small sample of leaves is both time-consuming and unreliable. On the other hand, collecting data automatically using UAVs (or ground robots) can provide a representative sample of images from the entire field efficiently and cost-effectively [33]. More broadly, UAV and remote sensing imagery have been widely used in various recognition tasks, including scene classification, object detection, and land-cover analysis. Recent studies have increasingly explored CNN–Transformer hybrids, ViT-based models,

and knowledge distillation to improve recognition accuracy, robustness, and computational efficiency. For example, quantitative regularization has been used to improve ViT-based remote sensing image classification under domain discrepancies and limited training data [34]. Other studies have investigated cross-modal and symmetry-aware knowledge distillation, where knowledge from ViT or ensemble teacher models is transferred to lightweight CNN students to improve classification accuracy while reducing model size and inference cost [35, 36]. In addition, causal feature alignment has been introduced for remote sensing object detection to improve robustness against environmental and sensor-induced variations [37].

In the domain of agricultural abnormality identification a methodology was proposed to detect nitrogen deficiency in maize plants and evaluate its severity using UAV-collected high-resolution RGB images by Zermas et al. [18]. In their study, the authors first tackled the challenge of generating a large dataset of high-quality annotated images for training DL-based object detection models. For specialized applications, such as nitrogen deficiency detection, domain experts are typically required to annotate many images, a time-consuming and costly process. To overcome this challenge, the authors developed a recommendation tool to identify candidate plants exhibiting nitrogen deficiency. With minimal interaction, this tool enabled annotators to efficiently generate a large training dataset. The recommendation system was developed using three different SVMs. The first SVM extracted green pixels by distinguishing them from other categories, such as yellow and soil-colored pixels. The second SVM further separated the yellow pixels from the remaining ones, while the final SVM determined whether the yellow pixels resulted from nitrogen deficiency. The generated dataset was then used to train a Faster R-CNN model based on the ResNet-50 architecture for nitrogen deficiency detection. Their approach achieved a mean average precision of approximately 82.3% at a 50% Intersection over Union (IoU). UAV-captured RGB images of maize fields were analyzed to detect drought-affected plants in [30]. The authors employed U-Net and U-Net++ architectures with a ResNet34 backbone to segment drought-affected areas. Rather than manually annotating the images to establish the ground truth, they generated segmentation masks by combining Otsu's method with a naive thresholding approach. The U-Net and U-Net++ models were then trained for drought-affected maize plant segmentation. By integrating the results of both models, they achieved an average Intersection over Union (IoU) of 0.71. In [38], CNN models were employed to analyze UAV-captured images for weed segmentation. Initially, a CNN model was trained on small field regions. Subsequently, pixel-wise weed maps were generated by applying the trained CNN to larger, whole-field orthomosaic images.

Deploying DL models in agricultural applications requires large amounts of training data, which are often not readily available. Data annotation is also frequently necessary, a time-consuming and labor-intensive process. Various methods have been proposed to address these challenges, including data augmentation [7] and transfer learning through pre-training followed by fine-tuning [6]. Standard data augmentation techniques (e.g., rotation, shifting, zooming, flipping, and brightness and color transformations) and transfer learning strategies (e.g., pre-trained models on large datasets or ensembles of pre-trained models) with fine-tuning have been commonly used in agricultural applications. Such approaches have been applied to a variety of problems such as nutrient deficiency detection [39, 40], disease detection [41, 42], and semantic segmentation [43]. Traditional data augmentation techniques, however, fail to offer the data variability required in agricultural applications. In [29], five different CNNs with standard data augmentation were evaluated for detecting nutrient deficiencies in sugar beets using high-resolution RGB images (top-down view of the entire plant). The results revealed that variations in growth stages and field conditions, especially differences in soil types, significantly impacted the performance of the CNN models. Transfer learning leverages the knowledge of a previously trained model to enhance performance on downstream tasks. Fine-tuning, which involves refining the model's weights using data from the application domain, typically leads to further improvements. However, when pre-trained models are applied to domains that differ significantly from those they were trained on, performance remains suboptimal, even with fine-tuning [44]. We have encountered this issue in our experiments, which can be mitigated by ensuring that the domains of the pre-trained model and the downstream task are similar.

A promising data augmentation strategy that can also help to address the shortage of annotated data, is synthetic data generation [45–47]. This approach offers a cost-effective way to create large datasets for training DL models. Additionally, annotations are trivially available from the renderer, eliminating manual annotation and providing highly accurate annotations. To enhance the realism of synthetic images, SR-UDA [8] can be used to align synthetic data with real field domains better, improving generalization performance. The main idea of domain adaptation is to align the source and target distributions. This can be performed either in the feature space (i.e., extract features that remain invariant in the source and target domains) or in the pixel space (i.e., map images from the source domain to the target domain). Here, we focus on pixel-based methods as feature-based methods depend on the downstream task.

Fei et al. [9] introduced a SR-UDA data augmentation method to generate photorealistic agricultural images by transforming synthetic 3D crop models into real-world crop domains. The key innovation of their work was a task-aware, semantically constrained CycleGAN model designed to map images from a synthetic agricultural domain to a real-world domain while preserving task-relevant semantics. The study focused on vineyard grape detection using YOLOv3, assuming a synthetically generated dataset with ground-truth labels in the source domain created via a 3D rendering engine. Additionally, they had a large collection of images from the target domain, though only a small subset was labeled. Initially, they trained an object detection model using synthetic source-domain data and fine-tuned it with the limited labeled data from the target domain. The synthetic images from the source domain were then adapted to the target domain for data augmentation purposes. This was achieved using a CycleGAN model with an additional task-specific semantic constraint loss to preserve spatial semantics such as grape position and size. For this, the object detection model (with frozen weights) was used to ensure that the detections agreed with the ground truth. As a result, the domain-adapted images remained well-aligned with the generated semantic labels in the source domain. The domain-adapted images were subsequently used to fine-tune the object detection model further. Their experiments demonstrated that using domain-adapted images improved grape detection performance compared to using unadapted synthetic images.

In related work, Cieslak et al. [10] introduced a procedural modeling technique for generating synthetic images of soybean plants and weeds in a field to support semantic segmentation tasks for autonomous weed control. Their model can simulate various growth stages of soybean plants, diverse soil conditions, and randomized field arrangements under diverse lighting conditions. Their approach involved modeling the shape and size of these plants, using computer graphics software (Blender) to simulate different lighting conditions and how light interacts with objects in the field. To improve the photorealism and practical utility of synthetic data, real-world textures, and environmental factors were incorporated into the procedural generation process. To mitigate the synthetic-to-real domain gap (i.e., the discrepancy between synthetic and real data distributions), domain adaptation was performed using SR-UDA based on the Contrastive Unpaired Translation (CUT) model [48], enabling the transformation of synthetic images to better align with real-world data. Their experiments showed that training semantic segmentation models with synthetic and real data in the agricultural domain could achieve equal or even superior performance compared to using real datasets of the same size. Surprisingly, however, the authors found that the domain-adapted images did not improve performance compared to synthetic images. They hypothesized that this was because (i) the ImageNet-1k weights used for pre-training did not capture features relevant to their semantic segmentation task, and (ii) the domain-adapted images contained artifacts that disrupted the semantic consistency between the images and their corresponding labels.

The above methods focused on data augmentation using synthetic data generation. SR-UDA was employed in this context to bridge the domain gap between synthetic and real data. However, RR-UDA can also be employed to bridge the performance gap due to the domain shift between source and target domains. Traditional techniques, for example, based on contrast stretching [49, 50], have been proposed in this context but with limited success. Magistri et al. [12] proposed an RR-UDA method for robust crop-weed-soil segmentation across various field conditions and crop types. Their approach used labeled images from the source domain and unlabeled images from the target domain. Similarly to Fei et al. [9], their goal was to generate domain-adapted images that closely resembled the target domain while preserving the semantic content of the source domain. However, unlike [9], they enforced the semantic constraint by jointly learning to (i) translate source domain images into the target domain and (ii) perform semantic segmentation in the target domain using only source domain labels. A modified version of the CUT model was used to generate both the domain-adapted images and corresponding segmentation masks in the target domain. The labels from the source domain and the predicted segmentation masks from the target domain were used to define an additional loss function based on a modified IoU function and enforce the semantic consistency between the source and target domains. The domain-adapted images and the source domain labels were then used to train a segmentation model in the target domain using CNNs. The method was validated and compared with other methods by performing experiments using both ground- and UAV-collected images across different field conditions and crops. Results for semantic segmentation were somewhat mixed, with the most favorable outcomes observed on the largest datasets. Interestingly, the Fréchet Inception Distance (FID), a metric commonly used to evaluate the quality of image-to-image translation, did not correlate with improved segmentation performance in the target domain.

In a related work [11], semantic consistency was maintained by ensuring that both source and domain-adapted images were segmented similarly. This was achieved by jointly training a CycleGAN model for domain adaptation alongside a supervised semantic segmentation model for the target domain, using domain-adapted images and source domain labels. Additionally, a separate semantic

segmentation model was trained in a supervised manner using source domain images and labels; however, its weights remained frozen during the optimization of the domain adaptation model and the target-domain semantic segmentation model. To enforce cross-domain semantic segmentation consistency, four loss functions, based on the same modified IoU function as in [12], were employed. They reported excellent performance across varying domains, including new field environments, different crops, and different sensors or robots, while relying solely on labeled data from the source domain. The above approach was motivated by the work in [14], where cross-domain semantic segmentation consistency was employed for domain adaptation in dense prediction tasks such as semantic segmentation, optical flow estimation, and depth prediction. In [15], an extra loss term based on the Fourier Transform (FT) phase was shown to further preserve semantic segmentation consistency and improve performance in weed segmentation. Recent remote sensing studies further emphasize the importance of preserving structural information under domain shifts. Gao et al. [51] showed that low-level patterns, such as lines and contours, can serve as domain-invariant features for multi-source remote sensing classification. More directly, Song et al. [52] used wavelet-based high-frequency guidance for heterogeneous remote sensing change detection, capturing shared structural details such as edges, textures, and object boundaries. These studies support the observation that pixel-level adversarial translation alone may be insufficient to preserve fine structural semantics, motivating additional semantic, structural, or frequency-domain constraints during domain adaptation.

RR-UDA has also been leveraged in several other agricultural applications, including leaf counting [53], fruit counting [54], and potato defect classification [55]. UDA methods have also been proven to be valuable in applications beyond agriculture. In [56], a similar approach to [10] was proposed for synthetic-to-real domain adaptation for semantic segmentation in robotics. To reduce the training time of the CUT model and enhance global consistency, they trained it using random patches rather than entire images. The resulting domain-adapted images achieved higher segmentation performance than synthetic images and performed comparably to real images. While combining real and domain-adapted images provided additional improvements, the gains were not significantly greater than those achieved by combining synthetic and real images. Hoffman et al. [13] proposed CyCADA, a domain adaptation approach combining cycle consistency with task-specific semantic consistency in a CycleGAN model as in [9, 12]. In addition to adapting representations at the pixel level (i.e., pixel-level consistency), they also adapt representations at the feature level (i.e., feature-level consistency) by adding a new loss function enforcing that both domain-adapted and target images yield similar levels of performance on the task-specific task. In their work, they considered improving the performance of object recognition (digit recognition across domains) and semantic segmentation (urban scenes across domains) by transferring real images between domains. Moreover, they applied CyCADA to transform synthetic images to a real domain, such as in [9, 10]. In [57], domain adaptation was used to build a robust pedestrian detection system in the thermal infrared spectrum. Despite the effectiveness of thermal infrared imaging in detecting pedestrians, particularly in low-light conditions, progress in this field has been hindered by the lack of a sufficiently large, annotated thermal infrared image dataset. To address this challenge, the authors leveraged the availability of large annotated image datasets in the visible spectrum and domain adaptation for data augmentation. Specifically, they employed a CycleGAN model to convert annotated pedestrian images from the visible spectrum into the thermal infrared spectrum, enhancing the training of a pedestrian detector in the thermal infrared domain. They trained the domain adaptation model and pedestrian detector in two different ways: (i) separately (two-stage) and (ii) jointly (end-to-end). They reported a reduction in the log-average miss rate by more than 12% on the KAIST detection benchmark, with the two-stage approach performing better than the end-to-end approach. A related approach was proposed in [58] for thermal infrared tracking.

In this study, we deploy RR-UDA to mitigate performance gaps caused by domain shifts between the source and target domains. We leverage RR-UDA for data augmentation in training and domain generalization in testing. Unlike traditional data augmentation schemes based on synthetic data, we augment the training data by merging real-world datasets, which inherently capture diverse plant growth stages, abnormalities, and field conditions. While synthetic data generation offers the advantage of automatically producing ground truth labels, creating highly realistic synthetic images that accurately reflect real field conditions and crop variations remains a significant challenge. Scaling synthetic data generation across different field conditions and crop types requires considerable time and effort. Our approach addresses these challenges by first mapping different image datasets to a *shared domain* using RR-UDA. As new data become available, they can be mapped to the same shared domain and added to the existing training dataset, continuously improving both the quantity and quality of the training set. Incoming new images can also be adapted to the same shared domain in order to improve generalization performance.

3 Methodology

Given UAV-captured high-resolution RGB images of a maize field, the objective is to determine whether an image contains abnormal corn plants and to quantify the overall level of abnormality. Using the UAV's position and orientation, it is also possible to estimate the geographic area in the field exhibiting signs of abnormality. However, precise localization of individual abnormalities or classification into specific abnormality types is beyond the scope of this work. To quantify the level of abnormality in an image, we apply a *ViT* regressor [17], denoted as ViT_{rg}, to non-overlapping patches of the image using a sliding window approach. Specifically, we slide an $W \times W$ window from left to right, top to bottom, with a stride equal to W and apply the ViT_{rg} to estimate the proportion of abnormal pixels relative to the patch size. Details on how this ratio is determined are provided in Sect. 4. An overall "Abnormality Probability" (AP) is then computed at the image level by summing the abnormality scores predicted for all non-overlapping patches. AP values above zero (or a properly chosen threshold) indicate the presence of abnormalities at the whole-image level. Thus, the output of the patch-level ViT_{rg} model can be used for both detecting and quantifying abnormalities in maize fields.

In addition to evaluating *ViT* model, we tested several alternative architectures, including ResNet, EfficientNet, and DenseNet. Among these, *ViT* consistently achieved the highest performance in our experiments. Consequently, we selected *ViT* to validate the proposed framework, although other models could be used as well. We have used *ViT*-B/16 architecture that is the base Vision Transformer model which takes images of 224×224 spatial resolution and processes them into 16×16 patches and incorporating positional embeddings, enabling it to effectively learn global spatial relationships. There are 12 transformer encoder layers and 12 attention heads and a hidden layer size of 768. The model leverages multi-headed self-attention mechanisms to capture both local features and high-level semantic information while preserving spatial relationships. In our implementation, the standard *ViT*-B/16 architecture was modified by replacing its final fully connected layer with a convolutional layer. This modification enables the model to process multiple image patches efficiently in a sliding-window manner, following the strategy used in OverFeat [59]. Specifically, the convolutional layer was configured with 768 input channels and one output channel, corresponding to the dimensionality of the *ViT*-B/16 feature representation and the scalar abnormality prediction, respectively. A kernel size of 1×1 and a stride of 1 were used, allowing the model to generate patch-level predictions while preserving the spatial arrangement of the extracted features.

The quality of AP estimates is based on the accuracy and robustness of the ViT_{rg} patch-level model. The model is trained using randomly extracted $W \times W$ patches from the training set as detailed in Sect. 4. To improve the performance of the ViT_{rg} patch-level model, we leverage traditional transfer learning and fine-tuning strategies. Most importantly, we use a data augmentation strategy based on RR-UDA as discussed earlier. DL models typically require large, diverse datasets to generalize effectively. In our application, training the ViT_{rg} patch-level model on data collected from a single field limits its generalization performance due to insufficient and non-representative data. However, training the ViT_{rg} patch-level model on a more comprehensive training set, including data collected from different fields, improves its generalization performance both on novel data from the same field or data from different fields. The key challenge is how to best combine data from different fields into a "unified" training set. Simply adding data collected from different fields to a common training set is not the best strategy. This is because variations in the data collection process (e.g., different lighting conditions) and field types (e.g., different soil types) will introduce irrelevant information in the training data, hurting training.

To address the challenges of domain variability, we propose transforming data collected under different maize field conditions into a *shared* domain before aggregating them into a unified training set. Specifically, we introduce a data augmentation strategy that maps images from various fields into a *shared* domain using an ensemble of CycleGAN models [16]. Assuming that data is collected from n different fields under varying conditions (e.g., different times, plant growth stages, or abnormality levels), we train a separate CycleGAN model for each field i, mapping its data into a common domain D. In principle, D can be any one of the n fields. The resulting unified training set is constructed by aggregating all transformed data in the shared domain D, allowing the ViT_{rg} model to be trained on a dataset that captures realistic, field-specific variations, unlike purely synthetic augmentation which may lack realism.

Figure 1 (Stage 1) illustrates the proposed augmentation scheme where y_q corresponds to the patch-level prediction. As new data becomes available, it can be added to the unified training data set by first adapting them to the shared domain D using one of the existing CycleGAN model or by training a new CycleGAN model. The primary motivation for employing an ensemble of CycleGAN models in this work is to mitigate domain imbalance issues. Nevertheless, it is possible to train a single CycleGAN model using techniques specifically designed to handle domain imbalance, such as the method proposed in [60]. Figure 1 (Stage 2) illustrates how test images are processed where Y_q corresponds to the whole-image AP prediction. The main idea

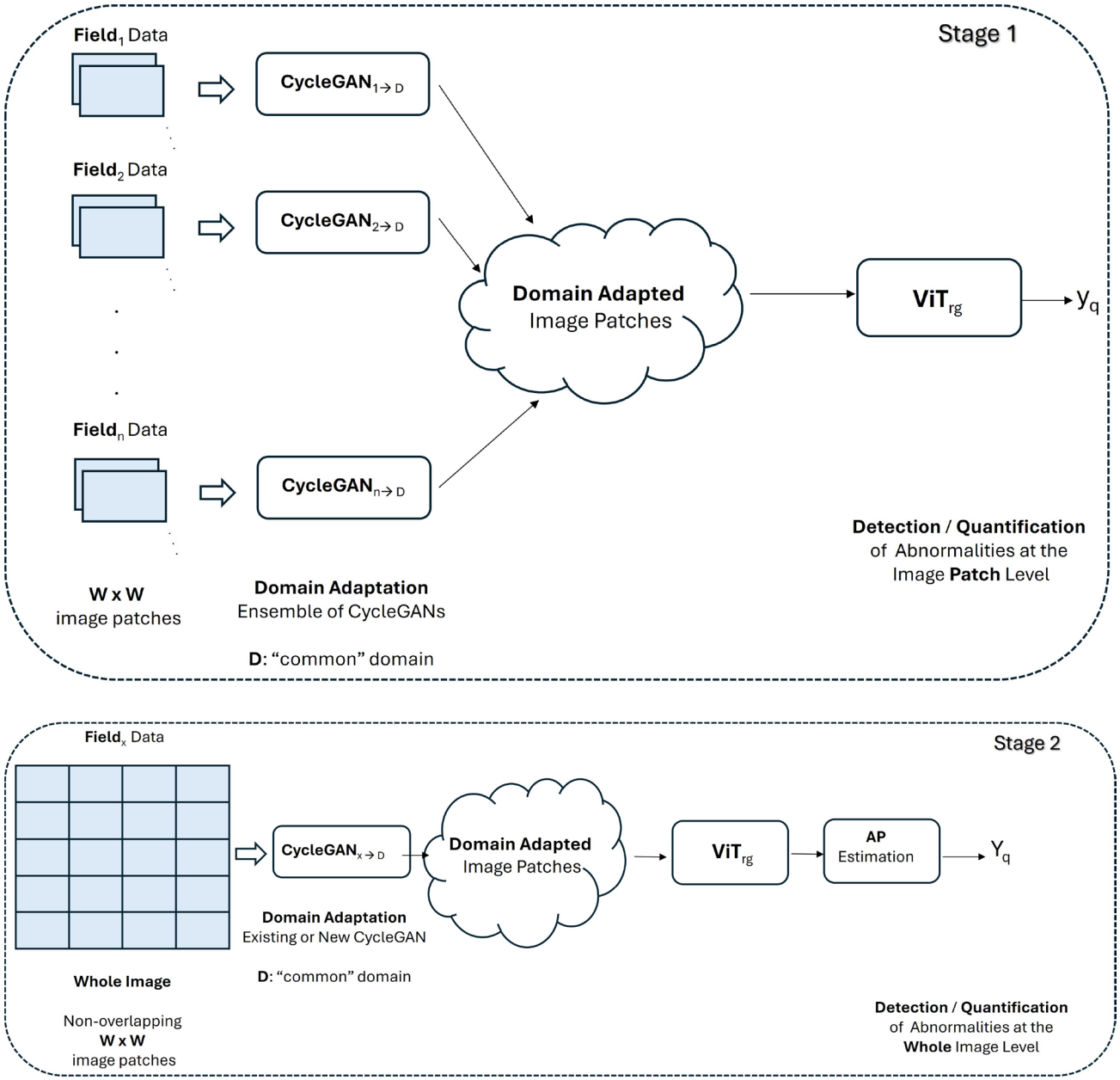


Fig. 1 The two stages of the proposed methodology. In Stage 1, image patches from different fields are transformed into a shared domain D. The domain adapted patches are used to train the ViT_{rg} model for patch-level detection and quantification of abnormalities. y_q corresponds to the patch-level predictions. In Stage 2, non-overlapping patches are extracted from a test image and adapted to the shared domain D. The AP is then estimated at the whole image level. Y_q corresponds to the whole-image level prediction

is adapting test images to the same shared domain D before estimating the AP value. This can be performed again by applying one of the existing CycleGAN models or by training a new CycleGAN model.

CycleGAN is a powerful unpaired image-to-image translation method; however, alternative variants such as CUT can also be employed, as discussed in Sect. 2. CycleGAN consists of two generator networks: G, which adapts images from the source domain X to the target domain Y, and F, which performs the inverse adaptation from Y to X. Additionally, two discriminator networks, D_Y and D_X, are used to distinguish between real and generated images in each domain. This bidirectional transformation ensures that the model captures domain-specific information without the need for paired training images. CycleGAN is trained using three primary loss functions: adversarial loss, cycle-consistency loss, and identity loss. The adversarial loss, which encourages the generator to produce images indistinguishable from real ones in the target domain, is defined for the adaptation of images from X to Y using G as:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = E_{y \sim P_{data}(y)}[logD_Y(y)] + E_{x \sim P_{data}(x)}[log(1 - D_Y(G(x)))]$$

Similarly, the adversarial loss function for mapping images from Y to X using F is given by:

$$\mathcal{L}_{GAN}(F, D_X, X, Y) = E_{x \sim P_{data}(x)}[logD_X(x)] + E_{y \sim P_{data}(y)}[log(1 - D_X(F(y)))]$$

To prevent mode collapse in the forward transformation, where all images from X might be mapped to a single image in Y, CycleGAN incorporates a cycle-consistency loss. This loss ensures that the inverse transformation accurately reconstructs the original images when mapped back from Y to X. The corresponding loss function is defined as:

$$\mathcal{L}_{cyc}(G, F) = E_{x \sim P_{data}(x)}[||F(G(x)) - x||_1] + E_{y \sim P_{data}(y)}[||G(F(y)) - y||_1]$$

To prevent unnecessary modifications to images that already belong to the target domain, an identity loss is introduced. This loss ensures that when an image from the target domain is passed through the generator, it remains unchanged. The identity loss is formulated as:

$$\mathcal{L}_{identity}(G, F) = E_{x \sim P_{data}(x)}[||F(x) - x||_1] + E_{y \sim P_{data}(y)}[||G(y) - y||_1]$$

Using the individual loss functions above, the total CycleGAN loss function is given by:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = \lambda_{GAN}(&\mathcal{L}_{GAN}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{GAN}(F, D_X, Y, X)) \\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{id}\mathcal{L}_{identity}(G, F)\end{aligned}$$

In the rest of the paper, we collectively refer to the combination of cycle-consistency loss and identity loss as the "reconstruction" loss.

Although domain adaptation can, in general, enhance the generalization performance of the *ViT* models, special attention must be given to only using high-quality adapted data for training and testing purposes. Specifically, transforming data from one domain X to another domain Y does not always yield high-quality adapted data, and many times, undesirable artifacts can be introduced during the domain adaptation process. Augmenting the training data set with low-quality adapted data will affect the performance of the *ViT* models. Similarly, using poorly adapted data during testing will affect accuracy. Therefore, it is important to filter out poorly adapted data for training and testing purposes; we refer to this idea as *selective* domain adaptation. In general, certain domain adaptation transformations are "easier" to learn than others, and the domain-transformed images do not suffer from significant artifacts. Different statistics (e.g., mean, median, standard deviation, and kurtosis) of the CycleGAN reconstruction loss can be used to assess the quality of domain-adapted data. In this work, we identify poorly domain-adapted data by calculating the Peak Signal-to-Noise Ratio (PSNR) between the original and domain-adapted images. Using PSNR scores, we compute the mean and standard deviation based on the loss values of the CycleGAN model; outliers can then be removed from the training and testing processes. We provide additional details and examples on this issue in Sect. 5.

4 Dataset

We have experimented with a publicly available data set which contains UAV captured, high resolution RGB images of maize plants at various growth stages, exhibiting abnormalities due to nutrient deficiencies (e.g., nitrogen deficiency) and dryness [18]. The primary assumption underlying the collection of this dataset is that a high-altitude flight can first be used to survey the entire field, identifying potential areas where plants may exhibit abnormalities (e.g., using the Normalized Difference Vegetation Index (NDVI)). The GPS locations of these areas can then be transmitted to a small-scale UAV as waypoints, enabling an automated low-altitude flight (less than 15 m) where high-resolution RGB images can be captured. The images were captured using Sentera 65R payload, which is capable of capturing 65MP images with a ground sampling distance (GSD) of 1.5 mm at 90 feet altitude, and 0.45cm GSD at the FAA-allowed maximum altitude of 400 feet. This high-resolution capability enables the capture of clear, detailed, and motion-blur-free imagery even at operationally safe altitudes (Fig. 2).

The images were captured at two different fields: Becker and Waseca, MN. Specifically, the dataset contains corn plants at different vegetative growth stages: V5, V6, V8, V10, and V12 where "V" denotes the vegetative stage and the corresponding number indicates the count of fully developed leaves [18]. The Becker field dataset includes plants at the V8 and V12 growth stages with a higher level of abnormality, whereas the Waseca field includes plants at the V5, V6, and V10 stages with a lower level of abnormality. Besides differences in terms of growth stages and abnormality levels, the two fields also differ in terms of soil type as shown in Fig. 3. Compared to the Waseca field, the plants in the Becker field have more leaves, which obscure the background in the Becker images. In contrast, the Waseca field has fewer leaves, revealing more of the background along with dead leaves and debris from previous cultivation.

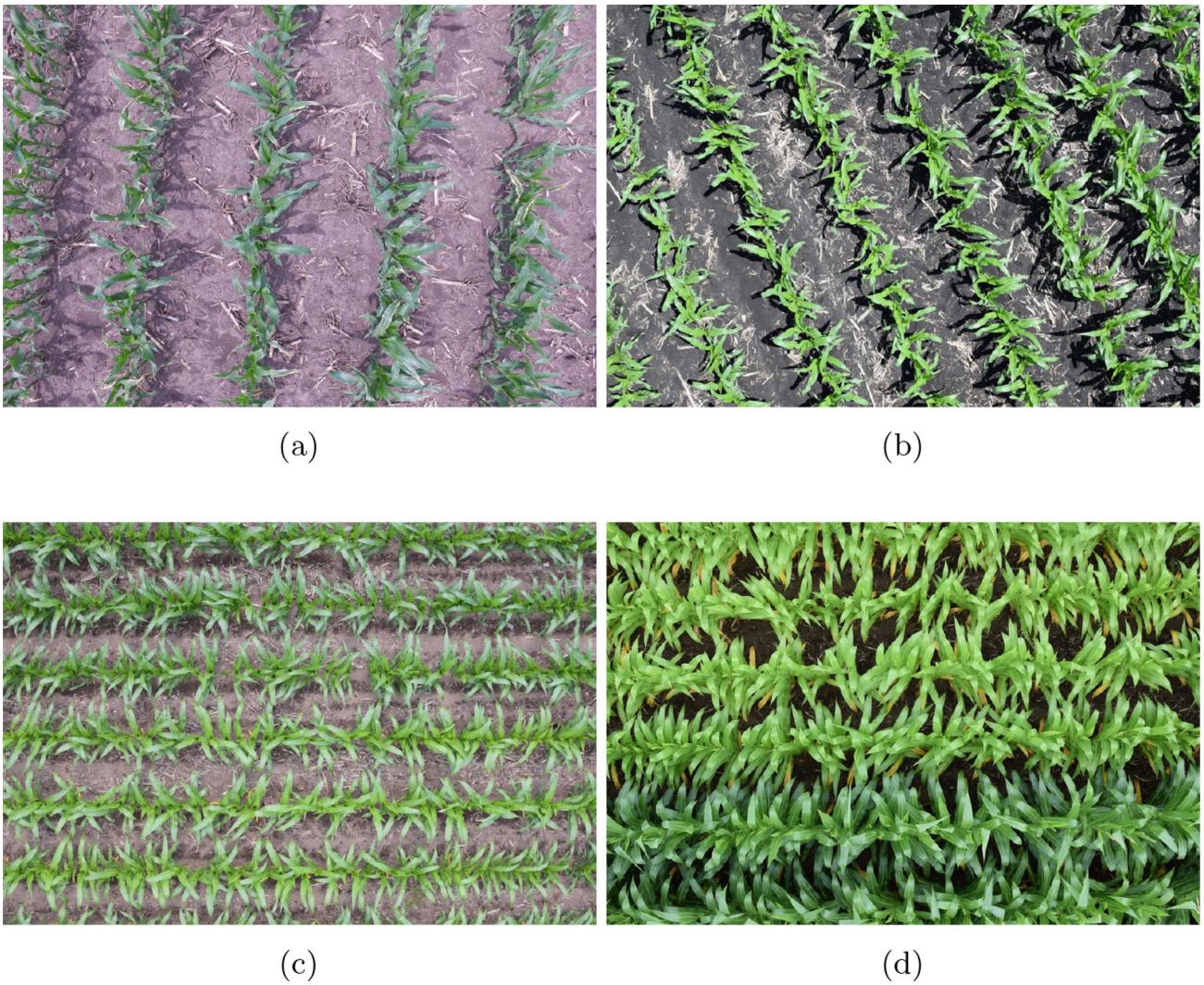


Fig. 2 Representative images from the Waseca (**a**, **b**) and Becker (**c**, **d**) fields

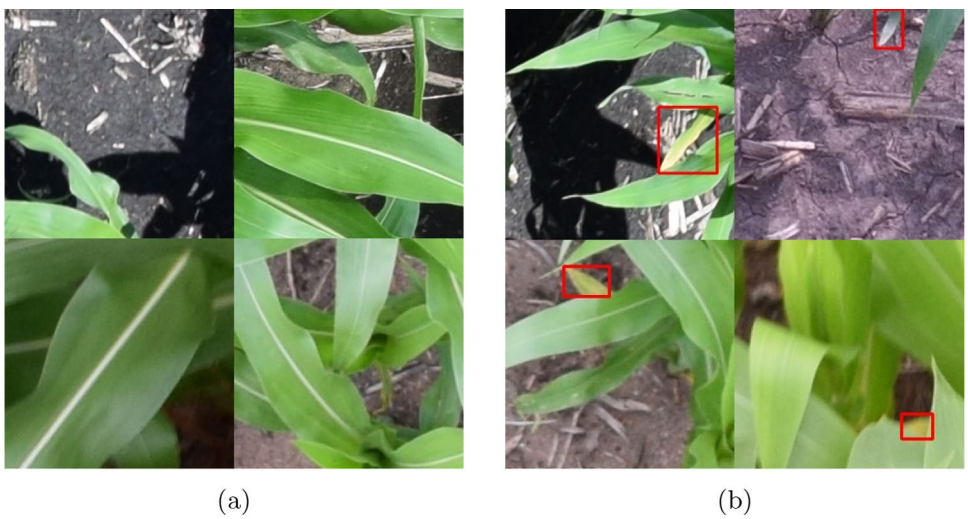


Fig. 3 Representative patch-level images from **a** the normal class, and **b** the abnormal class (i.e., abnormalities are enclosed by "red" boxes). Patches on the top row are from the Waseca field while patches on the bottom row are from Becker field

Additionally, the soil color in the Waseca field differs significantly from that in the Becker field. Image resolution varies across stages: V6, V8, and V10 images were captured at 4000×6000 pixels, V12 images at 6000×4000 pixels, and V5 images at 4000×3000 pixels. The original dataset did not contain any annotations for the abnormalities. We used Label Studio [61] to manually label abnormal plant leaves by drawing a rectangular bounding box around each leaf. Using bounding boxes allows for fast annotation of abnormalities but they lack localization accuracy which affects the precise quantification of abnormalities as discussed later. Using bounding box also overestimates the amount of abnormalities of the data.

The dataset contains 61, 64, 44, 64, and 48 images from the V5, V6, V8, V10, and V12 growth stages, respectively. To train the ViT_{rg} patch-level model, we extracted random image patches from 226 of the 281 images (i.e., 50, 51, 36, 51, and 38 from the corresponding growth stages). To test the performance of the patch-level models, we extracted random image patches from the remaining 55 images (i.e., 11, 13, 8, 13, and 10 respectively). Specifically, we randomly extracted patches of size $W \times W$ from both the training and test images. We chose $W = 250$ after some experimentation which was not exhaustive by any means. However, our objective in this study was to demonstrate the feasibility of the proposed framework rather than fully optimize its parameters. Patches containing abnormalities (i.e., containing any portion of a bounding box) were labeled as "abnormal," while those without containing any abnormalities (i.e., no portion of a bounding box) were labeled as "normal". These normal and abnormal patches were then used to train and test the ViT_{rg} patch-level model using a separate channel for each RGB color. The target values were calculated by dividing the total area of bounding boxes within a patch by the patch's total area. If a bounding box extended beyond a patch's boundaries, only the portion within the patch was considered. As previously noted, bounding boxes do not precisely localize abnormalities, which can lead to an overestimation of target values.

Table 1 Number of patch-level images (randomly extracted) from whole UAV images used for training and testing in the Becker and Waseca fields

	Growth stage	Training set		Testing set	
		Normal	Abnormal	Normal	Abnormal
Becker	V8	1124	1200	238	247
	V12	2401	3766	909	967
	Total	3525	4966	1147	1214
Waseca	V5	660	668	49	45
	V6	1302	1355	125	110
	V10	1815	1815	211	209
	Total	3777	3838	385	364

We randomly selected (x, y) coordinates from the full images to create 250×250 patches for our patch dataset. A patch was labeled as abnormal if any part of a bounding box annotation was contained in it; otherwise, it was labeled as normal. The number of generated patches for the abnormal class was approximately equal to the total number of bounding boxes present in the images for each growth stage. A similar approach was followed to extract the normal class. Table 1 shows the total number of randomly extracted normal and abnormal patches extracted for training and testing the ViT_{rg} patch-level model, categorized by field and growth stage. It should be noted that in the Becker field, fewer normal patches were extracted than abnormal ones due to the higher prevalence of abnormalities in V12 stage images. To maintain class balance during training, we applied data augmentation to increase the number of normal patches. Since our objective was to develop a general purpose framework for abnormality detection and quantification independently of growth-stage and field, the ViT_{rg} model was trained with patches from different growth stages and fields as discussed later in Sect. 5. Figure 3 shows some representative normal and abnormal patches extracted for this purpose.

As described in the methodology section, the ViT_{rg} patch-level model was applied on the whole images, using a non-overlapping sliding window approach, to compute the AP values. For this, we used the 55 test images mentioned earlier. Since whole images typically capture large areas (i.e., 110–130 plants on average), labeling an image as normal or abnormal might not be very useful compared to labeling a sub-image of the whole image, associated with a smaller area. Therefore, each test image was divided into quarters, each containing around 30–40 plants, effectively quadrupling the size of our test set (i.e., 220 test images).

5 Experimental results and analysis

We have performed a wide range of experiments to evaluate the proposed framework. Initially, we performed baseline experiments to evaluate the performance of the ViT_{rg} model on randomly extracted image patches. At the same time, we investigated the effectiveness of transfer learning and fine-tuning by pre-training on datasets less or more relevant to the domain of our problem. To further improve the generalization performance of ViT_{rg}, we then considered how to leverage domain adaptation using ensembles of CycleGAN models. By visualizing and analyzing the results from each CycleGAN model, we gained important insights on how to assess the relative difficulty among different domain transformations and inferred a strategy to remove poorly, domain adapted images (i.e., selective adaptation).

Next, we employed domain adaptation for data augmentation to further boost the performance of the ViT_{rg} patch-level model. The final ViT_{rg} patch-level model was used to predict AP for the detection and quantification of abnormalities at the whole image level. In all experiments, each patch-level ViT_{rg} model was trained for 200 epochs with a batch size of 64. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4} and an L2 regularization coefficient of 0.01. Mean Squared Error (MSE) was used as the loss function for the regression task. Each CycleGAN model was trained for 400 epochs with a batch size of 1, using the Adam optimizer with a learning rate of 1×10^{-4}.

5.1 Transfer learning and fine-tuning

Baseline experiments involved training and testing a separate ViT_{rg} patch-level model for each field using the randomly extracted patches described in Sect. 4. To prevent overfitting, we allocated 10% of the training data for validation purposes. This validation set was used not only to monitor overfitting but also to determine the optimal model weights that yielded the best performance.

To boost the performance of the ViT_{rg} patch-level models, we investigated transfer learning, a widely used approach for training deep learning models across various tasks more effectively. Typically, it involves leveraging the pre-trained weights of a model trained on a large dataset, such as ImageNet, and fine-tuning the weights of the last layer(s) for the downstream task. Using ImageNet pre-trained weights followed by fine-tuning, we achieved an improvement of about 2% for the Becker field, however, the performance for the Waseca field dropped by about 1%. We hypothesized that the drop in performance was because ImageNet contains mostly images outside of the domain of our application.

To validate our hypothesis, we considered the Plant Village dataset [62] for pre-training which contains 54,303 images from 38 classes of healthy and unhealthy plant leaves. Using Plant Village pre-trained weights followed by fine-tuning, we achieved an improvement on the Becker field by approximately 2.75% and an improvement on the Waseca field by about 3.07%. While it is possible to further improve these results using much larger, domain-related, datasets, we did not pursue this issue further as our primary focus is on exploring the benefits of domain adaptation. Consequently, we adopted Plant Village pre-training and fine-tuning to optimize the performance of ViT_{rg} patch-level models in all subsequent experiments.

5.2 Domain adaptation in maize fields

Cross-field experiments have confirmed that when training a model on data from field X, its performance will typically drop when tested on data from a different field Y. Domain adaptation can address this issue by mapping the data from field Y to the same domain as the data from field X. This should increase the cross-field performance of the ViT_{rg} patch-level model since the domain-adapted data from field Y will look much more similar to the data from field X. In general, domain adaptation can help to improve both single-field (i.e., train and test on the same field) and cross-field (i.e., train and test on different fields) performance. When data from multiple fields is available, they can be transformed to a shared domain D before training or testing. A ViT_{rg} parch-level model can then be trained using data from multiple fields instead of a single field, as shown in Fig. 1 (Stage 1).

The ViT_{rg} patch-level model should be expected to achieve higher accuracy and robustness in this case, as it will be trained with a much larger and comprehensive dataset (e.g., containing many "visually" different abnormalities and plants at different growth stages). At the same time, novel images can be analyzed for abnormalities by first adapting them to the same shared domain "D" in an effort to improve generalization performance as shown in Fig. 1 (Stage 2). To be successful, domain adaptation should preserve information related to the abnormalities present in each field while "normalizing" irrelevant information (i.e., information not related to abnormalities), such as differences due to lighting conditions and soil type. This will ensure that the ViT_{rg} patch-level model will perform consistently across diverse field types. In this section, we report a number of domain adaptation experiments between the Becker and Waseca maize fields using CycleGANs to validate the above insights. Each cycleGAN model was trained with image patches from the corresponding classes and fields.

First, we investigated how the choice of the shared domain D affects performance. In principle, any of the maize fields represented in the dataset can be chosen as the shared domain D; however, it should be expected that some maize fields might perform better than others. To explore this issue, we aimed to determine whether converting image patches from Becker to Waseca (i.e., B2W, assuming D = W) is more challenging than converting image patches from Waseca to Becker (i.e., W2B, assuming D = B). For this, we trained a CycleGAN model for each domain transformation (i.e., B2W and W2B) and analyzed the CycleGAN reconstruction values in each case. Figure 4 shows the histogram of CycleGAN's reconstruction loss values for the normal and abnormal classes for the B2W and W2B transformations.

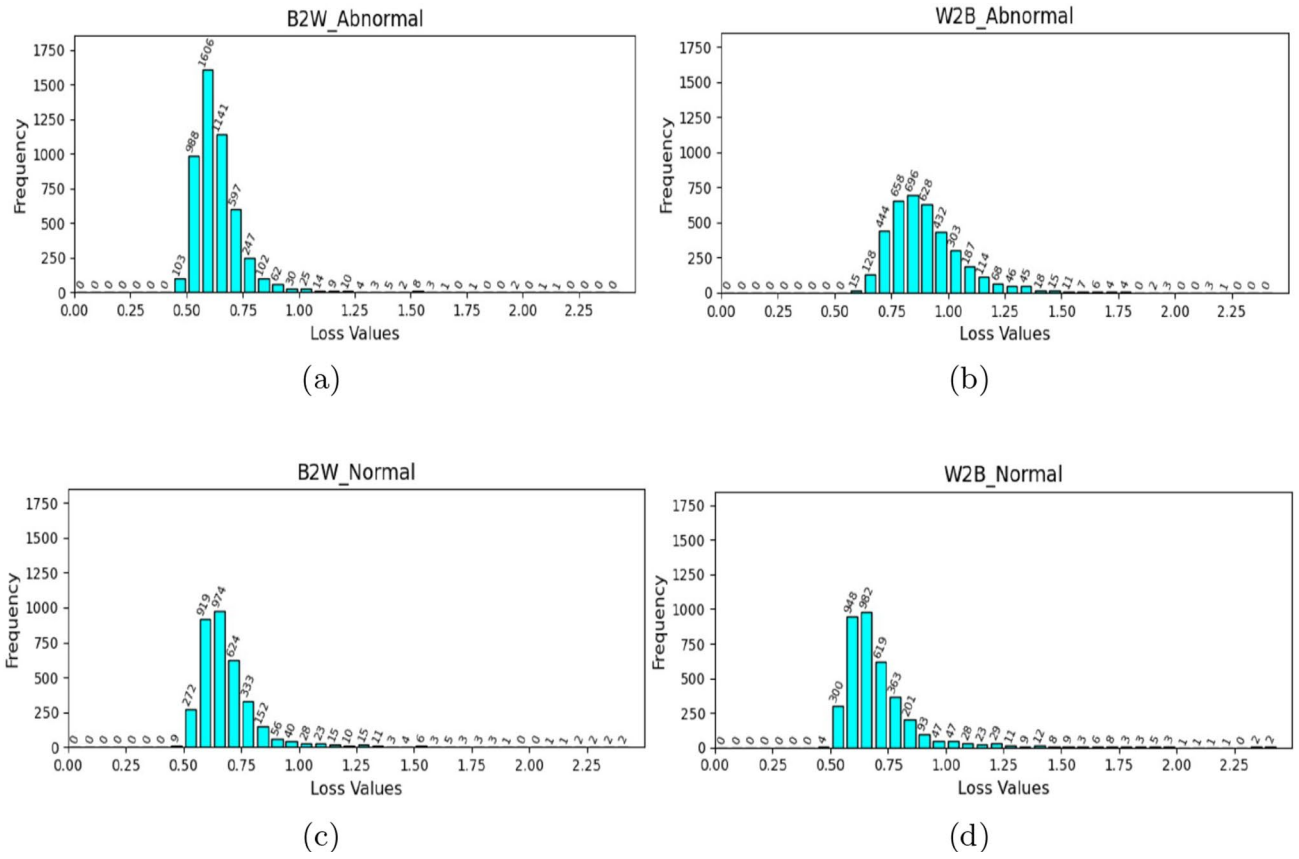


Fig. 4 Histogram of the CycleGAN reconstruction loss values in the case of the abnormal class for **a** B2W **b** W2B and normal class for **c** B2W and **d** W2B

Table 2 Statistics computed from the reconstruction loss histograms of the B2W and W2B domain adaptations for the normal and abnormal classes

	B2W abnormal	W2B abnormal	B2W normal	W2B normal
Mean	0.64	0.90	0.69	0.71
Median	0.59	0.84	0.65	0.65
Std	0.12	0.17	0.18	0.20
Skewness	**4.02**	1.79	4.77	4.75
Kurtosis	**31.99**	6.32	35.21	38.87

Table 2 summarizes some key statistics obtained by fitting a skewed Gaussian distribution to each histogram. As observed, the histograms and statistics for the normal class are highly similar between B2W and W2B, suggesting that both transformations present a comparable level of difficulty. However, for the abnormal class, the distributions differ significantly, indicating varying degrees of difficulty between the two transformations. In particular, the skewness and kurtosis values for B2W are substantially higher at 4.02 and 31.99 respectively than those for W2B which are 1.79 and 6.32. We speculate that these statistical differences could serve as indicators of the "difficulty" of domain adaptation transformations in a broader context, although further validation is required. The increased difficulty for the abnormal class may stem from its greater variability, as abnormalities in the Waseca field differ markedly from those in the Becker field (see Fig. 3). Given that more challenging transformations result in poorly adapted images (see Fig. 11), an optimal shared domain D could potentially be selected by favoring transformations that are easier to learn. Additional insights on this issue are provided in Sect. 5.3.

To further evaluate the model's domain adaptation performance, we created several visualizations to examine how well the domain-adapted images are distributed within the shared domain D. Specifically, we visualized the original source and target domain images, along with the domain-adapted images, using t-SNE visualization [63]. Figure 5 illustrates the t-SNE visualizations for the B2W and W2B domain adaptations. For these visualizations, we randomly selected 300 image patches (half from each field) and transformed the Becker image patches to the Waseca field and the Waseca image patches to the Becker field. The visualizations show that the domain-adapted images fall nicely within the shared domain D in each case. Interestingly, a few domain-adapted image patches remain within the source domain cluster. We hypothesize that this occurs

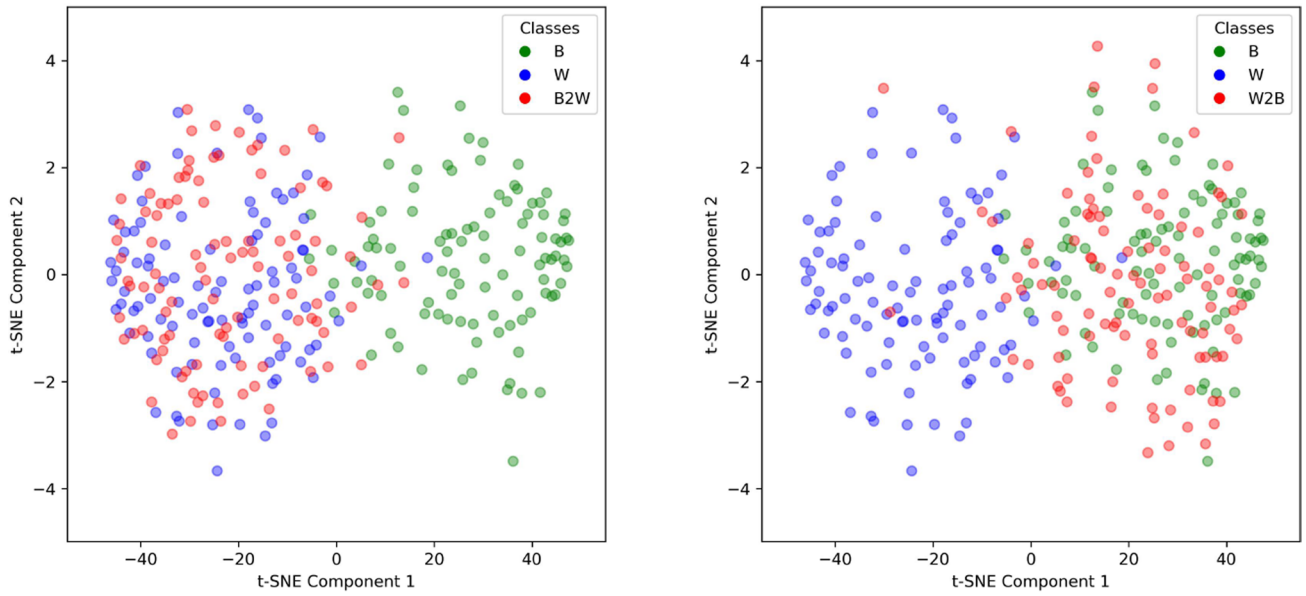


Fig. 5 t-SNE visualization of original and domain adapted image patches using **a** D = W, and **b** D = B. In **a**, we can observe that the B2W domain adapted image patches (red) are closer distributed to the Waseca image patches (blue) than the original Becker image patches (green). In **b**, we can observe that the W2B domain-adapted image patches (red) are closer distributed to the Becker image patches (green) than the original Waseca image patches (blue)

due to inherent similarities between Becker and Waseca field images. As observed, some source Waseca images naturally overlap with Becker images and vice versa. This overlap likely leads the CycleGAN model to generate certain domain-adapted images that remain within the same neighborhood as the source domain. Maximum Mean Discrepancy (MMD) was also computed to further evaluate the quality of the domain adaptation. MMD measures the discrepancy between two distributions, where a lower value indicates that the adapted image distribution is closer to the target-domain distribution. In this experiment, the MMD scores for the B2W and W2B adaptations were 0.04 and 0.01, respectively. These results suggest that both adapted distributions are closely aligned with their corresponding target domains. The lower MMD score for W2B indicates that the W2B adapted images are more closely matched to the Becker domain than the B2W adapted images are to the Waseca domain. Overall, the relatively small MMD values provide additional evidence that the adaptation process was effective in reducing the distribution gap between the source-adapted and target-domain images.

5.3 Data augmentation using selective domain adaptation

In this section, we present additional experiments to evaluate the benefits of selective domain adaptation for data augmentation. Specifically, we conducted experiments to quantify the extent of abnormality in each patch and to evaluate how domain adaptation-based augmentation improves performance. To ensure a fair comparison, we used the same source-domain test sets across all experiments. To objectively evaluate cross-field performance where the amount of abnormality might vary, we also report Normalized MSE (NMSE) scores (i.e., MSE scores divided by the variance of the ground-truth values).

Table 3 Same-field and cross-field patch-level quantification performance using original and selective domain-adapted data for data augmentation

Trained on	Train MSE	Valid MSE	Test MSE/ NMSE (Becker)	Test MSE/ NMSE (Waseca)
Becker	7.80	8.28	9.84/0.55	2.58 /0.71
Becker+W2B	6.68	5.87	**8.00/0.54**	2.13/0.87
Waseca	11.08	10.49	17.46/0.97	8.32/1.52
Waseca + B2W	11.11	9.80	9.01/0.50	**1.97/0.57**
Waseca + Becker	6.27	5.93	8.21/0.57	3.12/0.90

The results of these experiments are summarized in Table 3. Initially, we trained a separate ViT_{rg} model for each field using only the respective patch-level images from each domain. Using MSE and NMSE scores as performance metrics, the same-field performance on the Becker field was 9.84/0.55, while that on the Waseca field was 8.32/1.52. Next, we introduced data augmentation based on selective domain

adaptation in both fields. For the Becker field, the training set was expanded to include both original Becker data and selective domain-adapted Waseca data. When training the ViT_{rg} patch-level model with the augmented dataset, the MSE/NMSE scores dropped to 8.00/0.54. In the case of the Waseca field, there was an even higher drop of MSE/NMSE scores to 1.97/0.57. We also noticed cross-field performance improvements. For comparative evaluation, we additionally trained a model using a direct data aggregation strategy in which the Becker and Waseca images were merged without adapting these images to a shared domain. Under this setting, the model achieved test errors of 8.21/0.57 on the Becker field and 3.12/0.90 on the Waseca field. Although this approach yielded an improvement over training exclusively on Becker data for the Becker field, the performance remained inferior to that obtained using selectively domain-adapted data. Moreover, for the Waseca field, this strategy resulted in a substantial degradation in performance.

Figure 6 provides a visual comparison between the predicted abnormality scores produced by the proposed model and the corresponding patch-level ground truth abnormality values for representative test images. In Fig. 6a, the predicted score for the abnormal region is close to the ground truth value, indicating that the model can identify abnormality even when it is partially occluded by leaves. Figure 6b shows a normal image with a ground truth abnormality value of zero, for which the model predicts a low score of 0.04. This small deviation is expected, as exact agreement is difficult in a regression-based setting. In Fig. 6c, the model incorrectly interprets the plant tassel as an abnormal region, resulting in an overestimated prediction. Similarly, Fig. 6d shows another case where the model overestimates the abnormality score. Overall, these examples demonstrate both the ability of the model to approximate abnormality severity and some of its limitations in visually ambiguous cases.

5.4 Abnormality quantification in whole images

In this section, we focus on quantification of abnormalities in whole images using the methodology described in Sect. 3 and illustrated in Fig. 1. The generation of the test images used in these experiments are detailed in Sect. 4. For each test image, we assess the presence of abnormalities by estimating its AP value using the ViT_{rg} patch-level model. To enhance the visualization of our results, test images have been arranged in ascending order based on their ground truth AP values (i.e., images on the left correspond to lower AP values, while those on the right represent higher AP values). Figure 7a and c present the predicted AP values (in blue) alongside the ground truth values (in black) for the Becker and Waseca fields, when the same field is used for both training and testing, without any data augmentation. As shown, the model tends to overestimate AP values for images with lower levels of abnormality and underestimate them for images with higher levels of abnormality.

Applying data augmentation leads to improved predictions in both fields, as illustrated in Fig. 7b and d. It is important to note the difference in scale between the Becker and Waseca plots. In practice, the level of abnormality in the Waseca field is significantly lower than that in the Becker field, which makes the Waseca predictions appear more variable or noisy. For a more objective comparison, Table 4 presents the Mean Squared Error (MSE) and Normalized MSE (NMSE) values for each case. Interestingly, despite the apparent visual noise, the NMSE for the Waseca field is much lower (0.0002/0.0002) than that for the Becker field (0.0074/0.0073) in same-field evaluations.

Cross-field performance is illustrated in Fig. 8a and c. Similar to the same-field experiments, the predicted AP values tend to be overestimated for images with lower levels of abnormality and underestimated for those with higher levels of abnormality. Applying data augmentation leads to improved predictions in both fields, as illustrated in Fig. 8b,

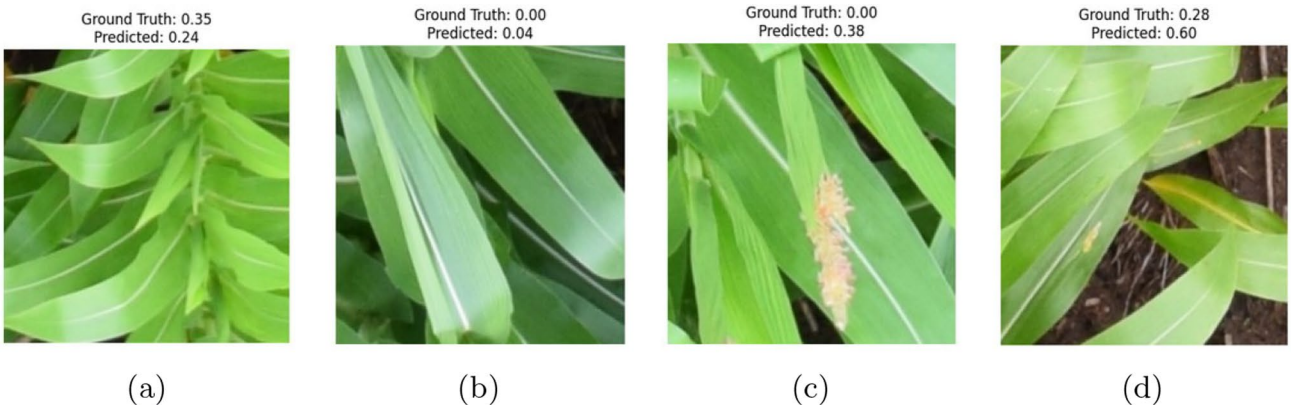


Fig. 6 Visual comparison of the proposed model's predictions with the corresponding ground-truth abnormality scores for representative test images

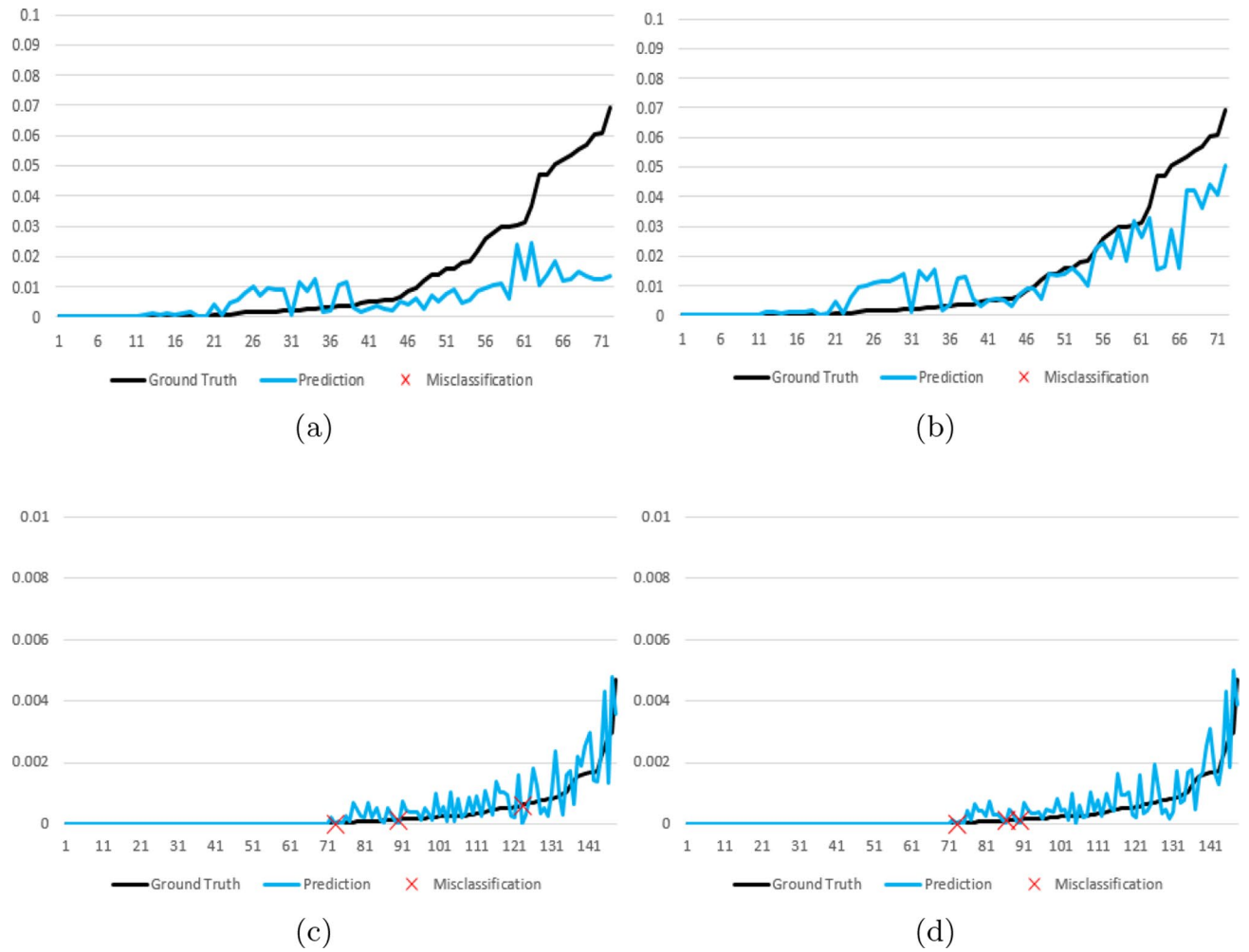


Fig. 7 Same-field, AP-based whole-image predictions: **a** trained without data augmentation on Becker field, **b** trained with data augmentation on Becker field (e.g., Becker+W2B data), **c** trained without data augmentation on Waseca field, and **d** trained with data augmentation on Waseca field (e.g., Waseca+B2W data). Test images have been arranged in increasing order with respect to the ground truth AP values. Ground truth AP values are shown in "black" while predicted AP values are shown in "blue". The "red" crosses indicate the test images misclassified. AP values have been scaled in the range [0,100]

Table 4 Same-field and cross-field, AP-based whole-image detection and quantification performance using the original test images

	Test Acc (%) (Becker)	Test MSE/NMSE (Becker)	Test Acc (%) (Waseca)	Test MSE/NMSE (Waseca)
Becker	100%	0.0001/0.0074	93.91%	1.9e−07/0.0002
Becker+W2B	100%	0.0001/0.0073	94.59%	1.2e−07/0.0001
Waseca	98.61%	0.0003/0.015	97.97%	1.8e−07/0.0002
Waseca+B2W	98.61%	9.9e−05/0.005	97.97%	1.6e−07/0.0002

d and further supported by the Mean Squared Error (MSE) and Normalized MSE (NMSE) values reported in Table 4. The Waseca predictions appear more noisy again, primarily because a different scale is used to visualize the results, reflecting the lower level of abnormality present in the Waseca field. It is worth noting, however, that the Normalized Mean Squared Error (NMSE) decreased by more than 50% (i.e., from 0.015 to 0.005) when training on Waseca and testing on Becker.

In addition to their role in abnormality quantification, AP values can also be used for abnormality detection, as non-zero values indicate the presence of abnormalities. To evaluate this potential, we conducted additional experiments assessing the effectiveness of AP-based detection. In our method, the smallest AP value among the abnormal training samples was used as a threshold, denoted T_q. A test image was classified as abnormal if its AP value exceeded T_q, and as normal otherwise. Misclassified test images are marked with a red cross in Figs. 7 and 8. Data augmentation had

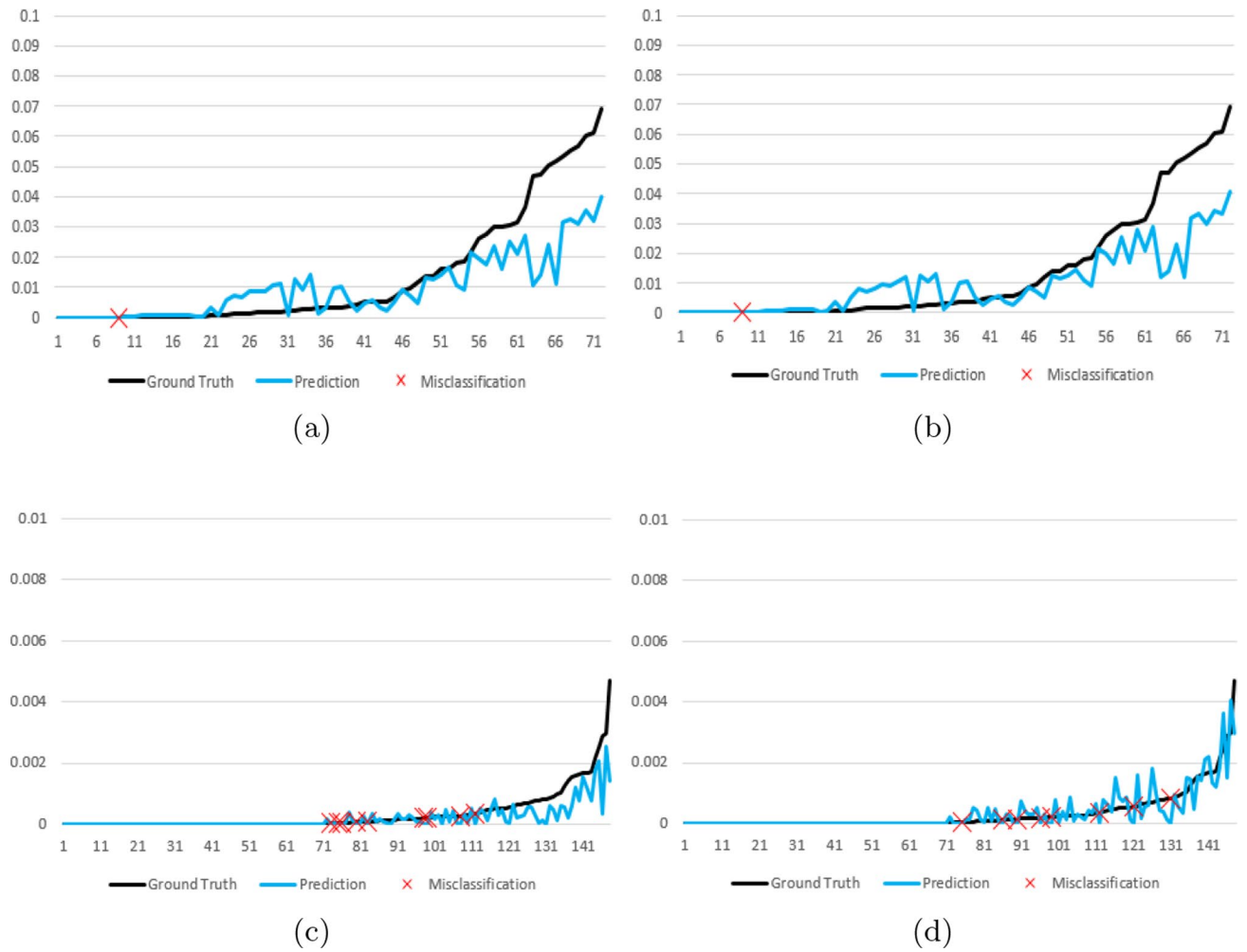


Fig. 8 Cross-field, AP-based whole-image predictions: **a** trained without data augmentation on Waseca field and tested on the Becker field, **b** trained with data augmentation on the Waseca field (e.g., Waseca+B2W data) and tested on the Becker field, **c** trained without data augmentation on Becker field and tested on the Waseca field, and **d** trained with data augmentation on Becker field (e.g., Becker+W2B data) and tested on the Waseca field. Test images have been arranged in increasing order with respect to the ground truth AP values. Ground truth AP values are shown in "black" while predicted AP values are shown in "blue". The "red" crosses indicate the test images misclassified. AP values have been scaled in the range [0,100]

no impact on detection accuracy in same-field experiments. In cross-field experiments, it improved detection accuracy from 93.91 to 94.59% when training on the Becker field and testing on Waseca. When training on Waseca and testing on Becker, the detection accuracy remained unchanged; however, the NMSE was reduced by more than 50%, as previously noted.

5.5 Abnormality quantification in domain-adapted whole images

In the previous section, we examined the effectiveness of domain adaptation for data augmentation in detecting and quantifying abnormalities at the whole-image level. Here, we shift our focus to Stage 2 of the proposed approach, as illustrated in Fig. 1. The objective of this stage is to further enhance generalization performance by aligning test images with the same shared domain D as the training data. Accordingly, all experiments presented in this section are conducted under cross-domain scenarios, where the training and test domains are different. Moreover, domain adaptation-based data augmentation is applied across all experiments reported in this section.

To assess the quality of domain adaptation on the test images, we computed histograms of the reconstruction loss, as shown in Fig. 9. When comparing these histograms with those for the training images in Fig. 4, it is evident that the domain-adapted test images exhibit lower quality than the domain-adapted training images. This discrepancy impacts the accuracy of abnormality quantification, as discussed later in this section, and highlights the importance of further

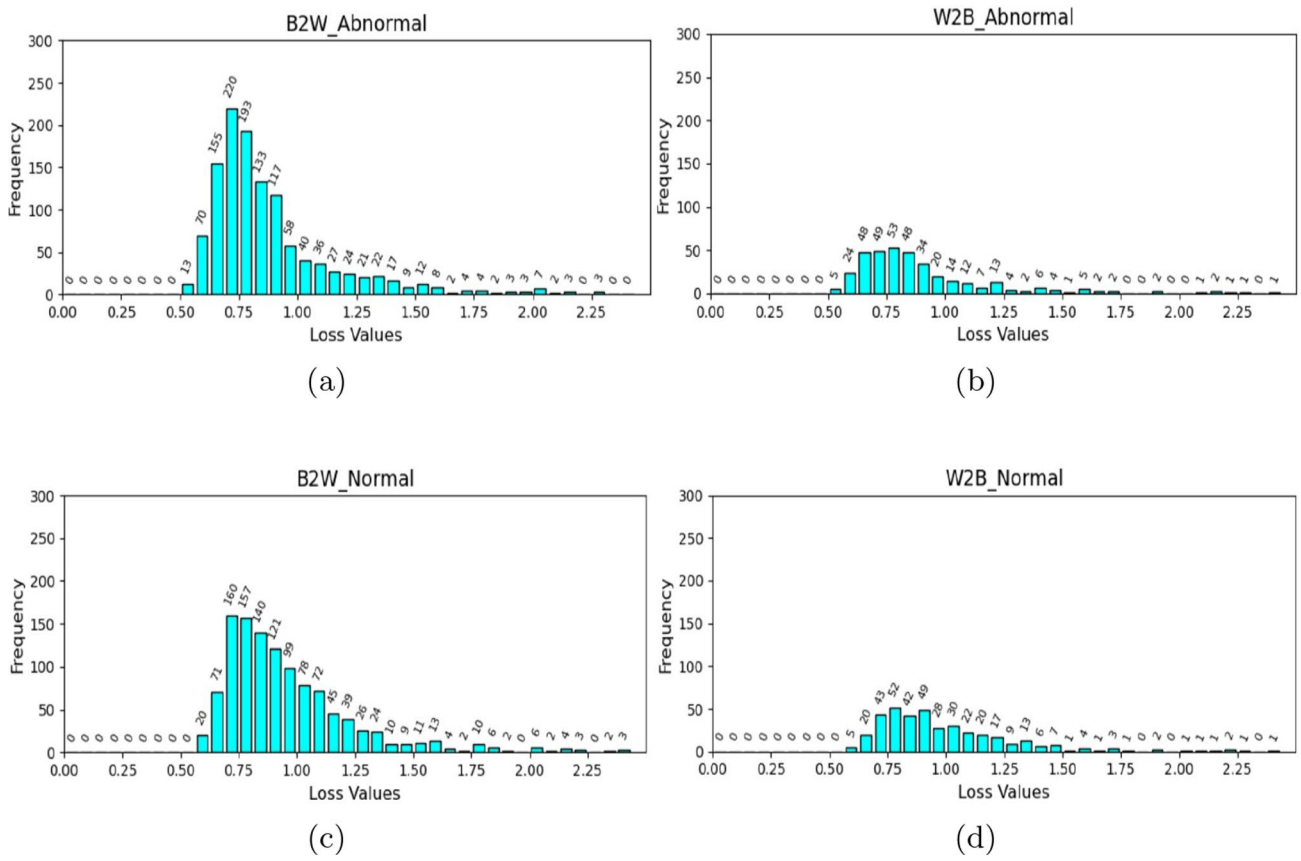


Fig. 9 Histogram of the CycleGAN reconstruction loss values on the test dataset in the case of the abnormal class for **a** B2W **b** W2B and normal class for **c** B2W and **d** W2B

improving the generalization performance of the domain adaptation process.

Figure 10b and d present the predicted AP values for domain-adapted test images, along with the corresponding ground truth values, for the B2W and W2B transformations. For ease of comparison, the predictions using the original, non-adapted test images are also shown in Fig. 10b and d. As can be observed, the model tends to overestimate the level of abnormality in the Waseca field when using the W2B data and underestimate the level of abnormality in the Becker field when using the B2W data. Once again, the difference in scale between the Becker and Waseca fields should be taken into account when interpreting these results.

Table 5 compares the detection and quantification performance between the original and domain-adapted test images. As observed, quantification performance slightly degrades when adapting the test images to the common domain D. Specifically, when testing on the Waseca field, the NMSE increased from 0.0001 to 0.001; for the Becker field, it increased from 0.005 to 0.017. We attribute this degradation to the lower quality of the domain-adapted test images. Although we attempted to mitigate this issue by removing poorly adapted test image patches using selective adaptation, the resulting AP values were unreliable due to the removal of, on average, nearly 10% of patches per test image. Evaluating the model on the complete test set, including challenging and poorly adapted samples, provides a more realistic assessment of the framework's robustness and generalization performance.

On the other hand, detection accuracy improved, increasing from 94.59 to 100% when adapting Waseca test images to the Becker domain (the shared domain), and from 98.61 to 100% when adapting Becker test images to the Waseca domain. This suggests that detection accuracy is less sensitive to the quality of domain-adapted test images, as it relies solely on thresholding the AP scores rather than the precision of patch-level quantification.

5.6 Ablation studies

5.6.1 Abnormality quantification models

To select the downstream abnormality quantification model, we compared four representative deep learning architectures: ResNet-50, DenseNet-121, EfficientNet-B0, and ViT. Tables 6 and 7 summarize their patch-level regression

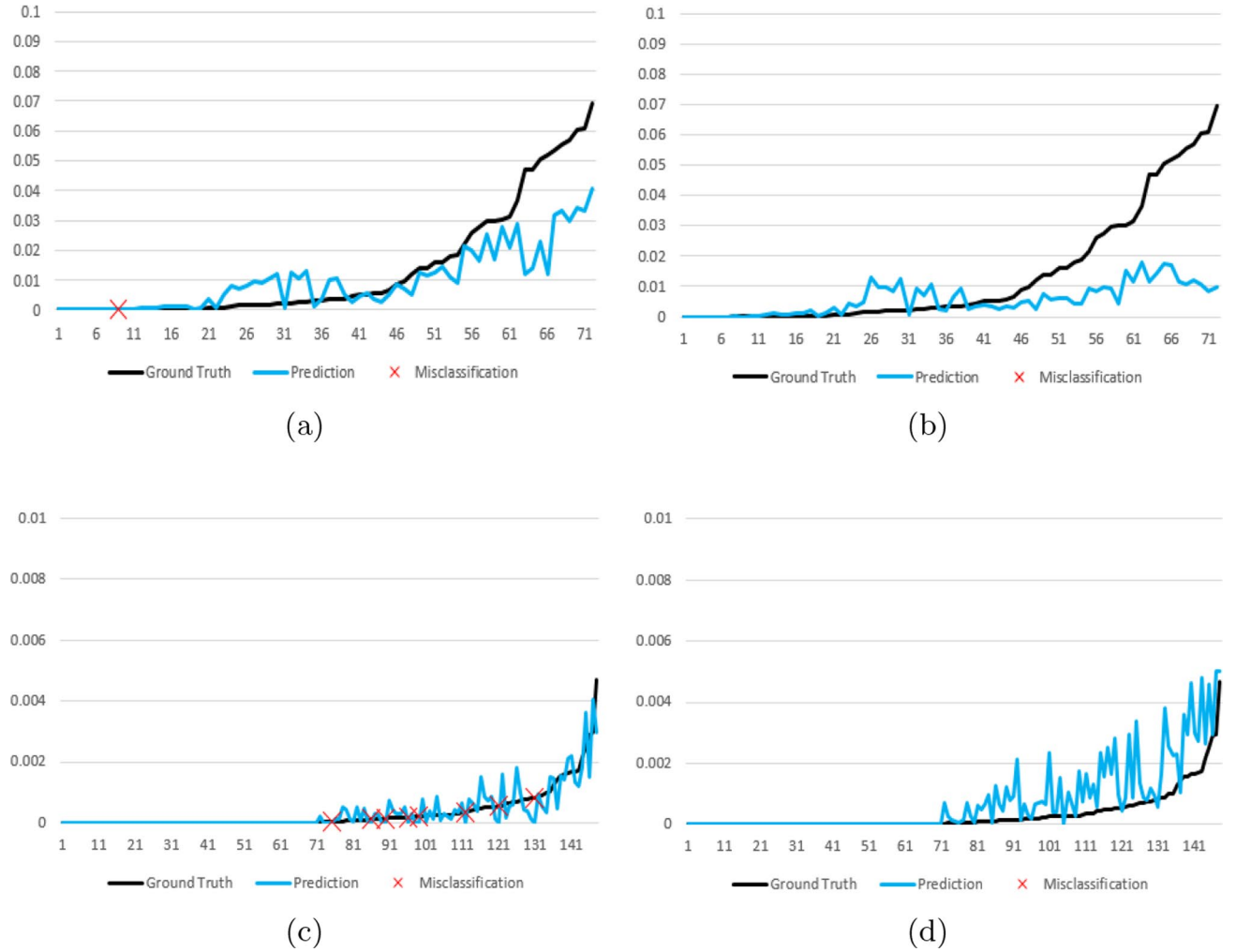


Fig. 10 Cross-field, AP-based domain-adapted whole-image predictions: **a** trained with Waseca+B2W data and tested on Becker data, **b** trained with Waseca+B2W data and tested on B2W data, **c** trained with Becker+W2B data and tested on Waseca data, and **d** trained with Becker+W2B data and tested on W2B data. Test images have been arranged in increasing order with respect to the ground truth AP values. Ground truth AP values are shown in "black" while predicted AP values are shown in "blue". The "red" crosses indicate the test images misclassified. AP values have been scaled in the range [0,100]

Table 5 Cross-field, AP-based whole-image detection and quantification performance using the domain-adapted test images

Trained on	Test Acc (%) (W2B)	Test MSE/ NMSE (W2B)	Test Acc (%) (Waseca)	Test MSE/ NMSE (Waseca)
Becker+W2B	100% (Becker)	6.9e−07/0.001 (Becker)	94.59% (B2W)	1.2e−07/ 0.0001 (B2W)
Waseca+B2W	98.61%	9.9e−05/0.005	100%	0.0003/0.017

Table 6 Patch-level abnormality quantification performance of different downstream models when trained on the Becker field and evaluated under same-field and cross-field settings

	Train MSE	Valid MSE	Test MSE (Becker)	Test MSE (Waseca)
ResNet-50	3.47	6.85	11.65	3.84
DenseNet-121	0.75	5.88	10.75	3.60
EfficientNet-B0	0.65	6.38	9.92	3.23
ViT	7.80	8.28	**9.84**	**2.58**

Table 7 Patch-level abnormality quantification performance of different downstream models when trained on the Waseca field and evaluated under same-field and cross-field settings

	Train MSE	Valid MSE	Test MSE (Waseca)	Test MSE (Becker)
ResNet-50	12.85	14.17	10.02	18.15
DenseNet-121	11.40	12.96	9.91	17.78
EfficientNet-B0	11.76	10.93	9.77	17.25
ViT	11.08	10.49	**8.32**	**17.46**

performance when trained on the Becker and Waseca fields, respectively.

When trained on the Becker field, ViT achieved the lowest test MSE on both the Becker and Waseca test sets, with scores of 9.84 and 2.58, respectively. Although CNN-based models such as DenseNet-121 and EfficientNet-B0 achieved lower training and validation errors, their higher test errors

indicate weaker generalization compared with ViT. This suggests that ViT was less prone to overfitting and provided better same-field and cross-field performance in this setting.

A similar trend was observed when training on the Waseca field. ViT achieved the lowest validation MSE and the lowest same-field test MSE on Waseca, with values of 10.49 and 8.32, respectively. EfficientNet-B0 achieved the lowest cross-field test MSE on Becker; however, its advantage over ViT was small. Overall, ViT provided the best balance across validation, same-field testing, and cross-field testing. Based on these results, we selected ViT as the default downstream abnormality quantification model in the proposed framework.

5.6.2 Selective domain adaptation

Since our framework utilizes domain adaptation for data augmentation (Stage 1), it is crucial to identify and remove poorly domain-adapted images from the training set of the detection and quantification models. At the same time, it is also equally important to exclude poorly domain-adapted test images (Stage 2) to preserve generalization performance. To achieve this, a reliable metric is needed to assess the quality of domain-adapted images. Among different metrics we considered (see below), the Peak Signal-to-Noise Ratio (PSNR) between the source and domain-adapted images has demonstrated satisfactory performance in this regard. Specifically, we need a metric that quantitatively evaluates how well an adapted image aligns with the target domain at the pixel level. PSNR is well-suited for this purpose, as it not only provides a measure of similarity but also offers insights into image quality. A higher PSNR score indicates that domain-adapted images closely resemble the target distribution with minimal distortions or artifacts, whereas a lower PSNR score suggests the presence of excessive artifacts and deviations.

Specifically, let us assume that there are N_S images I_i^S in the source domain, where $i = \{1, 2, \ldots, N_S\}$, and N_T images V_j in the target domain, where $j = \{1, 2, \ldots, N_T\}$. Given a source domain image I_i^S and its domain-adapted version I_i^T in the target domain, we calculate the PSNR score between I_i^T and a target domain image V_j as follows:

$$\mathrm{PSNR}_{i,j} = c \cdot log_{10}\left(\frac{L}{\sqrt{\mathrm{MSE}(I_i^T, V_j)}}\right)$$

where L represents the maximum possible pixel value ($L = 1$ in our experiments, as both domain-adapted and target images are normalized to the range [0,1]), c is a normalizing constant ($c = 20$ was used in our experiments), and MSE represents the Mean-Squared Error between I_i^T and V_j. Given a domain-adapted image I_i^T, we compute the $\mathrm{PSNR}_{i,j}$ score between I_i^T and each image V_j in the target domain, where $j = \{1, 2, \ldots, N_T\}$. Repeating this process for each domain-adapted image I_i^T, we then compute the average PSNR score μ_i for each I_i^T, $i = \{1, 2, \ldots, N_S\}$:

$$\mu_i = \frac{1}{N_T}\Sigma_{j=0}^{N_T}(\mathrm{PSNR}_{i,j})$$

In addition, we compute the overall average PSNR score, denoted as μ, and the standard deviation, σ, using all $N_S \times N_T \mathrm{PSNR}_{i,j}$ scores obtained in the previous step. A domain-adapted image I_i^T is considered to be poorly adapted if its individual average PSNR score μ_i fails below the threshold $\mu - c * \sigma$. Since a higher PSNR score indicates better translation quality from the source to the target domain, this process helps to identify images with excessive distortions or artifacts. In our experiments, we set $c = 1$. We refer to this process as *selective domain adaptation*, which provides performance improvements, as demonstrated in the following paragraph and further detailed in Sect. 5.3. Figure 11 shows some representative examples of poorly domain-adapted images that were effectively removed from the training set using the "selective" domain adaptation. These examples illustrate several issues when transforming images from one field to the other, such as hallucinated green leaves, soil-like appearances of leaves, and instances where abnormalities were lost entirely during the conversion process.

To evaluate the effectiveness of selective domain adaptation for data augmentation, we conducted abnormality quantification experiments using the ViT_{rg} patch-level model. We considered two experimental setups: (i) training on B+W2B image patches and testing on Becker image patches, and (ii) training on W+B2W image patches and testing on Waseca image patches. Training and testing were performed using the randomly extracted image patches as described in Sect. 4.

For comparison, these experiments were also repeated using non-selective domain adaptation. Quantification performance was assessed using Mean Squared Error (MSE). To explore the potential for filtering out low-quality domain-adapted images, we evaluated several visual similarity metrics, including Peak Signal-to-Noise Ratio (PSNR), Wasserstein distance, and Learned Perceptual Image Patch Similarity (LPIPS). As shown in Table 8, all selective domain adaptation methods outperformed the non-selective approach across both fields. Among the metrics evaluated, PSNR demonstrated the highest effectiveness, filtering out approximately 7.2% of low-quality images from W2B adaptations and 9.5% from B2W adaptations.

(a) (b)

Fig. 11 Representative examples of poorly domain-adapted image patches identified using the PSNR methodology for **a** the B2W transformation and **b** the W2B transformation. The top row shows the original image patches, while the bottom row presents their corresponding domain-adapted versions

Table 8 Performance comparison of different methods for filtering out poorly domain-adapted image patches

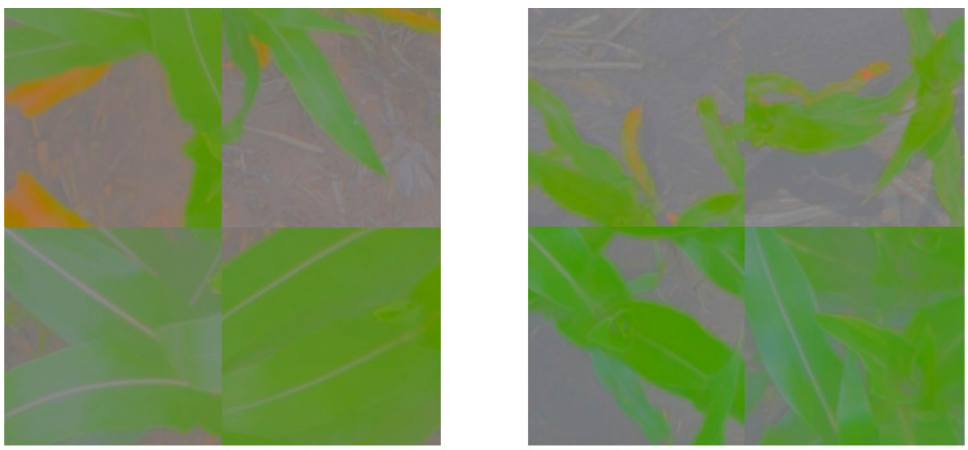
	Trained on Becker+W2B, tested on Becker		Trained on Waseca+B2W, tested on Waseca	
	Selective	Non-selective	Selective	Non-selective
Wasserstein	8.07	8.17	2.02	2.06
LPIPS	8.07		1.96	
PSNR	8.00		1.97	

"Selective" refers to training the ViT_{rg} model on original and domain-adapted image patches, while excluding image patches identified as poorly adapted. "Non-selective" training incorporates original and all domain-adapted image patches, without filtering. Quantification performance is assessed using MSE score

5.6.3 Domain adaptation using histogram matching

Motivated by the work of Pandey et al. [50], we compared the proposed domain-adaptation approach based on CycleGANs to an alternative domain-adaptation approach based on histogram matching. Histogram matching (HM) (or histogram specification), is a widely used technique in image processing to transform an image to match a specified histogram[64]. In the context of our application, we can transform images from different fields to match a specified histogram, which can be considered as a way to map data from different fields to a shared domain D.

We performed two experiments as before: B2W (i.e., D = W) and W2B (i.e., D = B). In the B2W experiment, we first computed the histogram of an "average" Waseca image computed by averaging the randomly extracted image patches from the Waseca field as described in Sect. 4. We then transformed the randomly extracted image patches from the Becker to match the "average" histogram from the Waseca field. In the W2B experiment, we transformed the randomly extracted image patches from the Waseca to match the "average" histogram from the Becker field. Following the methodology described in [50], the RGB image patches were converted to the LAB color space and then HM was performed only on the L channel. The results were then converted back to the RGB color space. The augmented images were incorporated alongside the original images to train the model. Figure 12 shows some representative domain-adaptation examples using HM.



(a) (b)

Fig. 12 Representative randomly extracted patches for **a** Becker field images are histogram matched based on Waseca field images, and **b** Waseca field images are histogram matched based on Becker field images. Top row represents sample abnormal images while the bottom row provides visualization for normal images

Table 9 Same-field patch-level quantification performance comparison between HM-based and CycleGAN-based domain adaptation

Trained on	Train MSE	Valid MSE	Test MSE
Becker+W2B (HM)	5.53	4.53	8.24
Becker+W2B (CycleGAN)	6.68	5.87	**8.00**
Waseca+B2W (HM)	8.94	9.15	2.48
Waseca+B2W (CycleGAN)	11.11	9.80	**1.97**

We used the randomly extracted patches from both the Becker and Waseca fields as training and test data. Test images were neither histogram-matched nor adapted using the CycleGAN model and were used solely to evaluate the model's performance within the same field. Table 9 compares the two different domain adaptation approaches, the first using histogram specification and the second using CycleGANs. The results show that CycleGAN-based domain adaptation outperforms the HM-based domain adaptation both on the Becker and Waseca fields.

6 Conclusions

In this study, we proposed a DL framework for detecting and quantifying abnormalities in maize fields using UAV-collected images. Our objective was to develop a general-purpose framework capable of effectively handling varying field conditions, including different growth stages, soil types, and lighting conditions. To address the challenges of limited training data for learning large DL models and ensuring generalization to new data, we leveraged domain adaptation. Specifically, we introduced a methodology to transform data from different fields into a shared domain, thereby enhancing the training process and improving generalization performance when analyzing new data. Additionally, we proposed a selective domain adaptation technique to filter out poorly adapted data, ensuring the quality of both training and evaluation phases. We have demonstrated the feasibility of the proposed approach through extensive experiments and comparisons on two different fields containing diverse conditions. Our framework can provide valuable insights to farmers by identifying field areas requiring closer inspection. Additionally, it facilitates data annotation by guiding annotators toward specific regions rather than the entire field, thereby improving the accuracy and efficiency of image labeling [18].

This work can be improved in several ways. *First*, we were unable to identify a more extensive dataset during this study that included data from multiple fields, which would have enabled a more comprehensive evaluation and exploration of additional research questions. Nevertheless, the dataset used in our experiments captures diverse field conditions, including varying growth stages, soil types, and lighting conditions. Despite the absence of multi-field data, extensive experiments have demonstrated the feasibility of the proposed framework. Future work will focus on validating our approach with larger, more diverse datasets to further support our findings. Addressing more challenging domain variations, such as differences in resolution, will also enhance the framework's applicability.

Second, while the *ViT* models used for regression performed well, alternative deep learning models may offer further improvements. Additionally, incorporating more advanced transfer learning techniques, such as self-supervised pretraining [44], is expected to enhance performance.

Third, using a separate domain adaptation model for each domain pair is computationally inefficient and complicates model selection when transforming novel data into the shared domain D. A more scalable solution could involve developing a single domain adaptation model capable of handling transformations across multiple source domains, for example, using models that can handle domain imbalances [60] or by building an "inverse" StarGAN model [65]. Further research is also needed to determine the optimal shared domain D, potentially by leveraging insights into the "difficulty" of adapting images between domains.

Fourth, exploring more advanced domain adaptation models, such as CycleGAN-Turbo [66] and CUT [48], is a promising direction. This is especially important in Stage 2 of our approch fr improving the generaliztion performance of domain adaptation.

Fifth, preserving semantic information during adaptation requires further attention. While patch-based domain adaptation methods, like ours, have demonstrated that they preserve semantic information [48, 56], incorporating task-specific semantic constraints [13] into the loss function could further improve structural and contextual integrity. Refining our methodology for identifying poorly domain-adapted images is also a key area for future research. We believe that a more effective method to identify poorly domain-adapted image patches will significantly benefit both Stages of the proposed methodology.

Finally, we are interested in extending our framework beyond agriculture, particularly to medical imaging, where domain adaptation could significantly improve model generalization across different imaging sources and conditions.

Acknowledgements This work was supported by the National Institute of Food and Agriculture/USDA, Award No. 2020-67021-30754.

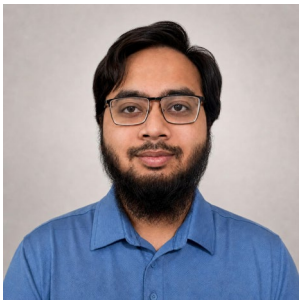
Author contributions A.H. conducted experiments, wrote manuscript, prepared visualizations D.Z. supervised and guided the research work, provided data G.B. supervised and guided the research work, wrote manuscript All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

References

1. FAO-UN: Plant production and protection. Food and Agriculture Organization of the United Nations (2024). https://www.fao.org/plant-production-protection/about/en
2. USDA: USDA forecasts us corn production down, soybean and cotton production up from 2023. Natl. Agricultural Stat. Service (2024). https://www.nass.usda.gov/Newsroom/2024/09-12-2024.php
3. Asseng, S., Asche, F.: Future farms without farmers. Sci. Robot. **4**, eaaw1875 (2019)
4. Pretto, A., et al.: Building an aerial–ground robotics system for precision farming. IEEE Robot. Autom. Mag. **28**(3), 29–49 (2020)
5. Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: a survey. Comput. Electron. Agric. **147**, 70–90 (2018)
6. Zhuang, F., et al.: A comprehensive survey on transfer learning. Proc. IEEE **109**, 43–76 (2020)
7. Xu, M., Yoon, S., Fuentes, A., Park, D.S.: A comprehensive survey of image augmentation techniques for deep learning. Pattern Recogn. **137**, 109347 (2023)
8. Liu, X., et al.: Deep unsupervised domain adaptation: a review of recent advances and perspectives. APSIPA Trans. Signal Inform. Process. **1**, 1–48 (2022)
9. Fei, Z., Olenskyj, A.G., Bailey, B.N., Earles, M.: Enlisting 3d crop models and gans for more data efficient and generalizable fruit detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),1269–1277 (2021)
10. Cieslak, M.: Generating diverse agricultural data for vision-based farming applications. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 5422–5431. CVPRW (2024)
11. Gogoll, D., Lottes, P., Weyler, J., Petrinic, N., Stachniss, C.: Unsupervised domain adaptation for transferring plant classification systems to new field environments, crops, and robots. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2636–2642 (2020)
12. Magistri, F., et al.: From one field to another—unsupervised domain adaptation for semantic segmentation in agricultural robotics. Comput. Electron. Agric. **212**, 108114 (2023)
13. Hoffman, J., et al.: CyCADA: Cycle-consistent adversarial domain adaptation. International Conference on Machine Learning (2017)
14. Chen, Y.-C., Lin, Y.-Y., Yang, M.-H., Huang, J.-B.: CrDoCo: Pixel-level domain transfer with cross-domain consistency. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1791–1800 (2019)
15. Bertoglio, R., Mazzucchelli, A., Catalano, N., Matteucci, M.: A comparative study of Fourier transform and cyclegan as domain adaptation techniques for weed segmentation. Smart Agric. Technol. **4**, 100188 (2023)
16. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2242–2251 (2017)
17. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: Transformers for image recognition at scale. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, 3–7 (2021)
18. Zermas, D., et al.: A methodology for the detection of nitrogen deficiency in corn fields using high-resolution RGB imagery. IEEE Trans. Autom. Sci. Eng. **18**, 1879–1891 (2020)
19. Huq, A., Zermas, D., Bebis, G.: Abnormality detection in maize fields using selective domain adaptation–driven data augmentation. Advances in Visual Computing—20th International Symposium, ISVC 2025, Las Vegas, NV, USA, November 17–19, 2025, Proceedings, Part I **16396**, 256–269 (2025)
20. Padilla, F.M., Gallardo, M., Peña-Fleitas, M.T., De Souza, R., Thompson, R.B.: Proximal optical sensors for nitrogen management of vegetable crops: a review. Sensors **18**, 2083 (2018)
21. Chore, A., Thankachan, D.: Nutrient defect detection in plant leaf imaging analysis using incremental learning approach with multifrequency visible light approach. J. Electrical Eng. Tech. **18**(2), 1369–1387 (2023)
22. Ushaa, K., Singh, B.: Potential applications of remote sensing in horticulture—a review. Sci. Hortic. **153**, 71–83 (2013)
23. Pandey, P., Ge, Y., Stoerger, V., Schnable, J.C.: High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. Front. Plant Sci. **8**, 1348 (2017)
24. Yuan, Y.: Diagnosis of nitrogen nutrition of rice based on image processing of visible light. In: IEEE International conference on functional-structural plant growth modeling, simulation, visualization and applications, 228–232 (2016)
25. Barbedo, J.G.A.: Digital image processing techniques for detecting, quantifying and classifying plant diseases. Springerplus **2**, 660 (2013)
26. Rahadiyan, D.: Classification of chili plant condition based on color and texture features. In: 7th International Conference on Informatics and Computing, 1–7 (2022)
27. Sosa, J., Ramírez, J., Vives, L., Kemper, G.: An algorithm for detection of nutritional deficiencies from digital images of coffee leaves based on descriptors and neural networks. XXII symposium on image, signal processing and artificial vision (STSIVA), 1–5 (2019)
28. Tan, L., Lu, J., Jiang, H.: Tomato leaf diseases classification based on leaf images: a comparison between classical machine learning and deep learning methods. AgriEngineering **3**, 542–558 (2021)
29. Yi, J., et al.: Deep learning for non-invasive diagnosis of nutrient deficiencies in sugar beet using RGB images. Sensors **20**, 5893 (2020)
30. Tejasri, N.: Drought stress segmentation on drone captured maize using ensemble u-net framework. In: IEEE International conference on functional-structural plant growth modeling, simulation, visualization and applications, 1–6 (2022)
31. Cardim Ferreira Lima, M.: Monitoring plant status and fertilization strategy through multispectral images. Sensors **20**, 435 (2020)
32. Debnath, S., et al.: Identifying individual nutrient deficiencies of grapevine leaves using hyperspectral imaging. Remote Sensing **13**, 3317 (2021)
33. Bechar, A., Vigneault, C.: Agricultural robots for field operations: concepts and components. Biosys. Eng. **149**, 94–111 (2016)
34. Song, H., Yuan, Y., Ouyang, Z., Yang, Y., Xiang, H.: Quantitative regularization in robust vision transformer for remote sensing image classification. Photogram. Rec. **39**, 340–372 (2024)
35. Song, H., et al.: Cmkd-net: a cross-modal knowledge distillation method for remote sensing image classification. Adv. Space Res. **75**, 8515–8534 (2025)
36. Song, H., et al.: Symmetrical learning and transferring: efficient knowledge distillation for remote sensing image classification. Symmetry **17**, 1002 (2025)
37. Song, H., Liang, L.: Causal-yolo: enhancing object detection in remote sensing imagery through causal feature alignment. Opt. Laser Technol. **200**, 115167 (2026)
38. Wang, Y.: Weed mapping with convolutional neural networks on high resolution whole-field images. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 505–514 (2023)
39. Talukder, M.S.H., Sarkar, A.K.: Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning. Smart Agric. Technol. **4**, 100155 (2023)

40. Guerrero, R., Renteros, B., Castaneda, R., Villanueva, A., Belupú, I.: Detection of nutrient deficiencies in banana plants using deep learning. 2021 IEEE International Conference on Automation/XXIV Congress of the Chilean Association of Automatic Control (ICA-ACCA), 1–7 (2021)
41. Zhang, X., Qiao, Y., Meng, F., Fan, C., Zhang, M.: Identification of maize leaf diseases using improved deep convolutional neural networks. IEEE Access **6**, 66 (2018)
42. Bansal, P., Kumar, R., Kumar, S.: Disease detection in apple leaves using deep convolutional neural network. Sensors **11**, 617 (2021)
43. Chiu, M.T.: Agriculture-vision: A large aerial image database for agricultural pattern analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2825–2835 (2020)
44. Wen, Y., Chen, L., Deng, Y., Zhou, C.: Rethinking pre-training on medical imaging. J. Vis. Commun. Image Represent. **78**, 103145 (2021)
45. Wood, E.: Fake it till you make it: Face analysis in the wild using synthetic data alone. In: Proceedings of the IEEE/CVF international conference on computer vision, 3681–3691 (2021)
46. Anderson, J.W., Ziolkowski, M., Kennedy, K., Apon, A.W.: Synthetic image data for deep learning. arXiv preprint arXiv:2212.06232 (2022)
47. Nikolenko, S.I.: Synthetic Data for Deep Learning. Springer, Berlin (2021)
48. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. European conference on computer vision, 319–345 (2020)
49. Lottes, P., Behley, J., Chebrolu, N., Milioto, A., Stachniss, C.: Joint stem detection and crop-weed classification for plant-specific treatment in precision farming. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 8233–8238 (2018)
50. Pandey, P., Best, N.B., Washburn, J. D.: Synthetically labeled images for maize plant detection in UAS images. International Symposium on Visual Computing, 543–556 (2023)
51. Gao, Y., Zhang, M., Li, W., Tao, R.: Distribution-independent domain generalization for multisource remote sensing classification. IEEE Trans. Neural Netw. Learn. Syst. **36**, 13333–13344 (2025)
52. Song, X., Gao, Y., Zhang, M., Li, W., Tao, R.: Wavehfg: high-frequency guidance for heterogeneous remote sensing image change detection with wavelet features. IEEE Trans. Geosci. Remote Sens. **64**, 1–14 (2026)
53. Giuffrida, M.V., Dobrescu, A., Doerner, P., Tsaftaris, S.A.: Leaf counting without annotations using adversarial unsupervised domain adaptation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2590–2599 (2019)
54. Bellocchio, E., Costante, G., Cascianelli, S., Fravolini, M.L., Valigi, P.: Combining domain adaptation and spatial consistency for unseen fruits counting: a quasi-unsupervised approach. IEEE Robot. Autom. Lett. **5**(2), 1079–1086 (2020)
55. Marino, S., Beauseroy, P., Smolarz, A.: Unsupervised adversarial deep domain adaptation method for potato defects classification. Comput. Electron. Agric. **174**, 105501 (2020)
56. Imbusch, B.T., Schwarz, M., Behnke, S.: Synthetic-to-real domain adaptation using contrastive unpaired translation. 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), 595–602 (2022)
57. Guo, T., Huynh, C.P., Solh, M.: Domain-adaptive pedestrian detection in thermal images. 2019 IEEE international conference on image processing (ICIP), 1660–1664 (2019)
58. Zhang, L., Gonzalez-Garcia, A., van de Weijer, J., Danelljan, M., Khan, F.S.: Synthetic data generation for end-to-end thermal infrared tracking. IEEE Trans. Image Process. **28**(4), 1837–1850 (2019)
59. Sermanet, P., et al.: Overfeat: Integrated recognition, localization and detection using convolutional networks. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014)
60. Patashnik, O., Danon, D., Zhang, H., Cohen-Or, D.: Balagan: cross-modal image translation between imbalanced domains. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2659–2667 (2021)
61. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2024). https://github.com/HumanSignal/label-studio. Open source software available from
62. Hughes, D., Salathé, M.: An open access repository of images on plant health to enable the development of mobile disease diagnostics (2015). arXiv:1511.08060 arXiv preprint
63. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 66 (2008)
64. Gonzalez, R.C., Woods, R.E.: Digital image processing 4th edition, global edition (2018)
65. Choi, Y., et al.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 8789–8797 (2018)
66. Parmar, G., Park, T., Narasimhan, S., Zhu, J.-Y.: One-step image translation with text-to-image models. arXiv:2403.12036 (2024)



Aminul Huq is currently pursing a Ph.D. degree in Computer Science and Engineering at the University of Nevada, Reno under the supervision of Dr. George Bebis. He obtained his M.Sc. from Tsinghua University, China and B.Sc. from Rajshahi University of Engineering & Technology, Bangladesh in 2019 and 2017 respectively in Computer Science and Engineering. His research interest lies in the field of computer vision and medical image analysis.

Dimitris Zermas, Ph.D. is a Machine Learning Manager at John Deere, where he leads the development of advanced computer vision and deep learning solutions for precision agriculture. Previously, he served as Director of Machine Learning at Sentera, where he established the company's ML capabilities and helped drive innovation in imagery-based analytics. Dr. Zermas received his Ph.D. in Computer Science and Engineering from the University of Minnesota, with research focused on integrating machine learning and computer vision for agricultural applications. His work spans areas such as 3D point cloud segmentation, UAV-based crop monitoring, and scalable infrastructure for model training and deployment. He is passionate about bridging academic research with real-world impact in agriculture through large-scale, intelligent systems.

George Bebis, Ph.D. received the B.S. degree in Mathematics and the M.S. degree in Computer Science from the University of Crete, Greece, in 1987 and 1991, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Central Florida, Orlando, in 1996. Currently, he is a Foundation Professor in the Department of Computer Science and Engineering (CSE) at the University of Nevada, Reno (UNR) and Director of the Computer Vision Laboratory (CVL). From 2013 to 2018, he served as Department Chair of CSE-UNR. His research interests include computer vision, image processing, pattern recognition, machine learning, and evolutionary computing. His research has been funded by NSF, NASA, ONR, NIJ, and Ford Motor Company. Dr. Bebis is an Associate Editor of the Machine Vision and Applications Journal and serves on the editorial board of the International Journal on Artificial Intelligence Tools, and the Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization journal. He has served on the program committees of various national and international conferences and is the founder/main organizer of the International Symposium on Visual Computing (ISVC) and the International Symposium on Mathematical and Computational Oncology (ISMCO).