Label-free Anomaly Detection in Aerial Agricultural Images with Masked Image Modeling

Sambal Shikhar Anupam Sobti
Plaksha University
Mohali, Punjab, India
sambal.shikhar@plaksha.edu.in
anupam.sobti@plaksha.edu.in

Abstract

Detecting various types of stresses (nutritional, water, nitrogen, etc.) in agricultural fields is critical for farmers to ensure maximum productivity. However, stresses show up in different shapes and sizes across different crop types and varieties. Hence, this is posed as an anomaly detection task in agricultural images. Accurate anomaly detection in agricultural UAV images is vital for early identification of field irregularities. Traditional supervised learning faces challenges in adapting to diverse anomalies, necessitating extensive annotated data. In this work, we overcome this limitation with self-supervised learning using a masked image modeling approach. Masked Autoencoders (MAE) extract meaningful normal features from unlabeled image samples which produces high reconstruction error for the abnormal pixels during reconstruction. To remove the need of using only "normal" data while training, we use an anomaly suppression loss mechanism that effectively minimizes the reconstruction of anomalous pixels and allows the model to learn anomalous areas without explicitly separating "normal" images for training. Evaluation on the Agriculture-Vision data challenge shows a ***6.3%*** *mIOU score improvement in comparison to prior state of the art in unsupervised and self-supervised methods. A single model generalizes across all the anomaly categories in the Agri-Vision Challenge Dataset [5].*

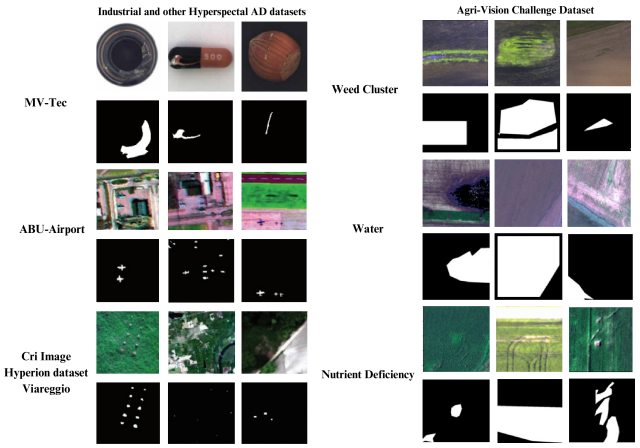


Figure 1. Comparison of anomaly datasets: The left column represents a variety of industrial and other hyperspectral anomaly detection (AD) datasets, including MV-Tec, ABU-Airport, and Cri Image Hyperion dataset of Viareggio. The right column displays the Agri-Vision Challenge Dataset, highlighting agricultural anomalies such as Weed Clusters, Water stress, and Nutrient Deficiency. This illustrates the complexity of agricultural anomalies, showcasing their large inter-class and intra-class variations and their occurrence at multiple scales, as opposed to more uniform and scale-consistent anomalies found in industrial datasets.

1. Introduction

In precision agriculture, Unmanned Aerial Vehicles (UAVs) have emerged as a pivotal tool for monitoring agricultural landscapes efficiently. UAVs provide much higher resolution images compared to satellite images, thus capturing fine grained details on the agricultural fields. Accurate anomaly detection in UAV images is crucial for the early identification of potential issues such as pest infestations, diseases, and nutrient deficiencies. The dynamic and diverse nature of agricultural fields further compounds the challenge, as anomalies can vary greatly in appearance due to factors such as crop type, growth stage, and environmental conditions as compared to other anomaly detection settings, compared in Figure 1. Thus, there is a need for a completely label free approach to training anomaly detection models so that it can be applied across different crops for different kinds of anomalies. Traditionally, *supervised learning* methods have been used for anomaly detec-

tion systems [2, 4, 11, 21]. These methods are inherently limited by their dependence on large sets of annotated data, which are labor-intensive to create and may not capture the full spectrum of possible anomalies. Even in case of *unsupervised* and *self supervised* methods [6, 12, 15, 29, 30] where explicit anomaly labels are not used, there is a dependence on using only "normal" data for training thus making it necessary for a user to curate normal data that does not contain any types of anomalies.

To leverage self-supervised learning through masked image modeling, we utilize Masked Auto-encoders (MAE) [9] to effectively learn normal features from unlabeled image samples. This "normality" then facilitates the detection of anomalies through higher reconstruction errors for patches containing anomalies. Incorporating a Swin Transformer-based Masked Autoencoder [7] enables our model to learn both local and global features, ensuring robust detection across a wide range of anomaly types. Our work introduces an approach that also learns to detect anomalies with abnormal samples within the training data. This inclusion allows users to simplify their data collection pipeline by removing the need to curate "normal" data. This enables the identification of a wide array of anomalies without training multiple models for detecting different anomalies (by removing anomalies of a particular type from the data). The effectiveness of our approach is demonstrated through extensive evaluation on the Agriculture-Vision Challenge dataset [5], showcasing significant improvement of 6.3% in mean Intersection over Union (mIOU) score and generalization across all the given anomaly classes.

2. Related Work

The Agriculture-Vision dataset[5] provides multispectral aerial images for fields at 10cm/pix resolution along with annotations for anomalies of 9 types - drydown, planet skip, water, weed cluster, nutrient deficiency, endrow, double plant, waterway, and, storm damage. Unlike other datasets that may include hyperspectral and multispectral data for general land cover classification [10] or crop type identification [23], Agriculture-Vision specifically targets the semantic segmentation of agricultural patterns for recognizing various field anomaly patterns crucial to farmers.

Supervised image segmentation approaches like FusePN [11] and AAFormer [21] have demonstrated competitive performance on detecting anomalies in UAV images. AAFormer uses a transformer based architecture. FusePN[11] uses a multimodal fusion approach fusing RGB and NIR bands of the image along in an encoder-decoder style architecture with additional modifications for inference efficiency. The limitations of supervised methods have steered research towards unsupervised and self-supervised learning approaches, where the focus shifts to learning from unlabeled data.

One-class classification (OCC) [22, 30] models provide another approach for anomaly detection utilizing high-level semantic information for anomaly identification in the feature space. OCC based anomaly detection uses high-level semantic information and distance metrics for anomaly scoring. However, they encounter challenges such as i) mode collapse and ii) overlook low-level structural features due to their focus on compact feature representation [18]. Anomaly Segmentation based on pixel Descriptors (ASD) [13] addresses anomaly segmentation in high spatial resolution (HSR) imagery by using deep one-class classification with discriminative pixel descriptors through abnormal sample generation, promoting descriptor compactness for normal data and diversity to prevent model collapse. ASD employs a multi-level, multi-scale feature extraction approach to capture low-level and semantic information.

Reconstruction-based anomaly detection methods like Attribute Restornation Network (ARnet) [28] and Deep Feature Reconstruction (DFR) [26] utilize autoencoders (AE) with an encoder-decoder architecture to capture the manifold of defect-free images to differentiate between normal and anomalous data based on reconstruction fidelity. These models, trained solely on normal imagery, are expected to yield higher reconstruction errors for anomalous inputs, using metrics such as mean square error (MSE) for anomaly quantification. Techniques to enhance anomaly detection capabilities include image degradation and subsequent restoration, notably through inpainting methods like RIAD [29], which mask parts of the image to challenge the model's reconstruction abilities. Despite their success, these methods struggle, as with progression in training it inevitably involves the anomalies in the reconstructed image. This is because models favor learning all the information from input, including both background and anomalies simultaneously. In terms of architectural elements, convolutions are prone to learn identity mapping (from input image to output image) as their receptive fields are biased towards learning local spatial features [14]. To address these limitations, recent approaches propose integrating Transformer based reconstruction method like IntRA (Inpainting Transformer) [17] which pose anomaly detection as a patch-inpainting problem and propose to solve it with a purely self-attention based approach discarding convolutions.Other transformer based approaches like MAE which mask $\sim$75% of the images patches and use the remaining to reconstruct the complete image. MAEDAY [20] leverages MAE for image-reconstruction-based anomaly detection method that utilizes a pre-trained model, enabling its use for Few-Shot Anomaly Detection (FSAD). We provide a class-wise comparison of these methods in Table 1.

3. Background

Objective - The anomaly segmentation of a high spatial resolution (HSR) image **X** with dimensions **H×W×B** (height, width, and Number of multi-spectral bands) is defined by a mapping function **f**, transforming **X** into an anomaly map **A** with dimensions **H×W** .

3.1. Masked Auto-encoder

A Masked Autoencoder (MAE) proposed by he et al. [9] learns image representations through self-supervised learning and masked image modelling where a Transformer based model learns to reconstruct an input image by reconstructing an input from a partially masked version of itself. We use masked image modelling framework and MAE for anomaly detection as given in Figure 2 to learn the background or normal patterns of an input image by reconstructing it from a subset of observed pixels. This process enables the MAE to identify deviations from the learned data distribution, which are indicative of anomalies. A Masked Auto-encoder (MAE) leverages Vision Transformer (ViT) [8] for both its encoder E and decoder D components. The ViT segments the input image into patches and applies self-attention mechanisms to capture complex and global features.

Consider an input image X with dimensions $H \times W \times B$, where H, W, and B denote the height, width, and number of bands, respectively. A binary mask M is applied to generate the masked image X_{masked}:

$$X_{\text{masked}} = X \odot M \tag{1}$$

The encoder E processes the image by dividing it into N patches, each with a fixed size $P \times P$:

$$\text{Patches}_E = \text{Patchify}(X_{\text{masked}}) \tag{2}$$

These patches are then flattened and linearly embedded to a dimension D, followed by adding positional embeddings to retain spatial information:

$$Z_0 = [\mathbf{x}_\text{p}^1 E; \mathbf{x}_\text{p}^2 E; \ldots; \mathbf{x}_\text{p}^N E] + \mathbf{E}_{\text{pos}} \tag{3}$$

Here, $\mathbf{x}_\text{p}^i E$ denotes the embedded patches, and $\mathbf{E}_{\text{pos}}$ is the positional embedding.

The sequence of embeddings Z_0 is passed through L Transformer layers to generate the latent representation Z_L:

$$Z_L = \text{Transformer}(Z_{l-1}), \quad \text{for } l = 1 \ldots L-1 \tag{4}$$

Each Transformer layer comprises multi-headed self-attention (MSA) [24] and multi-layer perceptrons (MLP), with layer normalization (LN) applied before each module and a residual connection after each:

$$Z'_l = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1} \tag{5}$$

$$Z_l = \text{MLP}(\text{LN}(Z'_l)) + Z'_l \tag{6}$$

The decoder D, structured similarly to E, reconstructs the original image from Z_L:

$$\hat{X} = \text{Patchify}^{-1}(\text{Transformer}_D(Z_L)) \tag{7}$$

The Transformer layers in D upsample the latent representations to the original resolution.

The reconstruction error E_{recon} between X and $\hat{X}$ serves as a measure for anomaly detection:

$$E_{\text{recon}} = ||X - \hat{X}||^2 \tag{8}$$

This error is evaluated per pixel to generate an anomaly map A by thresholding the map by θ :

$$A(i,j) = E_{\text{recon}}(i,j) \geq \theta \tag{9}$$

The anomaly detection in MAE is predicated on the assumption that the model, trained predominantly on normal data, will yield higher reconstruction errors for anomalies in X due to deviations from the learned patterns. The ViT architecture's self-attention mechanism allows the MAE to capture predominantly global features.

3.2. Swin Transformers

Swin Transformers [16] efficiently handles image representation by partitioning the input image into a grid of patches, which are then processed using self-attention within local windows. The local self-attention mechanism for a patch $P_{i,j}$ is defined as:

$$\text{SA}_{\text{local}}(P_{i,j}) = \text{Softmax}\left(\frac{Q_{i,j}K_{i,j}^T}{\sqrt{d}}\right)V_{i,j} \tag{10}$$

where $Q_{i,j}$, $K_{i,j}$, and $V_{i,j}$ are the query, key, and value matrices, respectively, and d is the dimensionality of the query and key.

The Swin Transformer expands the receptive field through a novel shifting mechanism that broadens the scope of self-attention across neighboring patches:

$$\text{Shift}(W_{i,j}) = W_{i+s,j+s} \tag{11}$$

where $W_{i,j}$ represents the original window of patches, and s is the shift size.

For multi-scale representation, patches are merged to form larger patches in deeper layers, reducing the resolution while expanding the receptive field:

$$P_{i,j}^{l+1} = \text{Transform}\left([P_{2i,2j}^l, P_{2i+1,2j}^l, P_{2i,2j+1}^l, P_{2i+1,2j+1}^l]\right) \tag{12}$$

where $P_{i,j}^l$ denotes a patch at layer l, and the Transform function fuses features from four adjacent patches into a new patch at layer $l+1$.

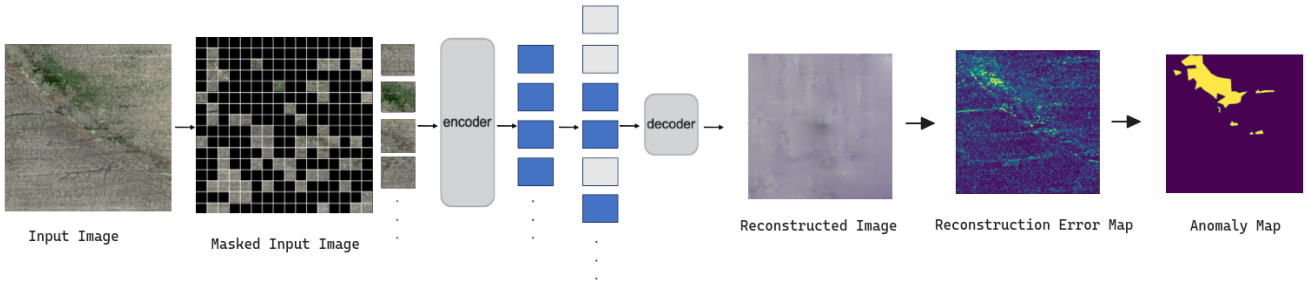


Figure 2. Input image is masked and the unmasked image patches are fed into the encoder which embeds each of those patches ,the decoder takes in embed patches along with masked patches to reconstruct the input image.A reconstruction error map is generated which is then used to generated the final Anomaly map

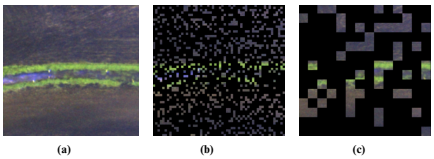


Figure 3. Comparison between masking methods. (a) original image (b) Normal random masking method (c) Window masking method.

This hierarchical approach enables Swin Transformers to capture global and local patterns and detect anomalies at multiple scales in agricultural fields to enhance their performance when compared to ViT based Masked Autoencoders for anomaly detection tasks.

3.3. SwinMAE (Swin Masked Auto-encoder)

To leverage Swin transformers ability to learn both local and global features and integrate it with masked image modelling framework, Swin Masked Auto-encoder [7] architecture replaces the Vision Transformer (ViT) typically used in MAEs with Swin Transformers. The masking strategy in Swin Masked Autoencoder involves a novel approach that maintains the number of patches in the input data during the encoding process whereas MAE only feeds unmasked patches into the encoder. Instead of removing masked patches, which could lead to a shortage of tokens necessary for subsequent processing steps like patch merging, the encoder replaces these masked tokens with a learnable vector. This method ensures a consistent number of tokens throughout the encoding process.

The Swin MAE's window masking strategy addresses the limitations of patch-based masking using MAE, particularly when using smaller patches like 4x4 used in the starting blocks of Swin Transformers as shown in 3 . This approach divides the image into larger, non-overlapping windows, each containing multiple patches, and masks these windows instead of individual patches. This method aims to prevent models from learning shortcuts, such as reconstructing masked areas through simple interpolation using neighbouring unmasked patches while also maintaining a consistent number of tokens throughout the encoder.

Swin MAE uses a light weight decoder with patch expanding layers to restore the image back to its original dimensions which is similar to Swin-Unet [3] . The decoder consists of Swin transformer blocks. Unlike MAE the masked tokens are not removed through out the encoder so there is no need add these masked tokens in the decoder input. The decoder uses a projection layer to finally restore the image back to its original dimension instead of a patch expanding layer used in Swin-Unet , just like MAE decoder.

3.4. Anomaly Suppression Loss

A Mean Square Error (MSE) based loss function allows the model to learn to reconstruct anomaly pixels as the training progresses. During the initial iterations, the reconstruction error for anomalies is higher than for the background pixels, but as more iterations are completed, reconstruction-based models are able to reconstruct the anomaly pixels. Wang et al. [25] introduced an adaptive-weighted loss function which aims to improve anomaly detection by modifying the training focus, emphasizing background pixel reconstruction over anomaly pixels. Note that Wang et.al. demonstrated the method using hyperspectral imagery in a non-agricultural anomaly setup. This method employs a weight map that adjusts the impact of each pixel on the loss based on its reconstruction error, calculated as

$$e_{i,j} = (x_{i,j} - \tilde{x}_{i,j})^2 \tag{13}$$

where $x_{i,j}$ represents the true pixel value, and $\tilde{x}_{i,j}$ is its reconstruction by the network. A reconstruction error map

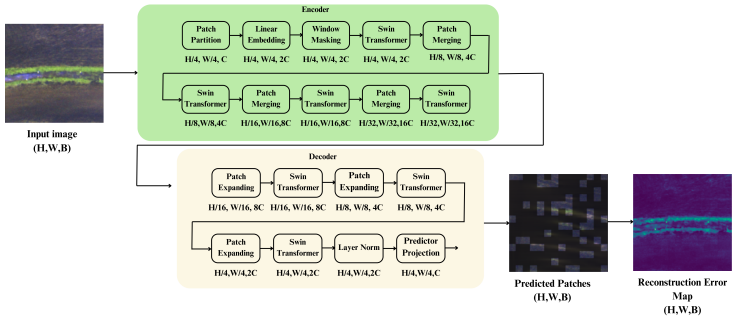


Figure 4. Architecture of the Swin Masked Autoencoder (Swin MAE) for anomaly detection. The encoder, leveraging Swin Transformer blocks, processes the input image through stages of patch partitioning, embedding, and window masking, followed by successive Transformer and merging layers to create high-dimensional token representations. The decoder employs a sequence of expanding, Transformer, and normalization layers before projecting back to the pixel space, resulting in the reconstructed image and its corresponding reconstruction error map.

E is constructed from these errors, serving as a basis for anomaly detection:

$$E = \begin{bmatrix} d_{1,1} & \dots & d_{1,W} \\ \vdots & \ddots & \vdots \\ d_{H,1} & \dots & d_{H,W} \end{bmatrix} \tag{14}$$

To prevent anomalies to get reconstructed, an adaptive-weighted loss function is utilized. The weight map W is derived by taking residual from the maximum error pixel :

$$w_{i,j} = \max(E) - e_{i,j} \tag{15}$$

$$W = \begin{bmatrix} w_{1,1} & \dots & w_{1,W} \\ \vdots & \ddots & \vdots \\ w_{H,1} & \dots & w_{H,W} \end{bmatrix} \tag{16}$$

After a few initial iterations the reconstruction error for anomalous pixels are very high when compared to background pixels , taking a residual from the maximum error pixel allows anomaly pixels to have less weight compared to the background pixels as the majority of anomaly pixels are closer to the maximum error. This weight map is updated periodically. The adaptive-weighted loss L is then computed as:

$$L = \sum_{i=1}^{H} \sum_{j=1}^{W} w_{i,j} e_{i,j} \tag{17}$$

By reducing the weights of pixels with large reconstruction errors early in training, the network is discouraged from focusing on anomalies, thus prioritizing background reconstruction. This method leads to an anomaly-suppressed model that can more accurately identify anomalies based on the reconstruction error map.

4. Proposed Method

We use the SwinMAE architecture (Section 3.3) along with the anomaly suppression loss (Section 3.4) to learn the "normal" feature embeddings from the farm images. The architecture of SwinMAE is divided into encoder and decoder as given in Figure 4. The encoder of the Swin MAE network begins by partitioning the input image into non-overlapping patches and mapping these patches into a high-dimensional embedding space through a linear transformation, which allows for more complex feature extraction.The patches then go through widow masking strategy. The Swin Transformer blocks, which form the core of the encoder effectively captures hierarchical features by using window based multi-head self-attention (W-MSA) followed by shifted (SW-MSA) window mutli-head attention. After each Swin Transformer block the embeddings are merged to reduce the number of patches by half while doubling the dimensionality of the embeddings. This operation aggregates information from adjacent patches and reduces the spatial resolution, while increasing the feature dimension. The encoder consists of 4 Swin transformer blocks as given in the original work [7]. The decoder in the Masked Autoencoder (MAE) network undertakes the task of reconstructing the input image from its condensed and partially masked representation produced by the encoder. It begins by processing the mixed embeddings through Swin Transformer blocks. Following this, the decoder employs a series of expanding operations that gradually restore the spatial resolution of the image. This is essentially the reverse of the encoder's patch merging process. As the resolution is increased, the complexity of the feature representation is reduced, aligning it closer to the original input space. Layer normalization steps interspersed within these operations ensure stable learning by maintaining a consistent scale of the features. The final stage involves a projection of the embeddings back to

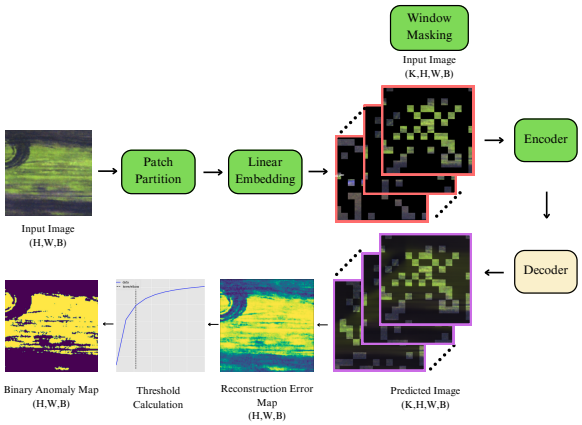


Figure 5. Anomaly Detection using Swin Masked Auto-encoder. An UAV image of shape Height,Width,Number of Bands (H,W,B) is input to the Swin MAE encoder where K window masked images are fed into the rest of the encoder comprised of Swin Tranformer and Patch Merging layers. The Swin MAE decoder then produces resulting K predicted images (K,H,W,B) . The predicted image is compared to the original input image to produce a reconstruction error map, which is thresholded using Knee-point calculation producing the final binary anomaly map delineating the detected anomalies within the image.

a space that mirrors the original image's patches, which are then reassembled to predict the full image, effectively filling in the masked regions with the learned information. This reconstructed output aims to be as close as possible to the original unmasked image.

Given an input image X, a subset of patches P representing 25% of X are fed into the SwinMAE model as given in Figure 5. This process is repeated K times with different random subsets, where $K = 32$ so that each patch is likely to be masked once, allowing for accurate reconstruction assessment for each pixel.For each iteration i, a reconstruction R_i is obtained and compared with X to compute a reconstruction error map E_i. The error for each pixel j in the error maps is averaged across all N reconstructions to obtain an averaged error map $\bar{E}$:

$$\bar{E}_j = \frac{1}{K} \sum_{i=1}^{K} E_{ij} \tag{18}$$

The final anomaly map A is produced by applying a threshold θ, determined by identifying the knee point [19] in the distribution of $\bar{E}$, to binarize $\bar{E}$ into anomalous ($A_j = 1$) and non-anomalous ($A_j = 0$) pixels:

$$A_j = \begin{cases} 1 & \text{if } \bar{E}_j \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

5. Experimental Setup

5.1. Dataset

We evaluate anomaly detection on Agriculture Vision challenge dataset.The Agriculture-Vision dataset [5] is a large aerial image database aimed at agricultural pattern analysis, designed to boost research in computer vision for agriculture. It contains 94,986 high-quality aerial images from 3,432 farmlands across the U.S., with each image including RGB and Near-infrared (NIR) channels with resolutions up to 10 cm per pixel.The dataset employs a cropping technique on large farm images with a window size of 512 × 512 pixels for annotations.The images are annotated with 9 types of field anomaly patterns such as such as double plant, drydown, endrow, nutrient deficiency, water,weed cluster,planter skip , storm damage and waterway which are crucial to farmers and serves as a benchmark for agricultural semantic segmentation, posing unique challenges due to the large inter-class and intra-class variations. A total of 56,944 images for training, 18,334 for validation and 19,708 for testing is created. We benchmark our methods on all 9 classes of anomalies, while previous results were only reported on 6 classes - Double Plant, Drydown, Endrow, Nutrient Deficiency, Water and Weed Cluster.

5.2. Evaluation Metrics

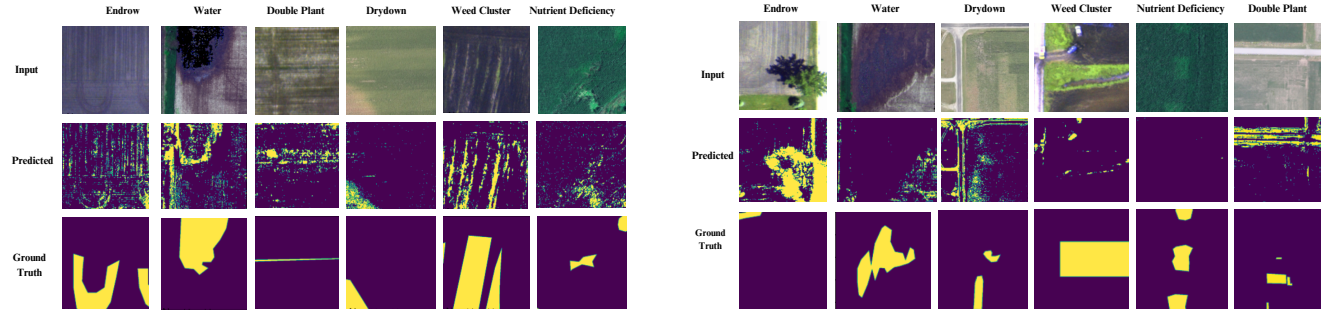
We evaulate the anomaly detection task a semantic segmentation task and use Intersection over Union (IoU), as it quantitatively assess how closely the predicted anomaly map aligns with the ground truth annotation.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{20}$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{21}$$

Method	Drydown	Double plant	Endrow	Weed cluster	ND	Water	Planter Skip	Waterway	Storm Damage	mIOU*
DSVDD[30]	30.8	14.7	10.1	3.0	24.7	26.0	-	-	-	18.2
RIAD[29]	**31.0**	25.6	27.3	38.2	25.7	**42.7**	-	-	-	31.7
ARNet[28]	30.6	15.3	25.5	15.9	26.4	9.6	-	-	-	20.5
GANomaly[1]	26.3	4.2	26.4	**41.9**	33.6	20.7	-	-	-	25.3
ASD[13]	34.6	24.8	25.7	19.7	31.7	40.4	-	-	-	29.4
InTra[17]	26.1	44.2	41/0	36.8	33.4	35.2	51.0	**47.1**	**42.5**	29.3
MAE	27.6	44.0	43.1	34.6	33.2	35.2	50.2	45.7	41.7	36.2
SwinMAE	27.9	46.3	43.3	36.7	33.8	37.4	51.5	45.3	41.9	37.5
SwinMAE + ASL (Ours)	28.1	**46.6**	**43.8**	37.8	**34.1**	37.8	**52.7**	46.9	42.3	**38.0**

Table 1. The comparative performance of anomaly segmentation methods on the Agriculture-Vision dataset. ASL = Anomaly Suppression Loss. SwinMAE based anomaly detection is able to beat existing benchmarks. Anomaly Suppression Loss (ASL) further improves accuracy through selective suppression of anomalies. *Since previous methods reported mean across 6 classes, we also reported mean IoU over 6 classes for a fair comparison. These numbers are reported by training in a leave-one-out fashion.



(a) Good examples

(b) Poor Examples

Figure 6. Qualitative results demonstrating the model's performance across six anomaly classes.

5.3. Implementation Details

The input images to SwinMAE are resized from 512x512 to 224x224 for computational efficiency and 4 bands were used, RGB and Near-infrared (NIR). The SwinMAE takes in the entire training data as input across all the classes and provides a single model for anomaly detection across the given classes.The encoder and decoder comprises of 4 Swin Transformer blocks.The initial input image patch size is 4x4. AdamW [27] optimizer with initial learning rate of 1e-3 and weight decay was used to train the model. Experiments were conducted on a single 24G A5000 GPU and on a machine with 256GB RAM.The model was trained on a total of 200 epochs and with a batch size of 64. We train SwinMAE for 20 epochs initially and then the weight maps are updated for anomaly loss compression. Also, to benchmark MAE, the training setup remains the same as SwinMAE .

6. Results

We compare SwinMAE with several state of the art anomaly detection algorithms ranging from convolutional, GAN-based, One-class classification (OCC), Transformers and MAE based models. To be specific quantitative analysis of Swin MAE in comparison to several models such as DSVDD [30], RIAD [29], ARNet [28], GANomaly [1], and the Anomaly Segmentation model based on Pixel Descriptors (ASD) [13], Inpainting Transformer [17] and MAE. Our quantitative experiments in Table 1 reveals that Swin MAE outperforms previous unsupervised and self supervised approaches across several anomaly categories. Supervised models like AAFormer [21] achieves an mIOU of 41.2 and Fuse-PN [11] achieves a dice score of 82.71 acting as strong baselines. For instance, in the Double plant and Endrow category, Swin MAE achieved an mIOU of 46.6 and 43.8 which is significantly better than other mod-

Training Data	DryD	DP	Endrow	WC	ND	Water	PSkip	WW	SD	mIOU	medIOU
Excluding anomalous samples	**45.83**	**46.9**	43.5	**37.9**	33.7	**45.97**	**53.3**	**47.5**	42.0	**44.0**	**45.8**
Including anomalous samples	28.1	46.6	**43.8**	37.8	**34.1**	37.8	52.7	46.9	**42.3**	36.44	42.3

Table 2. When anomalous samples are included in the dataset, the anomaly suppression loss helps maintain the IoU over most classes. DryDown and Water show losses since these are semantically very different from the input data distribution and spread widely in the image. The median is very close to the original paradigm while the mean is lower due to the two classes - Drydown and Water.

els. Swin MAE performs on par if not better across all the categories of anomaly. Moreover, when augmented with an anomaly suppression loss, the model's proficiency further increases, where it recorded a jump in all categories showing the efficacy of the anomaly suppression mechanism during training. A Masked autoencoder also performs well compared to the other models , which enforces the fact that masked image modelling is very effective in detecting anomalies across multiple categories. However as Swin MAE learns both local and global features it outperforms MAE in all the categories. It is interesting to note that anomalies which spread out in the images, e.g., weeds, water, etc. are not caught well with reconstruction based methods, perhaps, owing to the fact that despite 75% masking, the pattern is caught by the encoder and successfully reproduced, thus acting as the normal distribution.

Label-free training We also analyze training SwinMAE with and without anomaly samples in the training set for a given class, as detailed in Table 2. Note that all methods in Table 1 are run without anomalies in the training data, i.e., in a Leave-one-out fashion. In general, it is observed that with our method, accuracies remain almost the same (or increase slightly by a few points in some cases) for all classes of anomalies. The median IOU only reduces by 3.5 points. For two classes - drydown and water, the IoU increases significantly when these classes are omitted during training. This effect can be attributed to the distinctive nature of these two classes, i.e., water and drydown not being present at all in the underlying data, which are less likely to be reconstructed if they are absent from the training set due to their minimal distribution overlap with other types of anomaly classes which are present in the training set.

These results indicate that Swin MAE is robust across most of the anomaly categories while it is trained with anomalous samples included in the training set and significantly outperforms other models which are trained on anomaly free samples which enables us to have to a single model better generalizing to all categories. Swin MAE ability to model both local and global dependencies across patches enables it to detect anomalies better than other transformer based techniques like a simple MAE and InTra.

Qualitatively we can observe in Figure 6a, that SwinMAE is able to segment irregular anomaly patterns across multiple classes. Due to large intra-class and inter-class variance within every class in the Agriculture Vision Dataset [5], we observe that in a few classes like Nutrient Deficiency as shown in Figure 6a there are other miscellaneous anomaly patterns apart from nutrient deficient areas that get segmented as anomalies, similarly for classes like water and double plant. In Figure 6b there are objects such as Trees in (Endrow), Roads in (Drydown, Double Plant) and cars in (Weed) that are detected as anomalies because they deviate from normal agricultural patterns and global pattern of the image, however these are not agricultural anomalies. As our model is trained in a self-supervised setting and without agricultural anomaly specific supervision, any pattern that deviates from the global setting of the query image might be labeled as anomaly. This also solidifies the fact that Masked image modelling based transformer models are efficient global feature learners.

7. Conclusion and Future Work

We propose a masked image modelling based self supervision methodology to detect anomalies in agricultural fields using UAV images. We demonstrate the effectiveness of masked image modelling through SwinMAE[7] and MAE[9] methods by benchmarking it on the Agriculture Vision dataset and improving mIOU by a margin of 6.3% compared to other unsupervised and self supervised methods. We also show that using an anomaly suppression loss[25] adds robustness even when trained with training data containing anomalous samples. Typically, the anomaly detection methods train only on normal samples (excluding the anomalous classes' patches from the input). This improvement allows a simplification of the pipeline so that anomaly detection pipelines can be trained without filtering out "normal" samples. With this improved methodology, it is also possible to have a single model that can generalize across all the anomaly classes. Our work should provide a definitive direction in creating an anomaly detection system that generalizes to various anomalies and relies less on human interventions. As future work, we will investigate if the single model is able to find "new" anomalies beyond the one described in the dataset and if the method is useful in domains other than UAV images as well. We will also investigate if a generic model will be able to work across different types of crops.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019. 7

[2] Tanmay Anand, Soumendu Sinha, Murari Mandal, Vinay Chamola, and Fei Richard Yu. Agrisegnet: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. *IEEE Sensors Journal*, 21(16):17581–17590, 2021. 2

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021. 4

[4] Mang Tik Chiu, Xingqian Xu, Kai Wang, Jennifer Hobbs, Naira Hovakimyan, Thomas S Huang, and Honghui Shi. The 1st agriculture-vision challenge: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–49, 2020. 2

[5] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020. 1, 2, 6, 8

[6] Jun Kang Chow, Zhaoyu Su, Jimmy Wu, Pin Siang Tan, Xin Mao, and Yu-Hsing Wang. Anomaly detection of defects on concrete structures with the convolutional autoencoder. *Advanced Engineering Informatics*, 45:101105, 2020. 2

[7] Yin Dai, Fayu Liu, Weibing Chen, Yue Liu, Lifu Shi, Sheng Liu, Yuhang Zhou, et al. Swin mae: masked autoencoders for small datasets. *Computers in biology and medicine*, 161: 107037, 2023. 2, 4, 5, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 8

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[11] Shubham Innani, Prasad Dutande, Bhakti Baheti, Sanjay Talbar, and Ujjwal Baid. Fuse-pn: A novel architecture for anomaly pattern segmentation in aerial agricultural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2021. 2, 7

[12] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. 2

[13] Jingtao Li, Xinyu Wang, Hengwei Zhao, Shaoyu Wang, and Yanfei Zhong. Anomaly segmentation for high-resolution remote sensing images based on pixel descriptors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4426–4434, 2023. 2, 7

[14] Haijun Liu, Xi Su, Xiangfei Shen, Lihui Chen, and Xichuan Zhou. Bigset: Binary mask-guided separation training for dnn-based hyperspectral anomaly detection. *arXiv preprint arXiv:2307.07428*, 2023. 2

[15] Yunfei Liu, Chaoqun Zhuang, and Feng Lu. Unsupervised two-stage anomaly detection. *arXiv preprint arXiv:2103.11671*, 2021. 2

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[17] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, pages 394–406. Springer, 2022. 2, 7

[18] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation, 2021. 2

[19] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011. 6

[20] Eli Schwartz, Assaf Arbelle, Leonid Karlinsky, Sivan Harary, Florian Scheidegger, Sivan Doveh, and Raja Giryes. Maeday: Mae for few-and zero-shot anomaly-detection. *Computer Vision and Image Understanding*, page 103958, 2024. 2

[21] Yao Shen, Lei Wang, and Yue Jin. Aaformer: a multi-modal transformer network for aerial agricultural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1711, 2022. 2, 7

[22] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004. 2

[23] Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[25] Shaoyu Wang, Xinyu Wang, Liangpei Zhang, and Yanfei Zhong. Auto-ad: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder. *IEEE*

Transactions on Geoscience and Remote Sensing, 60:1–14, 2021. 4, 8

[26] Jie Yang, Yong Shi, and Zhiquan Qi. Dfr: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122*, 2020. 2

[27] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, pages 10665–10673, 2021. 7

[28] Fei Ye, Chaoqin Huang, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 24:116–127, 2020. 2, 7

[29] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2, 7

[30] Zheng Zhang and Xiaogang Deng. Anomaly detection using improved deep svdd model with data structure preservation. *Pattern Recognition Letters*, 148:1–6, 2021. 2, 7