Research Square

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

DFINE-AGS: A Representation-Preserving Transformer Detector for Potato Detection in Complex Field Environments

qiyu zhang
Qingdao Agricultural University
guangfei liu
Qingdao Agricultural University
huan zhang
Qingdao Agricultural University
ranbing yang
Hainan University
zhiguo pan
200401043@qau.edu.cn

Qingdao Agricultural University

Article

Keywords: Potato, object detection, Transformer detector, multi-scale feature fusion, edge inference

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11513428/v1

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

Additional Declarations: No competing interests reported.

DFINE-AGS: A Representation-Preserving Transformer Detector for Potato Detection in Complex Field Environments

Qiyu Zhang[1], Guangfei Liu[1], Huan Zhang[1],Ranbing Yang[2*], Zhiguo Pan[1*]
[1] School of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao, China
[2] School of Mechanical and Electrical Engineering, Hainan University, Haikou, China

Abstract

Reliable recognition of potato tubers in open-field scenes remains difficult because targets frequently exhibit soil adhesion, partial burial, mutual occlusion, large apparent-scale variation, uneven illumination, and strong visual similarity to surrounding soil and residues. We propose DFINE-AGS, a D-FINE-derived detector formulated as a coordinated representation-preservation framework for these field-induced degradations. GC-BRA combines global semantic anchoring, content-dependent region routing, and local spatial compensation to improve target-background discrimination; SPDA-Conv preserves phase-structured local evidence before spatial reduction; and the P3/P4/P5 Zoom_cat-ScalSeq-ASF pathway models ordered cross-scale interactions between fine spatial detail and high-level semantics. On the self-collected potato dataset, DFINE-AGS achieved 90.4% precision, 91.5% recall, a 90.9% F1-score, and 90.0% mAP@0.5. TensorRT inference on an NVIDIA Jetson Orin Nano reached 37.4 FPS with 90.3% mAP@0.5. An independent field evaluation containing 1,297 manually verified potato targets yielded 89.67% precision and 90.36% recall. These results demonstrate that DFINE-AGS provides a robust and edge-deployable detection framework for complex field environments while retaining a favorable accuracy-efficiency trade-off.
Keywords: Potato; object detection; Transformer detector; multi-scale feature fusion; edge inference

1 Introduction

Potato is the fourth most important food crop worldwide and also serves as both a staple and vegetable crop in China. Major production areas include Dingxi (Gansu), Ulanqab (Inner Mongolia), and Bijie (Guizhou). Potato tubers supply starch, protein, vitamin C, potassium, and several bioactive compounds, which underpins their nutritional and economic importance[1-2]. Demand for fresh, processed, and high-quality potato products has continued to expand as potato production and consumption have become more diversified.

Harvest timing has a direct effect on potato yield and market quality. Mature plants are generally harvested after canopy senescence and completion of skin set, when dry-matter accumulation is high and tubers are less susceptible to mechanical injury[3]. Allowing fully mature tubers to remain in the soil over an extended time may induce secondary growth and greening. It also increases their vulnerability to cold damage, diseases and the build-up of glycoalkaloids[4-5]. For this reason, the ability to rapidly spot tubers suitable for field harvest supports automated harvesting operations and reduces quality losses that could otherwise be avoided.

Potato harvesting in China is still dominated by manual assistance or partially mechanized operations, both of which are affected by seasonal labor availability and increasing labor cost[6-7]. As agricultural equipment becomes more automated, machine vision is increasingly used to provide the perception required by field machinery. For a harvesting platform, the vision subsystem can locate tubers, estimate target number, and supply target information for route planning or subsequent manipulation[8].

Field imagery is substantially more difficult than controlled laboratory imagery. Tubers can have color and texture close to those of surrounding soil, stones, and crop residues, so target-background contrast is often weak[9]. They may also overlap, remain partly buried, or be hidden by neighboring material, leaving only incomplete contours. In addition, illumination, shadow, soil moisture, cultivar, and maturity alter appearance from one scene to another[10-11]. Rule-based approaches built around thresholds, edges, or morphology are sensitive to these changes because the relevant visual features must be specified in advance. A detector intended for harvesting therefore needs to retain fine target structure while remaining robust to background and environmental variation[12].

Deep learning has changed agricultural object detection by allowing visual features to be learned directly from data rather than prescribed manually[13]. CNN-based detectors are now common in crop recognition, disease monitoring, and machine-vision systems. One-stage detectors, particularly members of the YOLO family, are attractive for field use because they offer a practical compromise between computational cost and real-time accuracy.

Potato-oriented studies have mainly improved speed or robustness through lightweight backbones, attention blocks, and feature-fusion designs. Zhang et al.[14] replaced the conventional backbone of YOLOv4 with MobileNetV3 and reported 91.4% average accuracy at 23.01 FPS in a complex harvesting environment. Wang et al.[15] used an improved YOLOv4 to distinguish potato-related clods and stones; its average accuracy exceeded Faster R-CNN by 11.2 percentage points while running at 78.49 FPS. More recently, Kim et al.[11] applied YOLOv12 to potato harvesting and sorting, using R-ELAN and area-based attention to improve difficult-class recognition. These CNN-based designs are efficient, but severe occlusion and visually confusing backgrounds can still weaken feature discrimination.

Transformer detectors provide a different route because self-attention can model dependencies between distant image regions[16-18]. D-FINE is a real-time DETR-family detector that reformulates box regression with fine-grained

distribution refinement[19]. When transferred directly to field potato images, however, three limitations are apparent. Its AIFI-based encoder performs fixed-form feature interaction rather than selecting context according to scene content; the P4/P5 prediction pathway makes limited use of shallow spatial detail needed for small or partially visible tubers; and repeated downsampling can suppress weak boundaries and local texture. These observations motivated changes to the encoder, the scale-fusion pathway, and the downsampling operation.

Accordingly, DFINE-AGS is formulated as a coordinated representation-preservation framework rather than as a collection of independent plug-in modules. Three successive degradation processes are considered. First, contextual ambiguity can weaken target-background separation during feature interaction; GC-BRA therefore combines global semantic anchoring with content-dependent regional routing so that aggregation is concentrated on regions relevant to the current scene[20]. Second, repeated spatial reduction can irreversibly suppress weak contours and local texture; SPDA-Conv reorganizes spatial samples before convolution so that phase-structured local evidence is preserved during downsampling[21]. Third, large apparent-scale changes can create inconsistent representations across pyramid levels; the P3/P4/P5 Zoom_cat-ScalSeq-ASF pathway aligns, models, and adaptively reweights ordered scale states to maintain continuity between shallow detail and deeper semantics[22]. In this way, the three modifications operate at different stages of the representation pipeline while addressing complementary field-induced failure modes.

The contributions of this work are summarized as follows:

(1) A failure-driven formulation is introduced in which complex-field potato detection is decomposed into contextual ambiguity, spatial-detail degradation, and cross-scale inconsistency. This formulation provides a unified rationale for modifying the encoder, downsampling, and feature-aggregation stages of D-FINE.

(2) GC-BRA is developed as a three-part contextual interaction mechanism consisting of global semantic anchoring, region relevance routing, and local spatial compensation. The design aims to suppress visually confusing background interactions while retaining local evidence around partially visible tubers.

(3) SPDA-Conv and the ZSA pathway preserve complementary information at two different stages: SPDA-Conv retains phase-structured spatial evidence before resolution reduction, whereas ZSA treats P3/P4/P5 as an ordered scale sequence and adaptively integrates detail and semantic information across levels.

(4) The detector is evaluated through controlled ablation, comparison with representative real-time detectors, condition-specific field analysis, Jetson Orin Nano deployment, and independent field validation, allowing both recognition robustness and computational efficiency to be assessed.

2 Materials and Methods

The experimental workflow covers four stages: dataset construction and preprocessing, development of DFINE-AGS, controlled detector evaluation, and edge/field deployment. Model comparisons and ablation experiments were performed on the same potato dataset. The selected model was then transferred to an embedded platform and integrated with a small tracked machine to assess real-time field perception. The overall workflow is summarized in Fig. 1.

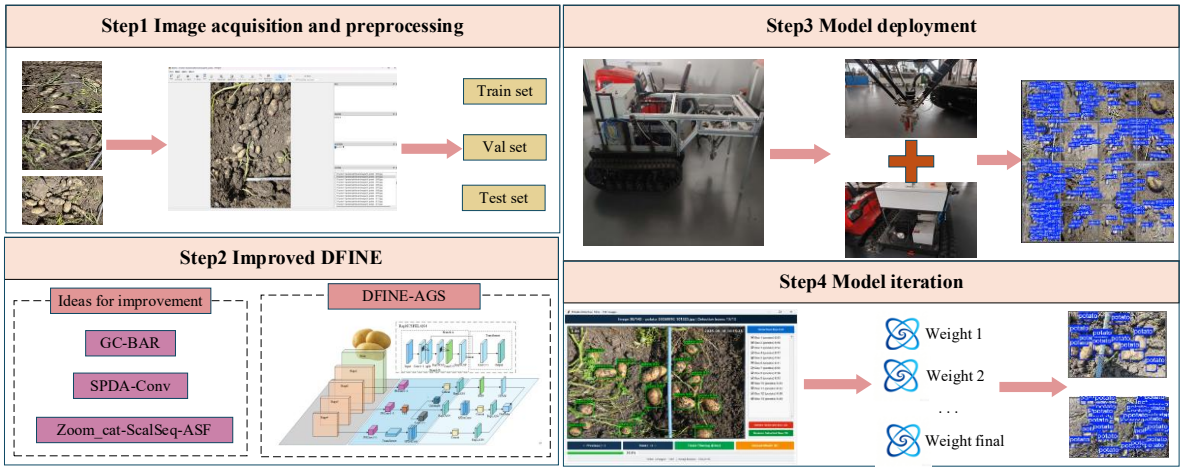


Fig. 1. Overall Research Approach

2.1 Experimental Data and Analysis

The dataset was assembled from field photographs containing exposed potato tubers together with soil clods, plant residues, and drip-irrigation tape. Images were acquired in Ulanqab, Inner Mongolia, in September 2025 at a native resolution of 4096 x 3072 pixels. The collection location and representative field scenes are shown in Fig. 2.

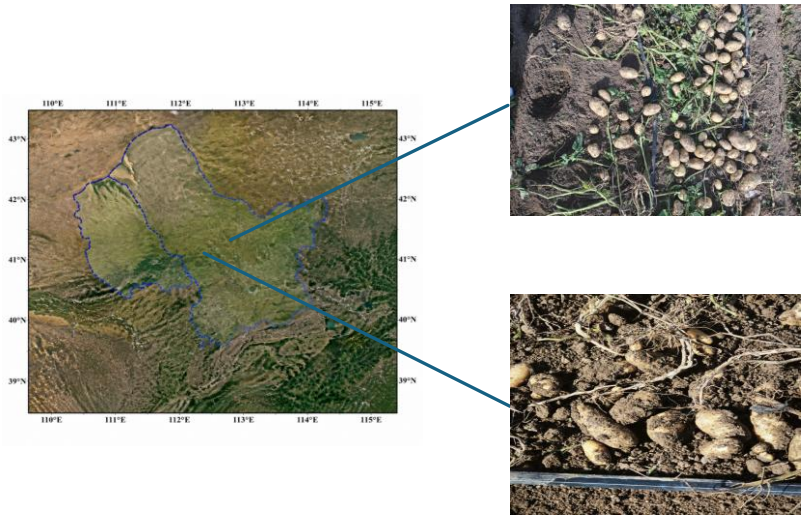


Fig. 2. Locations Where the Dataset Was Collected

Images were collected in both sandy-loam and clay-soil plots so that the dataset would include two common background types. In total, 5,255 images were retained, with the two soil categories represented in approximately equal proportions. The data were divided into 3,678 training images, 1,051 test images, and 526 validation images (7:2:1). Augmentation was applied only to the training partition. Mosaic augmentation combined four randomly transformed images, while Mixup-style hybrid augmentation blended two images at a random ratio. Directional filtering was also applied to the collected images, and the filtered training images were added to the original training set. This procedure doubled the training partition to 7,356 images; the validation and test partitions were left unchanged. An example of the filtering result is provided in Fig. 3.

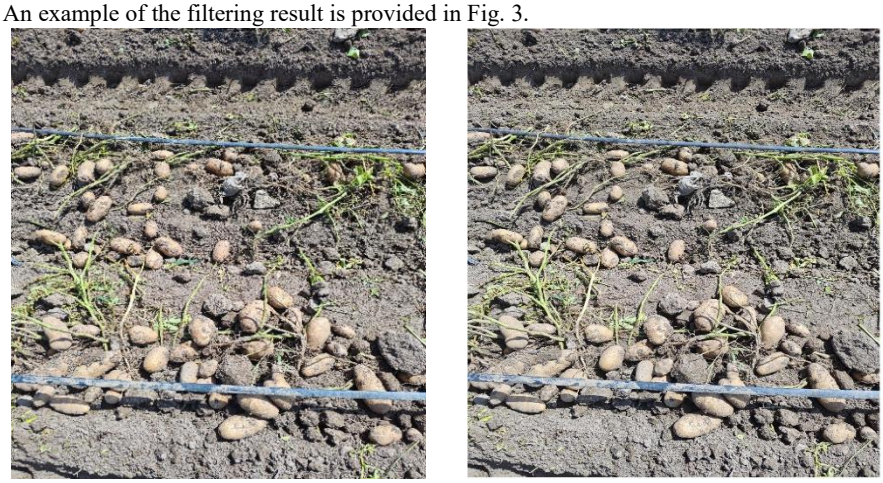
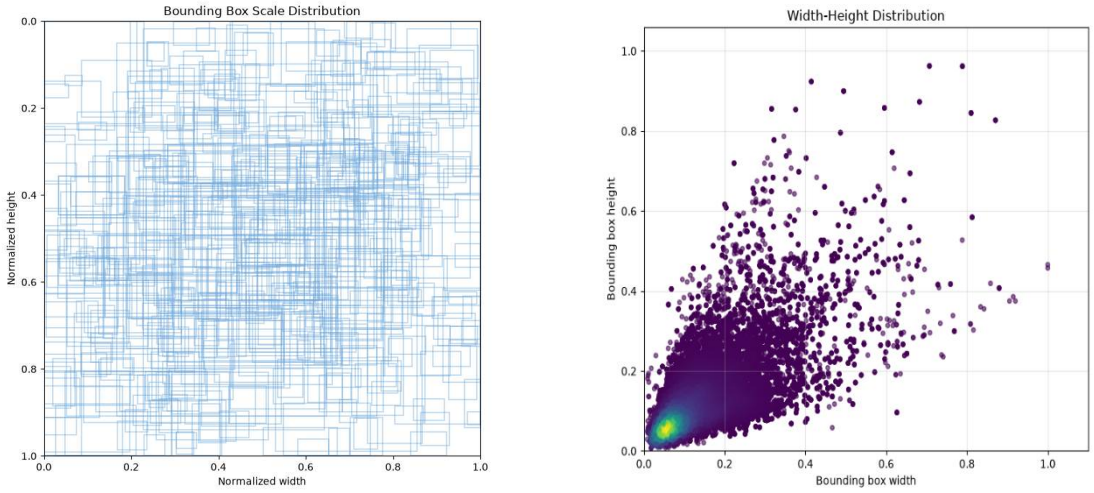


Image before filtering Filtered Image

Fig. 3. Illustration of the noise reduction effect achieved through directional filtering

We also examined the annotation geometry before model training. Fig. 4(a) shows that box centers were concentrated mainly within the effective central field of view, although the covered area expanded substantially for larger targets. This spread reflects variation in camera-to-target distance, pose, partial burial, and occlusion. The dataset therefore contains a mixture of small, medium, and large instances rather than a single dominant target size. Small boxes are especially common, which makes preservation of fine spatial cues important. Fig. 4(b) plots normalized box width against height. Most instances lie in the low-width/low-height region, but the distribution extends over a broad range and follows only a loose positive width-height trend. Irregular tuber shape, orientation, overlap, and truncation all contribute to this dispersion. The annotation statistics thus support the need for multi-scale fusion and boundary-aware feature extraction.



(a) (b)

Fig. 4. Statistical analysis of annotation distribution in the dataset. (a) Bounding box scale distribution; (b) Width–height distribution of bounding boxes.

2.2 Preparation of Laboratory Equipment

Training was performed under Linux using PyTorch and Python in PyCharm. The workstation contained an AMD EPYC 7K62 processor and an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. The main software and hardware settings are listed in Table 1.

Table 1: Experimental Environment Configuration

Category	Configuration
CPU	AMD EPYC 7K62 48-Core Processor
GPU	Nvidia 4090-24G
System environment	Linux 6.1.0-33-amd64
Framework	Pytorch2.1.0
Programming	Python3.10.14

For embedded testing, the trained detector was transferred to a compact field-perception setup consisting of an Intel RealSense D456 depth camera, an NVIDIA Jetson Orin Nano Super (8 GB), and a 7-inch display. This platform was used for real-time deployment and simulated field experiments.

3 DFINE-AGS Potato Detection Model

3.1 Standard DFINE Model

DFINE-N was selected as the lightweight baseline because its computational scale is suitable for embedded deployment. The network can be viewed as three functional blocks: a hierarchical convolutional backbone, a hybrid encoder for multi-scale feature interaction, and an end-to-end Transformer decoder. The backbone produces feature maps at multiple resolutions; the encoder mixes local spatial information with higher-level semantic context; and the decoder predicts object categories and boxes. Bipartite matching based on the Hungarian algorithm assigns predictions to ground-truth objects during training, so the inference pipeline does not require conventional non-maximum suppression.

The DFINE-N backbone begins with a StemBlock and is followed by four HG_Stage blocks. StemBlock performs early feature extraction and spatial reduction. Each HG_Stage stacks lightweight convolutional operations and includes squeeze-and-excitation channel weighting, allowing informative channels to be strengthened while less useful responses are suppressed.

3.2 Enhanced DFINE Architecture

DFINE-AGS modifies D-FINE-N according to a three-stage representation-degradation analysis. The first stage concerns contextual ambiguity: potato regions can be locally similar to soil and residues, so GC-BRA replaces fixed-form encoder interaction with scene-dependent contextual routing. The second stage concerns irreversible spatial-detail loss: weak contours and texture can disappear during resolution reduction, so SPDA-Conv preserves phase-structured local samples before convolutional compression. The third stage concerns cross-scale inconsistency: target evidence can shift markedly between P3, P4, and P5 as apparent tuber size and visible area change, so ZSA aligns these levels, models their ordered scale relationship, and reweights the fused response according to image content. The three components therefore act at different representation stages and are designed to be complementary rather than redundant. Fig. 5 gives the resulting architecture, and Table 2 lists the stage-level configuration.

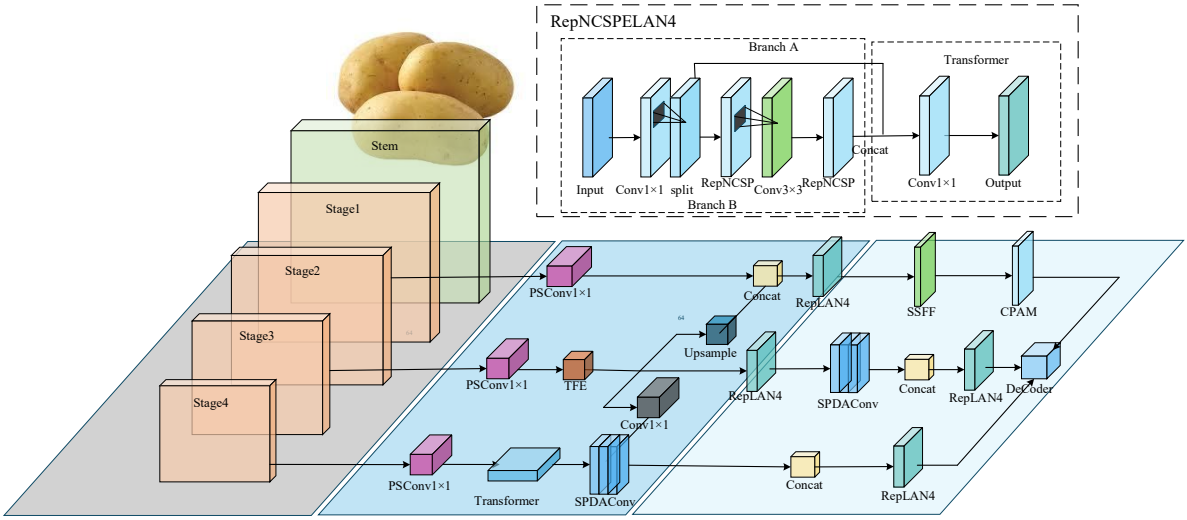


Fig. 5. DFINE-AGS Network Architecture (where TFE stands for Zoom_cat, SSFF stands for ScalSeq, CPAM stands for asf_attention, and RepLAN4 stands for RepNCSPELAN4)

Table 2: Stage-wise configuration of the proposed DFINE-AGS network.

Stage	Input Level	Operator	Kernel/Scale	Type
1	P2/4	SPDA-Conv	3×3	Downsampling
2	P3/8	GC-BRA	-	Encoder
3	P4/16	SPDA-Conv	3×3	Downsampling
4	P4/16	GC-BRA	-	Encoder
5	P5/32	SPDA-Conv	3×3	Downsampling
6	P5/32	GC-BRA	-	Encoder
7	P3-P5	Zoom_cat	Multi-scale	Feature Fusion
8	P3-P5	ScalSeq	Multi-scale	Feature Interaction
9	P3-P5	ASF	Attention	Adaptive Fusion

3.3 Structure of the GC-BRA

The original AIFI block performs self-attention within a predefined feature-interaction scheme. In field potato imagery, however, the most discriminative evidence is often unevenly distributed: a weak local patch may resemble soil, while its relationship with surrounding tuber regions provides the information needed for correct recognition. GC-BRA addresses this mismatch through three coupled operations. First, a global semantic anchor summarizes scene-level context and provides a prior for subsequent interaction. Second, region-level correlations are used to construct content-dependent routing weights, so each region preferentially aggregates information from semantically relevant regions rather than interacting uniformly with the full feature map. Third, a lightweight depthwise-convolution branch preserves local positional and texture cues that can be weakened by region-level aggregation. The resulting mechanism combines global context, selective non-local interaction, and local compensation within one encoder block[23].

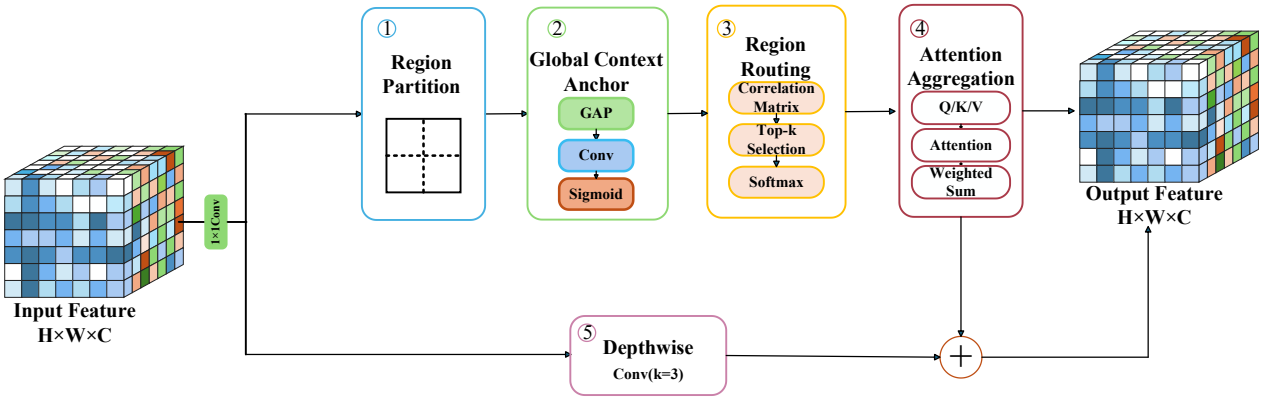


Fig. 6. Schematic Diagram of the GC-BRA Structure

GC-BRA first obtains region-level semantic representations through the global context enhancement module, as shown in Equation 1. Here, X denotes the input feature map, X_i denotes the i-th local feature, $\hat{\tau}_i$ denotes the enhanced regional feature after fusing global context information, $GAP(\square)$ denotes the global average pooling operation, which is used to extract overall semantic information from the feature space, and W_g denotes the learnable weight matrix, which is used to adjust the global context information; $\sigma(\square)$ denotes the activation function, which

is used for nonlinear feature mapping.

$$\hat{\tau}_i = GAP(X_i) + \sigma(W_g GAP(X)) \tag{1}$$

Dynamic routing is then performed using the regional correlation matrix, as shown in Equation 2. Here, A_{ij} represents the semantic correlation weight between region i and region j ; and $\hat{\tau}_i$, $\hat{\tau}_j$ denote the enhanced regional feature representations; T denotes the matrix transpose operation; C denotes the feature channel dimension; and $\sqrt{C}$ represents a scaling factor used to stabilize the attention calculation.

$$A_{ij} = \frac{\hat{\tau}_i^T \hat{\tau}_j}{\sqrt{C}} \tag{2}$$

Selection of key regions is achieved through soft routing weights, as shown in Equation 3. Here, α_{ij} represents the weight of region j's contribution to the information of region i.

$$\alpha_{ij} = Soft\max(A_{ij}) \tag{3}$$

Finally, regional attention is used to aggregate information from different regions, as shown in Equation 4. Here, Y_i represents the i-th region-enhanced feature output by the GC-BRA module; $Attention(\square)$ denotes the regional attention computation process; and Q_i, K_j, V_i represent, respectively, the information that the current region needs to match, the feature information of the candidate regions, and the actual feature content being passed; $DWConv$ denotes a deep convolution operation used to enhance local spatial position information.

$$Y_i = \sum_{j \in \Omega i} \alpha_{ij} Attention(Q_i, K_j, V_j) + DWConv(X) \tag{4}$$

GC-BRA can therefore be interpreted as a coarse-to-fine contextual discrimination mechanism. Global semantic anchoring establishes the overall scene prior, correlation-based routing determines where information should be collected, and the local branch restores spatial evidence needed for precise boundaries. Importantly, these operations address different failure modes: global context alone cannot resolve all soil-tuber ambiguities, routing alone may lose pixel-level detail, and local convolution alone lacks long-range context. Their combination is particularly relevant to partially buried, occluded, or weak-texture potatoes, for which recognition depends simultaneously on surrounding context and the remaining local boundary evidence[24-25].

3.4 The Structure of SPDA-Conv

Repeated downsampling is a critical source of information loss for small and partially visible tubers. Conventional stride-based reduction performs sampling and feature mixing simultaneously; consequently, responses located at different spatial phases inside a local neighborhood can be discarded before their importance is evaluated. SPDA-Conv decouples these operations. A space-to-depth transformation first maps neighboring sampling phases into separate channel groups, phase-sensitive attention then estimates their relative contribution, and convolutional fusion is applied only after this information has been retained. The module is therefore designed as a phase-preserving downsampling operator rather than simply as an alternative convolution.

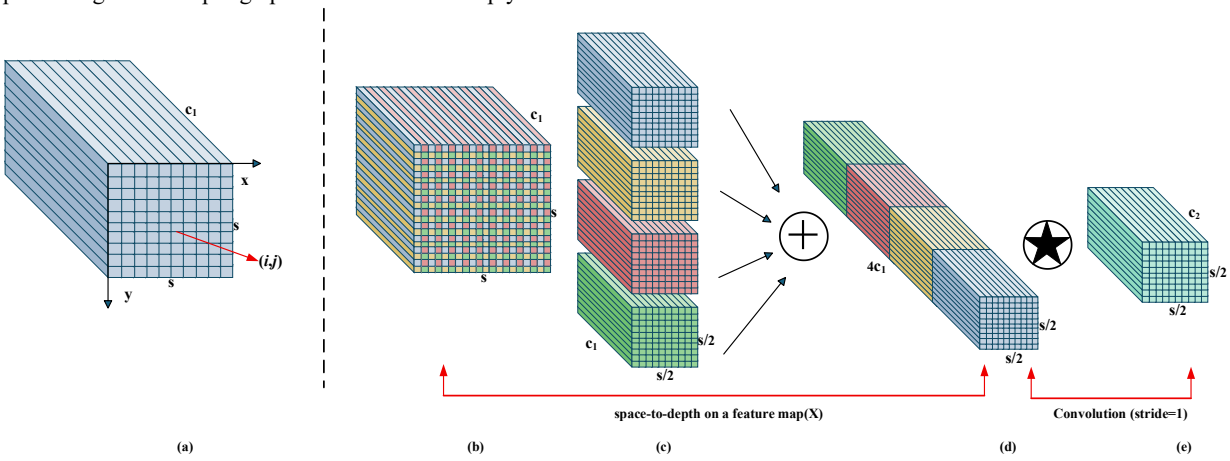


Fig. 7. Diagram of the SPDA-Conv Architecture

SPDA-Conv contains three successive operations: phase-preserving spatial rearrangement, phase-specific weighting, and convolutional feature mixing. The first operation ensures that the four sampling offsets of a local neighborhood remain explicitly represented after the spatial size is reduced. The second operation assigns adaptive weights to these phase components, allowing structurally informative edge or texture responses to receive greater emphasis. The final convolution mixes the weighted channels and maps them to the required output dimension. By separating preservation from compression, SPDA-Conv reduces the probability that weak but useful local evidence is removed at the beginning of a downsampling stage.

$$X_{sd} = Concat(X_{00}, X_{10}, X_{01}, X_{11}) \tag{5}$$

$$w_i = Soft\max(MLP(GAP(X_i))) \tag{6}$$

$$Y = Conv_{3\times3}(Concat(w_1X_1, w_2X_2, w_3X_3, w_4X_4)) \tag{7}$$

Here, X_{sd} represents the reconstructed feature obtained after the Space-to-Depth operation; X_{ij} represents the four phase subfeatures extracted from the input feature map at different spatial sampling positions; and $Concat(\cdot)$ represents the channel-dimension concatenation operation, which combines the four spatial subfeatures into a new feature representation. w_i denotes the attention weight corresponding to the *i*-th phase, and X_i denotes the *i*-th phase feature. Y denotes the final output feature of the SPDA-Conv module, and w_iX_i denotes the enhanced feature after applying the phase weights; $Conv_{3\times3}$ denotes a 3×3 convolution, used for feature fusion and channel mapping.

From a representation perspective, the main distinction between SPDA-Conv and direct strided convolution is therefore not the final spatial resolution but the information available before compression. Direct striding selects and mixes samples while reducing resolution, whereas SPDA-Conv first retains all phase offsets in the channel domain and then learns which components should dominate the compressed representation. This mechanism is expected to be most beneficial for small targets, incomplete contours, and soil-covered tubers, where a limited number of boundary pixels may determine whether the object remains distinguishable at deeper feature levels.

Table 3: Comparison of different downsampling strategies.

Downsampling	APsmall (%)	AP@0.5 (%)	Params (M)	GFLOPs	Inference time (ms)
Strided Conv	78.6	84.4	3.64	6.93	2.10
SCDown	79.3	84.9	3.72	7.12	2.15
SPD-Conv	82.2	86.3	3.80	7.43	2.18
Space-to-depth only	81.6	85.9	3.77	7.35	2.16
SPDA-Conv	83.7	87.0	3.86	7.56	2.21

As shown in Table 3, the choice of downsampling strategy has a noticeable influence on both detection accuracy and computational cost. Conventional strided convolution and SCDown provide relatively low computational complexity, but their performance on small targets remains limited because spatial reduction may discard weak boundary and texture information. SPD-Conv and the space-to-depth operation improve feature preservation to some extent, leading to higher APsmall and AP@0.5. Among the evaluated strategies, SPDA-Conv achieves the best overall detection performance, with APsmall and AP@0.5 reaching 83.7% and 87.0%, respectively. Although it introduces a moderate increase in parameters, GFLOPs, and inference time, the additional computational cost remains limited. These results indicate that the phase-aware weighting mechanism of SPDA-Conv is effective in retaining discriminative local structures during downsampling, particularly for small and partially visible potato targets.

3.5 The Structure of ZSA

The neck of DFINE-N aggregates features across resolutions, but conventional addition or concatenation treats pyramid levels mainly as independent tensors. This assumption is restrictive for potato detection because target evidence changes continuously with apparent scale: fine boundaries are strongest in P3, while more abstract shape and contextual cues dominate P4 and P5. DFINE-AGS therefore formulates P3, P4, and P5 as an ordered scale-state sequence and introduces the Zoom_cat-ScalSeq-ASF (ZSA) pathway. Zoom_cat performs scale alignment while retaining high-resolution information, ScalSeq models transitions among adjacent scale states through three-dimensional convolution, and ASF performs content-dependent channel and spatial reweighting. The resulting pathway is intended to preserve representation continuity as target evidence shifts from fine detail to high-level semantics.

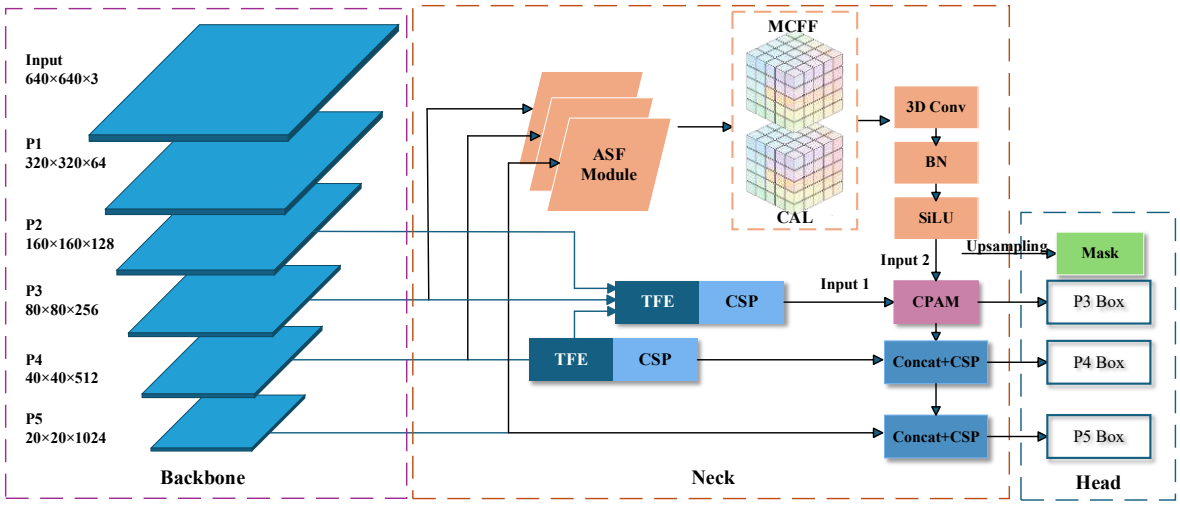


Fig. 8. ZSA Schematic Diagram

The Zoom_cat principle is described by Equation 8. Here, P_i represents multiscale features from different stages of the backbone network, $Up(\square)$ denotes the upsampling operation, and $Concat(\square)$ denotes the feature concatenation operation.

$$F_z = Concat(P_3, Up(P_4), Up(P_5)) \tag{8}$$

The principle of ScalSeq is shown in Equation 9. Here, $Conv_{3D}$ denotes a 3D convolution.

$$F_{sc} = Conv_{3D}(Concat(F_3, Up(F_4), Up(F_5))) \tag{9}$$

The principle behind ASF Attention is shown in Equation 10. Here, A_c and A_s represent the channel attention weights and spatial attention weights, respectively, and F_{out} denotes the final output features enhanced by ASF.

$$F_{out} = F_{sc} \times A_c \times A_s + F_{sc} \tag{10}$$

The three ZSA components consequently implement a progressive align-interact-select strategy. Zoom_cat first establishes a common spatial basis without forcing all information through aggressive downsampling. ScalSeq then treats the aligned P3/P4/P5 features as an ordered sequence rather than unrelated maps, allowing evidence to propagate across neighboring scale states. ASF finally selects the channels and spatial locations that are most informative for the current target configuration. This ordered scale-state modeling is particularly useful when camera distance, partial burial, or overlap causes the same class to occupy substantially different image areas. Instead of relying on a single feature level, the detector can combine fine P3 structure with stronger P4/P5 semantics according to scene content.

Table 4: Progressive ablation of the ZSA pathway.

P3	Zoom_cat	ScalSeq	ASF	APsmall (%)	AP@0.5 (%)	AP@0.5:0.95 (%)
×	×	×	×	79.3	84.9	55.8
✓	×	×	×	80.7	85.3	56.4
✓	✓	×	×	81.9	85.7	56.9
✓	✓	✓	×	83.0	86.0	57.4
✓	✓	✓	✓	84.0	86.3	57.9

Table 4 presents the progressive contribution of each component in the ZSA pathway. Introducing the P3 feature level first improves APsmall from 79.3% to 80.7%, indicating that higher-resolution shallow features provide useful spatial details for small-target recognition. After Zoom_cat is added, APsmall and AP@0.5 increase further to 81.9% and 85.7%, respectively, showing that resolution alignment and multi-scale aggregation improve information exchange among different feature levels. The addition of ScalSeq produces a further gain, confirming the benefit of explicitly modeling interactions across the ordered P3-P5 scale sequence. Finally, ASF increases APsmall to 84.0% and AP@0.5:0.95 to 57.9%, demonstrating that adaptive channel- and spatial-level reweighting further enhances the selection of informative fused features. Overall, the progressive improvement across the four configurations shows that P3, Zoom_cat, ScalSeq, and ASF provide complementary rather than redundant contributions to multi-scale representation.

3.6 Model Parameter Configuration

All detector experiments used the same training environment and were implemented within the DEIM codebase[26]. DFINE-N with an HGNetv2-N backbone served as the starting network for DFINE-AGS. Input images were resized to 640 x 640 pixels. Optimization used AdamW with a base learning rate of $6 \times 10^{\wedge -4}$ and weight decay

of $1\times10^{\wedge-4}$; the beta parameters were 0.9 and 0.999. A lower learning rate was assigned to the backbone to avoid destabilizing the pretrained low-level representation during fine tuning.

Training lasted 100 epochs. The learning rate was held constant during the initial 40 epochs and then reduced according to the schedule used in this study, with a decay factor of 0.5. Mixup augmentation was active from epochs 4 to 50 to expose the detector to a wider range of background and scale combinations[27]. Toward the end of training, strong augmentation was disabled and exponential moving averaging was used for model stabilization. Table 3 summarizes the training configuration.

Table 5:Training configurations of DFINE-AGS.

Configuration	Setting
Network architecture	DFINE-AGS
Backbone	HGNetv2-N
Framework	DEIM
Input size	640×640
Epochs	100
Optimizer	AdamW
Learning rate	6×10^{-4}
Backbone learning rate	4×10^{-4}
Weight decay	0.0001
β coefficients	(0.9,0.999)
LR scheduler	Flat + decay
Flat epochs	40
Mixup epochs	4–50
EMA epochs	92–100

Fig. 9 illustrates the learning-rate schedule used in the experiments. A short warm-up preceded the main optimization stage, after which the learning rate remained stable during early training and decreased during the later epochs. Strong data augmentation was removed near the end of optimization so that the final epochs could refine the detector on image statistics closer to those of the evaluation data. The complete run contained 100 epochs and used AdamW throughout[28].

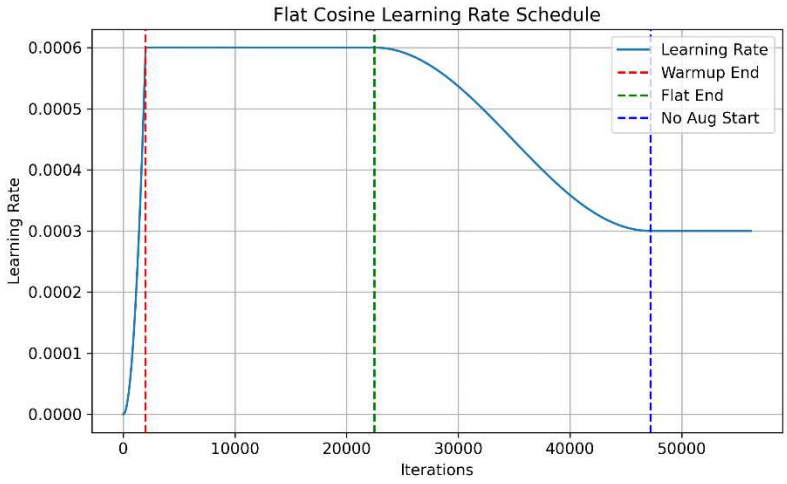


Fig.9. Learning rate scheduling strategy of the proposed DFINE-AGS model.

4 Experimental Setup and Analysis of Results

4.1 Experimental Evaluation Metrics

Detector performance was evaluated with precision (P), recall (R), F1-score, mean average precision at an IoU threshold of 0.5 (mAP@0.5), and inference time[29]. For object detection, TP denotes a ground-truth potato that is correctly matched to a prediction, FP denotes a predicted box that does not correspond to a valid ground-truth target, and FN denotes a ground-truth potato that is not detected. Precision therefore measures the proportion of predictions that are correct, whereas recall measures the proportion of annotated potatoes recovered by the detector. The equations used to compute these metrics are given below.

$$\mathrm{Precision}=\frac{TP}{TP+FP} \tag{11}$$

Recall is computed as TP/(TP+FN) and reflects target coverage: a high value indicates that few annotated potatoes are missed. Because increasing detector sensitivity can also introduce additional false positives, recall is interpreted together with precision rather than in isolation.

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{12}$$

Precision and recall emphasize different error types. A detector can increase recall by accepting weaker predictions, but this may increase false alarms; conversely, a strict confidence threshold can improve precision while discarding difficult targets. Their joint behavior is therefore more informative than either metric alone for the present field-detection task.

The F1-score is the harmonic mean of precision and recall and provides a single measure of their balance:

$$F_{1\mathrm{score}} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{13}$$

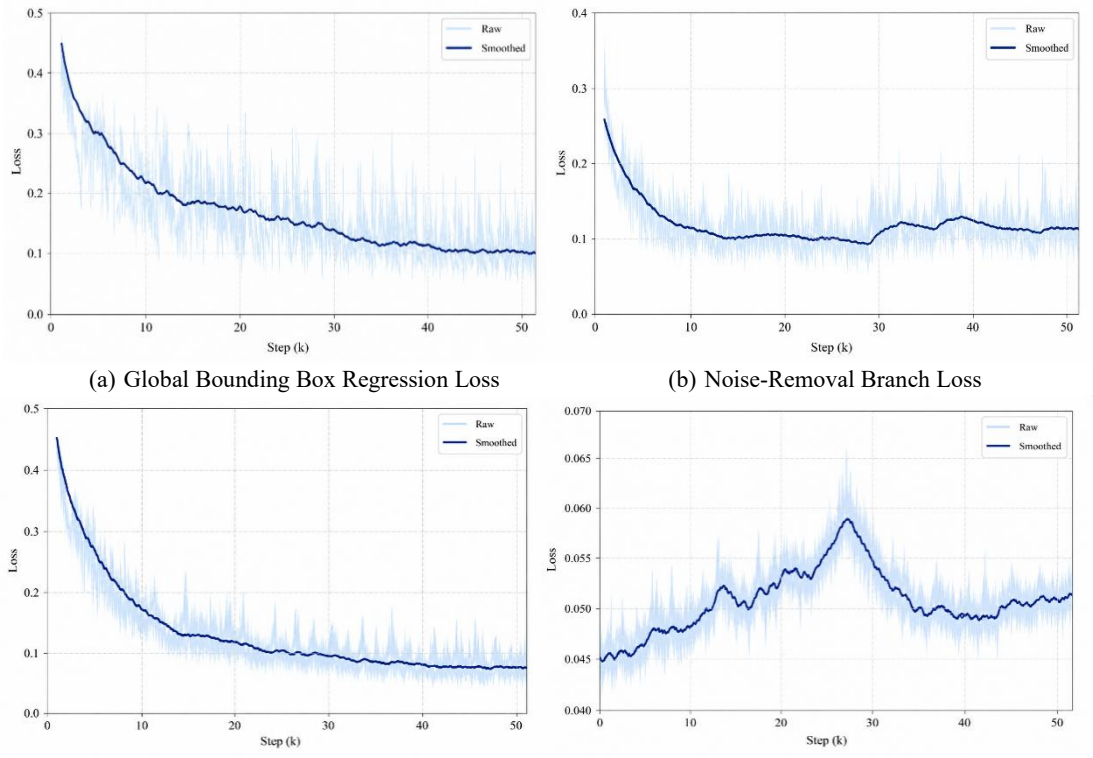
Average precision (AP) is obtained from the precision-recall relationship for a class. Because the present dataset contains one object class (potato), mAP@0.5 is numerically equal to the AP calculated at IoU = 0.5. The corresponding equations are given below.

$$AP_{\mathrm{i}} = \frac{P}{N} \tag{14}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{T} AP_I}{T} \times 100\% \tag{15}$$

Inference time was measured as the time required to process an image under the specified hardware and software configuration. It is reported alongside accuracy metrics because a detector intended for edge deployment must satisfy both recognition and latency requirements.

Training losses were inspected to confirm that the optimization process remained stable. In Fig. 10, the principal box-regression loss decreases rapidly in the early epochs and then approaches a plateau. The encoder auxiliary loss and the denoising losses follow the same overall pattern, which indicates that the main and auxiliary prediction branches converge together rather than diverging. The distribution-related auxiliary loss fluctuates more strongly at the beginning of training but subsequently settles to a low level. No branch shows persistent growth or unstable oscillation, suggesting that the final DFINE-AGS configuration is numerically well behaved under the selected training schedule.



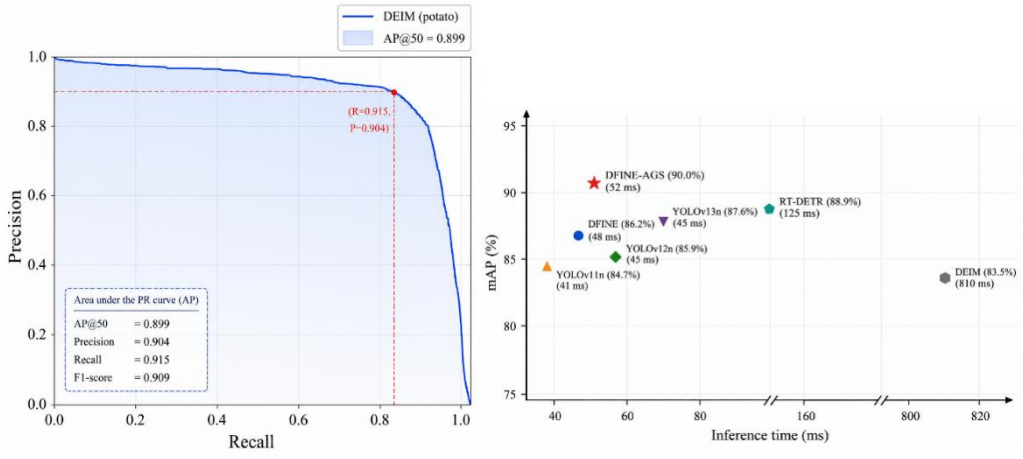
(a) Global Bounding Box Regression Loss (b) Noise-Removal Branch Loss

(c) Encoder Prediction Loss (d) Distributed Regression Loss

Fig. 10. Convergence curves of the multi-branch loss function during DFINE-AGS training

The test-set precision-recall curve of DFINE-AGS is shown in Fig. 11(a). The area under the curve at IoU = 0.5 is 0.899, indicating that high precision is retained over a broad range of recall values. At the operating point used for the reported test results, DFINE-AGS reached 0.904 precision, 0.915 recall, and an F1-score of 0.909. Fig. 11(b) places accuracy and inference time in the same view for the compared detectors. The proposed model occupies a favorable region of this trade-off: its accuracy is higher than that of the tested YOLO, RT-DETR, and DEIM

configurations while latency remains compatible with real-time use. These results support the subsequent ablation and deployment analyses.



(a) (b)

Fig. 11. Performance evaluation of the proposed DFINE-AGS model and comparison with different detection models. (a) Precision–Recall curve of DFINE-AGS; (b) Comparison of mAP and inference time among different models.

4.2 Overall Ablation and Mechanism Analysis

Ablation experiments were used to isolate the contribution of each representation-preservation mechanism[30]. Starting from D-FINE-N, all variants were trained and evaluated using the same data split and optimization settings[31-32]. GC-BRA improved the baseline when used independently, supporting the hypothesis that content-dependent regional interaction is beneficial when target and background appearances are similar[33-34]. This gain is consistent with the intended role of GC-BRA: global semantic anchoring supplies scene context, regional routing suppresses irrelevant interactions, and local compensation preserves boundary evidence. A stronger mechanism-level assessment should further separate these three internal operations, but the present overall ablation establishes that replacing the baseline interaction block provides a measurable benefit under the controlled protocol.

SPDA-Conv produced an additional improvement by changing how information is retained before spatial compression. Rather than attributing this gain only to increased model capacity, the design hypothesis is that explicit preservation of local sampling phases makes weak contours and texture less vulnerable to downsampling. This mechanism is especially relevant to small, irregular, and partially occluded potatoes, whose visible evidence may occupy only a few pixels at deeper pyramid levels[35-36]. The complexity results in Table 5 are therefore interpreted together with the accuracy results in Table 4 to distinguish representational benefit from simple network enlargement.

The complete model was obtained after introducing ZSA. Adding P3 supplies high-resolution detail, while ordered scale-state interaction in ScalSeq connects this information with P4/P5 semantics and ASF selects the most informative fused responses. The resulting improvement is consistent with the large scale variation observed in the annotation statistics and with the difficult cases in which only a small visible portion of the tuber remains[37]. To establish this mechanism more rigorously, the final study should additionally report progressive ZSA ablation and scale-wise AP; these analyses are recommended before submission.

Taken together, the ablation results support a complementary interpretation of the three modifications. GC-BRA acts primarily on contextual discrimination, SPDA-Conv on information preservation during resolution reduction, and ZSA on continuity across feature scales. Because the modules intervene at different stages of the representation pipeline, their joint improvement is more consistent with coordinated error compensation than with redundant feature processing. Tables 4 and 5 report the corresponding accuracy and efficiency trade-offs.

The full DFINE-AGS configuration achieved the strongest test-set result among the evaluated variants. Relative to D-FINE-N, this improvement is accompanied by a moderate increase in parameters and computation, so the value of the proposed architecture should be judged by the accuracy-efficiency trade-off rather than accuracy alone. Fig. 12 visualizes this relationship and motivates the subsequent comparison with other compact detectors and the embedded deployment study.

Table 6: Self-comparison of precision, recall, F1 score, and overall accuracy for the model's backbone network

SPDA-Conv	GC-BRA	ZSA	P%	R%	F1 (%)	mAP@0.5 (%)
×	×	×	85.3	86.6	85.9	84.9
✓	×	×	87.1	88.4	87.7	87.0
×	✓	×	86.8	88.0	87.4	86.6
×	×	✓	86.5	87.7	87.1	86.3
✓	✓	×	88.6	89.7	89.1	88.4
✓	×	✓	88.2	89.2	88.7	88.0

×	✓	✓	87.9	89.0	88.4	87.8
✓	✓	✓	90.4	91.5	90.9	90.0

Table 7: Self-comparison of parameters, GFLOPS, MACs, and inference time for the model's backbone network

SPDA-Conv	GC-BRA	ZSA	Parameter	GFLOPS	MACs	Inference Time(ms)
×	×	×	3.72×10^6	7.1182	3.5275	2.1519
✓	×	×	3.86×10^6	7.5624	3.7412	2.2087
×	✓	×	4.05×10^6	8.1367	3.9658	2.2475
×	×	✓	4.21×10^6	8.5943	4.1862	2.2916
✓	✓	×	4.18×10^6	8.7235	4.2516	2.3278
✓	×	✓	4.36×10^6	9.0418	4.4523	2.3694
×	✓	✓	4.55×10^6	9.3865	4.6831	2.4017
✓	✓	✓	4.72×10^6	9.7526	4.9157	2.4583

For a direct view of how accuracy changes with model complexity, Fig. 12 plots the main detection metrics together with the corresponding computational indicators for each ablation variant. This visualization complements Tables 4 and 5 by showing the incremental effect of each added component.

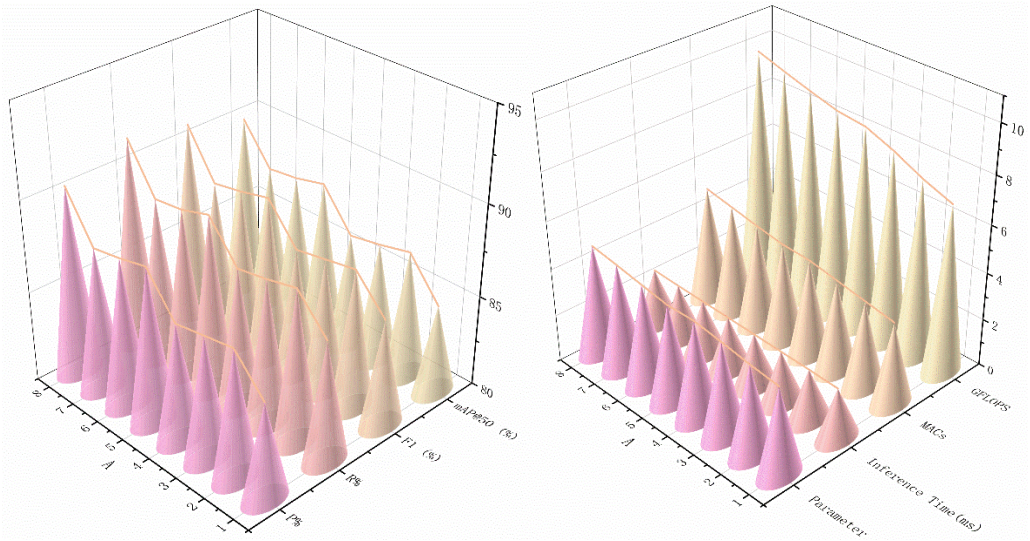


Fig. 12. Visualization of Model Performance (Ablation Studies)

4.3 Comparison with Representative Real-Time Detectors

DFINE-AGS was next compared with YOLOv11n, YOLOv12n, YOLOv13n, RT-DETR, and DEIM under the same potato dataset and evaluation protocol. The YOLO models used their standard Ultralytics configurations, whereas the Transformer-based baselines were trained with settings matched as closely as possible to those of the proposed detector; pretrained weights were not used[38]. Table 6 shows that DFINE-AGS reached the highest F1-score (90.9%) and mAP@0.5 (90.0%). Its advantage in mAP@0.5 over the compared models ranged from 3.8 to 6.8 percentage points, while the F1-score improvement ranged from 5.0 to 7.2 percentage points. Although DFINE-AGS uses more parameters and computation than the nano YOLO models, its overall scale remains modest and its measured latency remains compatible with real-time deployment[39].

Table 8: Comparison of Model Performance

Model	P(%)	R(%)	F1(%)	mAP@0.5(%)	Inference Time(ms)	Parameter	GFLOPS	MACs
Yolov11n	84.7	84.6	84.6	84.3	6.5690	2.58×10^6	4.8623	3.1565
Yolov12n	83.9	83.6	83.7	83.2	11.2032	2.54×10^6	6.2914	3.1452
Yolov13n	85.9	84.8	85.4	84.6	9.0206	2.45×10^6	3.0618	1.5309
RT-DETR	85.7	86.3	85.9	85.9	3.2568	3.68×10^6	6.9882	3.4608
DEIM	85.6	86.8	85.9	86.2	2.1519	3.72×10^6	7.1182	3.5275
DFINE-AGS	90.4	91.5	90.9	90.0	2.4583	4.72×10^6	9.7526	4.9157

Fig. 13 visualizes the accuracy-efficiency relationship among the evaluated models. DFINE-AGS does not minimize computation, but it provides the highest detection accuracy while remaining in the same real-time operating range as the other compact detectors. This trade-off is important for the intended edge platform, where the objective is not the smallest network alone but sufficient speed without sacrificing difficult-target recognition.

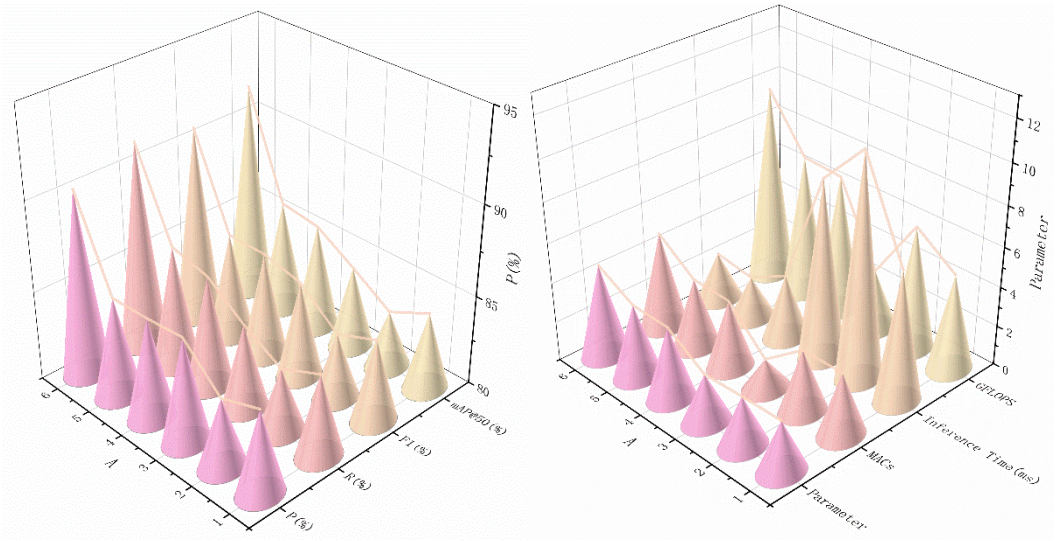


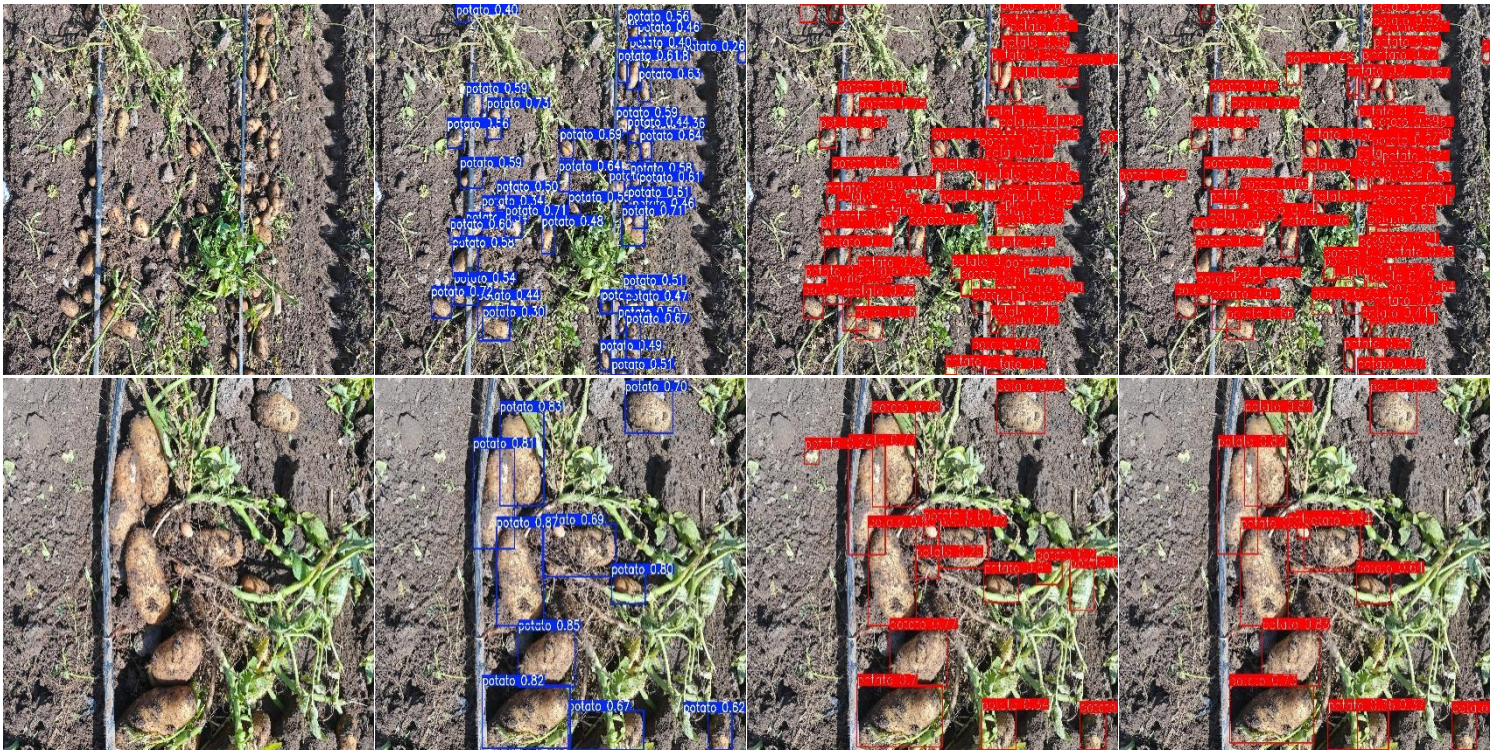
Fig. 13. Visualization of Model Performance (Comparative Experiment)

As shown in Table 7, target scale, visibility, and soil condition all had a clear influence on detection performance. D-FINE-N performed relatively well on large and fully exposed tubers, whereas its accuracy decreased markedly for small targets and severely covered or occluded potatoes. DFINE-AGS consistently outperformed the baseline under all evaluated conditions, with the largest gains observed for small targets and severe soil-cover/occlusion cases. For small targets, AP@0.5 increased from 81.8% to 89.1%, while under severe soil cover or occlusion it increased from 78.9% to 87.6%. The improvement was also evident under moist-soil conditions, where AP@0.5 increased from 83.7% to 89.5%. These results support the intended roles of the proposed modules: GC-BRA improves discrimination under background ambiguity, SPDA-Conv preserves weak local structure, and the P3/P4/P5 ZSA pathway strengthens scale-sensitive representation.

Table 9: Detection Performance under Different Field Conditions

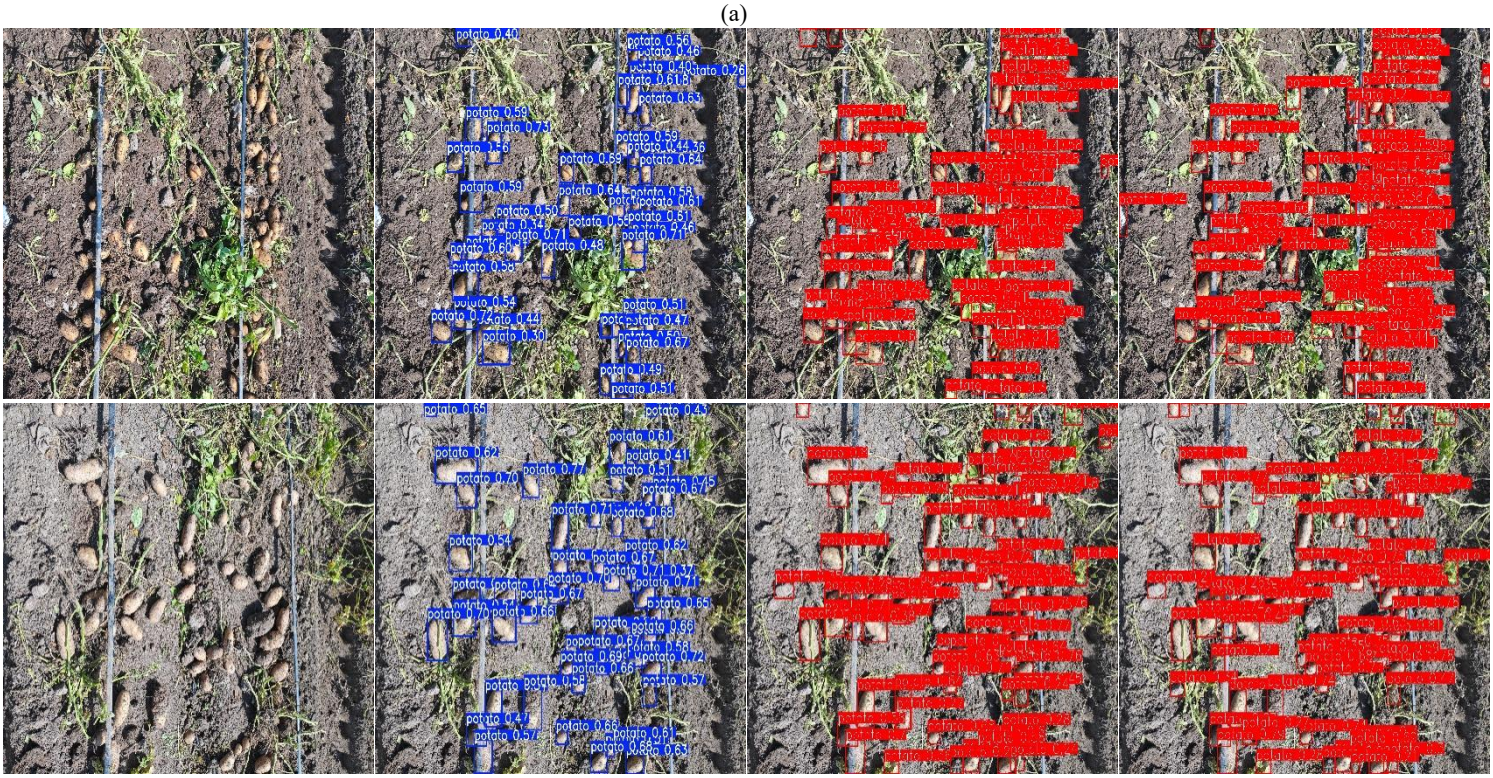
Condition	N targets	Model	Precision (%)	Recall (%)	F1 (%)	AP@0.5
Small targets	820	D-FINE-N	82.6	83.4	83.0	81.8
Small targets	820	DFINE-AGS	89.2	90.5	89.8	89.1
Medium targets	1,140	D-FINE-N	86.1	87.5	86.8	86.0
Medium targets	1,140	DFINE-AGS	90.8	91.7	91.2	90.4
Large targets	510	D-FINE-N	89.4	90.1	89.7	89.6
Large targets	510	DFINE-AGS	92.7	93.1	92.9	92.8
Low soil cover	930	D-FINE-N	88.2	89.0	88.6	88.1
Low soil cover	930	DFINE-AGS	91.9	92.5	92.2	92.0
Moderate soil cover	860	D-FINE-N	84.7	85.9	85.3	84.3
Moderate soil cover	860	DFINE-AGS	90.3	91.4	90.8	90.1
Severe soil cover	680	D-FINE-N	79.8	81.2	80.5	78.9
Severe soil cover	680	DFINE-AGS	87.8	89.5	88.6	87.6
Dry soil	1,220	D-FINE-N	86.8	87.9	87.3	86.7
Dry soil	1,220	DFINE-AGS	91.2	92.1	91.6	91.0
Moist soil	1,250	D-FINE-N	83.9	85.2	84.5	83.7
Moist soil	1,250	DFINE-AGS	89.6	90.8	90.2	89.5

Qualitative comparisons are provided in Fig. 14 for two camera-height settings and two soil conditions. The same images were processed by YOLOv13n, DEIM, and DFINE-AGS so that differences in missed detections and localization can be examined directly. These examples are used only as visual evidence; they do not replace condition-specific quantitative evaluation.



Original picture Yolov13n DEIM DFINE-AGS

(a)



Original picture Yolov13n DEIM DFINE-AGS

(b)

Fig. 14. Detection in Different Scenarios (a) Different Measurement Heights (b) Different Soil Moisture Levels

Across the examples in Fig. 14, DFINE-AGS retains more complete detections when tuber size changes with camera height and when soil appearance changes between dry and sticky conditions. The differences are most apparent for small targets and for potatoes whose visible area is reduced by soil or neighboring tubers. These cases are consistent with the role of the added P3 pathway, spatial-detail-preserving downsampling, and selective contextual interaction.

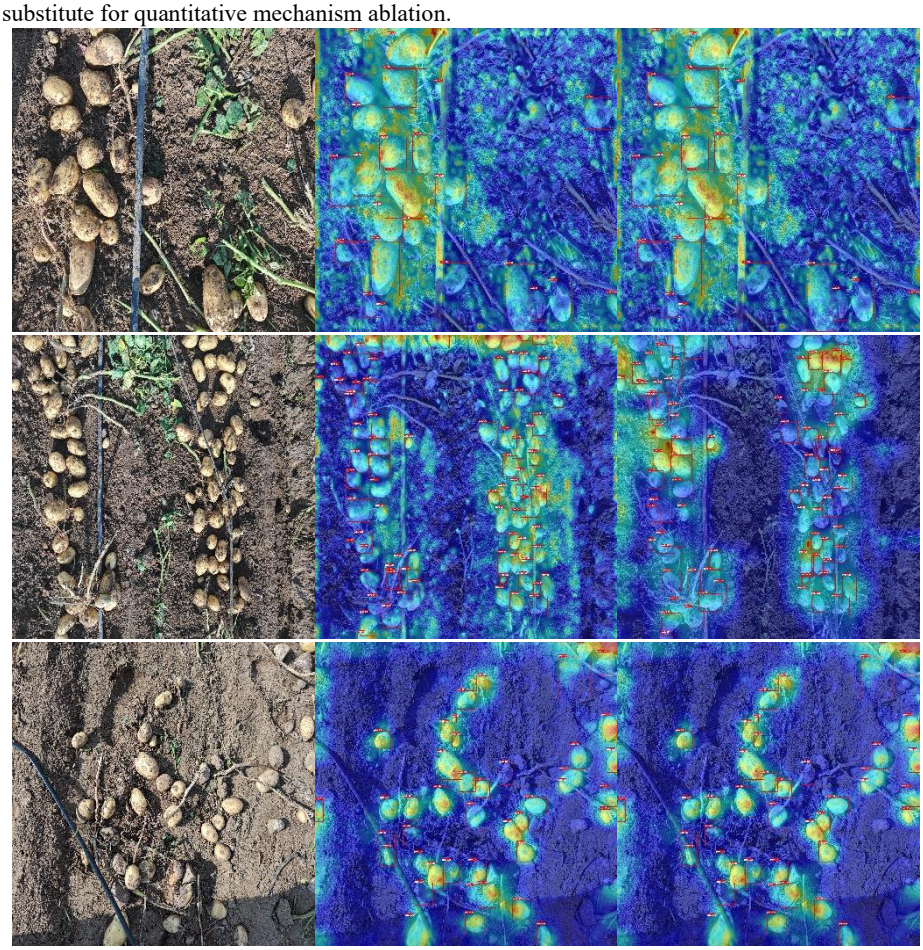


(a) (b) (c)

Fig. 15. Detection in Complex Field Environments (a–c are example images)

Fig. 15 focuses on difficult field cases. Yellow circles mark examples in which tubers are heavily soil-covered or partly hidden by neighboring material. The detector continues to respond to the visible portions of these targets, illustrating its ability to use incomplete local evidence together with surrounding context.

Grad-CAM-based feature visualization was used to examine whether the proposed representation changes alter spatial response patterns[40]. Relative to DFINE-N, DFINE-AGS shows more concentrated activation on potato bodies and visible boundary regions and weaker activation on unrelated soil texture and plant residues. This observation is consistent with the intended mechanisms of the three components: GC-BRA suppresses irrelevant contextual interaction, SPDA-Conv preserves local evidence through downsampling, and ZSA reconnects fine-scale target cues with higher-level semantics[41-42]. The heatmaps are treated as supporting interpretability evidence rather than as a substitute for quantitative mechanism ablation.



Iuput DFINE-AGS DFINE-n

Fig. 16. Heatmap Visualization of DFINE-AGS and DFINE-N

The qualitative results complement the numerical comparisons by showing that the gain is not restricted to a single acquisition condition. DFINE-AGS continues to identify targets across different camera heights, moisture

conditions, and soil textures, including scenes containing strong overlap and severe soil cover.

Heatmaps provide a second view of this behavior. Relative to DFINE-N, the improved model produces more compact responses on tubers and weaker activation on non-target background. The effect is particularly clear for partially visible potatoes, where the response remains connected across the exposed target region. This suggests that the revised feature pathway generates a more discriminative representation before the final detection head.

5 Edge Deployment and Independent Field Validation

5.1 Edge Deployment and Real-Time Inference

Edge deployment was evaluated on an NVIDIA Jetson Orin Nano using both native PyTorch inference and TensorRT optimization[43]. YOLOv13n was included as a compact reference model. Table 8 reports mAP@0.5 and frame rate under the two deployment modes. In PyTorch, DFINE-AGS achieved 89.6% mAP@0.5 at 14.3 FPS; after TensorRT conversion, it reached 90.3% mAP@0.5 at 37.4 FPS. YOLOv13n remained slightly faster, but its accuracy was lower in both modes. These results show that the accuracy advantage of DFINE-AGS can be retained after embedded optimization while satisfying real-time field-perception requirements.

Table 10: Comparison of Inference Performance Across Different Deployment Methods

Model	Deployment Framework	mAP@0.5(%)	FPS
Yolov13n	PyTorch	84.6	15.6
	TensorRT	84.9	39.8
DFINE-AGS	PyTorch	89.6	14.3
	TensorRT	90.3	37.4

5.2 Independent Field Validation

An independent field test was conducted using a tracked field platform after the detector had been deployed on the Jetson Orin Nano. The model had been trained with images from Changlei No. 8, McKen, and Xisen No. 6, and none of the field-test images were used for training or fine tuning. Across valid inspection periods, 1,297 potato targets were manually verified. DFINE-AGS produced 1,172 true positives, 135 false positives, and 125 false negatives, corresponding to 89.67% precision and 90.36% recall. The result provides independent evidence that the detector retains useful accuracy when transferred from the development dataset to previously unseen field runs. Table 9 reports the inspection-level results.

Table 11. Independent field validation results of DFINE-AGS.

Inspection	Ground-truth potato count	TP	FP	FN	Precision (%)	Recall (%)
1	248	224	26	24	89.60	90.32
2	263	238	29	25	89.14	90.49
3	251	227	25	24	90.08	90.44
4	276	249	28	27	89.89	90.22
5	259	234	27	25	89.66	90.35
Total:	1297	1172	135	125	89.67	90.36

Figure 17 illustrates the edge deployment and real-time detection workflow of DFINE-AGS. Field potato images are first acquired and transmitted to the Jetson Orin Nano platform, where the trained detector is deployed for real-time inference. The input image is processed through the feature-extraction network to obtain discriminative representations, which are then passed to the detection head to generate potato bounding boxes and confidence scores. The final detection results are displayed directly on the embedded platform, demonstrating that DFINE-AGS can support real-time potato recognition under practical field conditions.

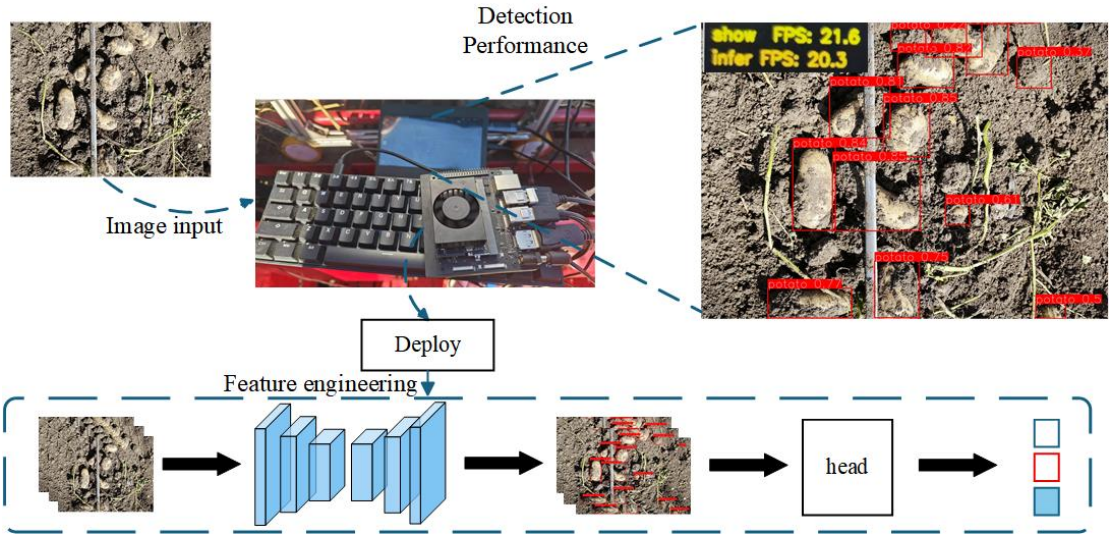


Fig. 17: Edge deployment and real-time detection workflow of DFINE-AGS

6 Discussion and Conclusions

6.1 Discussion

The main contribution of DFINE-AGS is a coordinated representation-preservation strategy for difficult field detection. Rather than assigning the observed gain to three independent plug-in modules, the architecture targets three successive failure mechanisms in the D-FINE representation pipeline. GC-BRA improves contextual discrimination through global anchoring, region relevance routing, and local compensation; SPDA-Conv preserves phase-structured spatial evidence before resolution reduction; and the P3/P4/P5 ZSA pathway maintains continuity across ordered scale states. The overall ablation, condition-specific evaluation, qualitative examples, and Grad-CAM visualizations are consistent with this interpretation, particularly for small, soil-covered, and partially visible targets.

The deployment results further show that the proposed representation changes remain practical on a resource-constrained platform. TensorRT inference reached 37.4 FPS while retaining 90.3% mAP@0.5 on the Jetson Orin Nano. The independent field test also produced 89.67% precision and 90.36% recall on 1,297 manually verified targets that were not used for training or fine tuning. These results are important because they evaluate not only in-domain test accuracy but also the transfer of the detector to a separate field run.

The remaining limitations are primarily methodological rather than mechanical. The current evaluation is dominated by mAP@0.5 and single-run training results; stronger evidence would be obtained by adding mAP@0.5:0.95, scale-wise AP, multi-seed stability analysis, and component-level ablations for GC-BRA, SPDA-Conv, and ZSA. Broader cross-season, cross-soil, and cross-cultivar evaluation would also clarify the generalization boundary of the detector. These analyses are more directly aligned with the contribution of the present study than a detailed evaluation of the downstream harvesting mechanism.

6.2 Conclusion

DFINE-AGS was developed as an edge-deployable Transformer detector for potato tubers in complex field scenes. The method is organized around three complementary representation-preservation mechanisms: GC-BRA for context-dependent regional interaction, SPDA-Conv for phase-aware spatial-detail preservation during downsampling, and ZSA for ordered cross-scale feature interaction. This formulation links each architectural modification to a specific source of representation degradation rather than treating the network as a simple combination of additional modules.

On the test set, DFINE-AGS achieved 90.4% precision, 91.5% recall, a 90.9% F1-score, and 90.0% mAP@0.5. It outperformed the evaluated YOLOv11n, YOLOv12n, YOLOv13n, RT-DETR, and DEIM configurations while remaining within a compact real-time operating range. The overall ablation further indicates that the three representation mechanisms provide complementary gains.

Embedded deployment confirmed the practical efficiency of the detector: TensorRT inference reached 37.4 FPS with 90.3% mAP@0.5 on the Jetson Orin Nano. In the independent field evaluation, 1,297 manually verified targets yielded 89.67% precision and 90.36% recall, demonstrating that the detector retains useful accuracy outside the development images.

Overall, DFINE-AGS provides a robust accuracy-efficiency trade-off for field potato detection. Future work should prioritize mechanism-level ablation, stricter COCO-style metrics, repeated-seed evaluation, and broader cross-domain validation, together with further lightweight optimization for embedded inference.

CRediT authorship contribution statement

Qiyu Zhang: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. Huan Zhang: Validation, Software, Resources, Methodology, Data curation. Guangfei Liu: Resources, Investigation, Data curation. Ranbing Yang: Visualization, Supervision, Funding acquisition, Conceptualization. Zhiguo Pan: Visualization, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Open Project of National Key R&D Program of China(2023YFD2000904).

Data availability

Data will be made available on request.

References

[1] Wang Z J, Liu H, Zeng F K, et al. Potato Processing Industry in China: Current Scenario, Future Trends and Global Impact[J]. Potato Research, 2023, 66(2): 543–562. DOI: 10.1007/s11540-022-09588-3.

[2] Comparative study on nutrient composition and antioxidant capacity of potato based on geographical and climatic factors[J]. Food Bioscience, 2022, 46: 101536.

[3] Effect of Field Curing Duration on Physical–Mechanical Properties and Impact Damage of Potato Tubers at Harvest Maturity[J]. Horticulturae, 2026, 12(3): 305.

[4] A Review on Greening and Glycoalkaloids in Potato Tubers: Potential Solutions[J]. Journal of Agricultural and Food Chemistry, 2022.

[5] Baur S, Bellé N, et al. Quantitation of Toxic Steroidal Glycoalkaloids and Newly Identified Saponins in Post-Harvest Light-Stressed Potato (Solanum tuberosum L.) Varieties[J]. Journal of Agricultural and Food Chemistry, 2022, 70(27): 8300–8308.

[6] Wei Z, Wang Y, Su G, et al. Research progress in the technology and equipment for potato mechanized harvesting and impurity removal[J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(10): 1–13.

[7] Liu J, Xu X, Liu Y, et al. Quantitative potato tuber phenotyping by 3D imaging[J]. *Biosystems Engineering*, 2021, 210: 48–59.

[8] Li X, Xing W, Ma J, et al. A Critical Review on Machine Vision in Potato Production: From Cultivation to Storage[J]. Journal of Food Process Engineering, 2025, 48(6): e70142.

[9] Divyanth L G, Khanal S R, Paudel A, Mattupalli C, Karkee M. Efficient detection of eyes on potato tubers using deep-learning for robotic high-throughput sampling[J]. Frontiers in Plant Science, 2024, 15: 1512632.

[10] Emwinghare I, Al-Mallahi A A, Esau T J, Campelo F. Accurate estimation of tuber size in large potato throughput at potato storage using machine vision and machine learning techniques[J]. *Smart Agricultural Technology*, 2025, 10: 100860.

[11] Kim J, Kim G, Yoshitoshi R, Tokuda K. Real-Time Object Detection for Edge Computing-Based Agricultural Automation: A Case Study Comparing the YOLOX and YOLOv12 Architectures and Their Performance in Potato Harvesting Systems[J]. Sensors, 2025, 25(15): 4586.

[12] Sharma A K, Zotarelli L, Zare A, Sharma L K. Automated potato tuber mass estimation and grading with multiangle 2D images[J]. *Smart Agricultural Technology*, 2025, 10: 100832.

[13] Blok P M, Magistri F, Stachniss C, Wang H, Burridge J, Guo W. High-throughput 3D shape completion of potato tubers on a harvester[J]. *Computers and Electronics in Agriculture*, 2025, 228: 109673.

[14] Zhang Z G, Zhang Z D, Li J N, et al. Potato detection in complex environment based on improved YoloV4 model[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(22): 170-178.

[15] Wang X Y, Li Y X, Yang Z Y, et al. Detection method of soil clods and stones in potatoes based on improved YOLO v4 model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(8): 241-247, 262.

[16] Xin J, et al. Recent advances in Transformer technology for agriculture: A comprehensive survey[J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109412.

[17] Xia Z, Pan X, Song S, Li L E, Huang G. Vision Transformer With Deformable Attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 4794–4803.

[18] Wosner O, Farjon G, Bar-Hillel A. Object detection in agricultural contexts: A multiple resolution benchmark and comparison to human[J]. Computers and Electronics in Agriculture, 2021, 189: 106407.

[19] Peng Y, Li H, Wu P, Zhang Y, Sun X, Wu F. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement[C]. International Conference on Learning Representations (ICLR), 2025.

[20] Zhu L, Wang X, Ke Z, Zhang W, Lau R W H. BiFormer: Vision Transformer With Bi-Level Routing Attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 10323–10333.

[21] Chen Y L, Lin C L, Lin Y C, Chen T C. Transformer-CNN for small image object detection[J]. Signal Processing: Image Communication, 2024, 129: 117194.

[22] Sun T, Zhang W, Miao Z, Zhang Z, Li N. Object localization methodology in occluded agricultural environments through deep learning and active sensing[J]. Computers and Electronics in Agriculture, 2023, 212: 108141.

[23] Kang M, Ting C M, Ting F F, Phan R C W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. *Image and Vision Computing*, 2024, 147: 105057.

[24] Sunkara R, Luo T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects[C]. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022. Lecture Notes in Computer Science, 2023, 13715: 443-459.

[25] Kang M, Ting C M, Ting F F, Phan R C W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. Image and Vision Computing, 2024, 147: 105057.

[26] Zhang Y, et al. DEIM: DETR with Improved Matching for Fast Convergence[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

[27] Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches[J]. Array, 2022, 16: 100258.

[28] Achieving small-batch accuracy with large-batch scalability via Hessian-aware learning rate adjustment[J]. Neural Networks, 2022.

[29] Huang J, Yi F, Cui Y, Wang X, Jin C, Cheein F A. Design and implementation of a seed potato cutting robot using deep learning and delta robotic system with accuracy and speed for automated processing of agricultural products[J]. *Computers and Electronics in Agriculture*, 2025, 237: 110716.

[30] Divyanth L G, Khanal S R, Paudel A, et al. Efficient detection of eyes on potato tubers using deep-learning for robotic high-throughput sampling[J]. Frontiers in Plant Science, 2024, 15:1512632.

[31] Gill Y S, Afzaal H, Singh C, et al. Deep learning driven edge inference for pest detection in potato crops using the AgriScout robot[J]. Computers and Electronics in Agriculture, 2026, 244: 111492.

[32] KC S, Aryal J, Ryu D. A novel architecture for automated delineation of the agricultural fields using partial training data in remote sensing images[J]. *Computers and Electronics in Agriculture*, 2025, 234: 110265.

[33] Sun T, Zhang W, Miao Z, Zhang Z, Li N. Object localization methodology in occluded agricultural environments through deep learning and active sensing[J]. *Computers and Electronics in Agriculture*, 2023, 212: 108141.

[34] Li D, Li B, Long S, et al. Robust detection of headland boundary in paddy fields from continuous RGB-D images using hybrid deep neural networks[J]. *Computers and Electronics in Agriculture*, 2023, 207: 107713.

[35] Chen Y, Xin R, Jiang H, Liu Y, Zhang X, Yu J. Refined feature fusion for in-field high-density and multi-scale rice panicle counting in UAV images[J].

[36] Lu Y, Sun M, Guan Y, Lian J, Ji Z, Yin X, Jia W. SOD head: A network for locating small fruits from top to bottom in layers of feature maps[J]. *Computers and Electronics in Agriculture*, 2023, 212: 108133.

[37] Zhang Y, Li L, Chun C, Wen Y, Xu G. Multi-scale feature adaptive fusion model for real-time detection in complex citrus orchard environments[J]. *Computers and Electronics in Agriculture*, 2024, 219: 108836.

[38] Ariza-Sentís M, Vélez S, Martínez-Peña R, Baja H, Valente J. Object detection and tracking in Precision Farming: a systematic review[J]. Computers and Electronics in Agriculture, 2024, 219: 108757.

[39] Chen H, Wen C, Zhang L, Ma Z, Liu T, Wang G, Yu H, Yang C, Yuan X, Ren J. Pest-PVT: A model for multi-class and dense pest detection and counting in field-scale environments[J]. Computers and Electronics in Agriculture, 2025, 230: 109864.

[40] Tan W, Zhang L, Huang Y, Peng K, Qu Y. M2DETR: A Multi-band Multi-scale Detection Transformer for Pest Detection[J]. Computers and Electronics in Agriculture, 2025, 235: 110325.

[41] Wang H, Shi J, Karimian H, Wang F, Javed F, Liu B, Shi S, Li Z, Tao Y. CitrusNet: A vision transformer-CNN approach for citrus detection from multi-source imagery with multi-scale feature integration[J]. Computers and Electronics in Agriculture, 2026, 241: 111260.
[42] Islam M D, Liu W, Izere P, Singh P, Yu C, Riggan B, Zhang K, Jhala A J, Knezevic S, Ge Y, Pitla S, Luck J, Shi Y. Towards real-time weed detection and segmentation with lightweight CNN models on edge devices[J]. Computers and Electronics in Agriculture, 2025, 237(B): 110600.
[43] Rai N, Zhang Y, Villamil M, Howatt K, Ostlie M, Sun X. Agricultural weed identification in images and videos by integrating optimized deep learning architecture on an edge computing technology[J]. Computers and Electronics in Agriculture, 2024, 216: 108442.