SOFIA: SELF-ADAPTIVE OPTIMIZATION FOR FEDERATED LEARNING WITH INPAINTING-BASED ANOMALY-AWARE PRETRAINING FOR ROBUST DEFECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual defect detection is critical for quality control in modern manufacturing, where data are often distributed across multiple production sites and subject to strict privacy constraints. In industrial settings such as pharmaceutical production, these challenges are further amplified by limited defect samples and strict data-sharing restrictions, making centralized training impractical while standard federated learning struggles with heterogeneous data distributions and limited defect samples. We introduce SOFIA, a privacy-preserving federated defect detection framework that jointly addresses these challenges by combining client-adaptive hyperparameter optimization based on Horse Self-adaptive Optimization with object-oriented inpainting-based pretraining. The former adapts local training to client-specific data characteristics, while the latter generates semantically localized synthetic defects to improve representation learning under limited data. Experiments on pharmaceutical and metal surface defect detection benchmarks show that SOFIA improves detection performance over centralized and federated baselines. Ablation results further indicate that the two mechanisms play complementary roles: adaptive optimization improves federated training, while inpainting-based pretraining provides its largest benefit when integrated with the federated optimization framework. We further study how SOFIA behaves as the number of participating clients grows, finding consistent gains in performance and stability as more clients participate, suggesting favorable scalability as federation size increases. These findings highlight the importance of jointly addressing optimization and representation learning for scalable and reliable visual defect detection in distributed, privacy-constrained industrial settings.

1 INTRODUCTION

Visual defect detection is a key component of modern manufacturing, supporting automated quality control in applications ranging from metal surface inspection to electronics and pharmaceutical production Ren et al. (2022); Qiao et al. (2025); Mordia & Verma (2022); Srihari et al. (2025); Xu et al. (2025); Vijayakumar et al. (2025). While recent convolutional and transformer-based detectors achieve strong performance Zhou et al. (2023a); Shang et al. (2023), their success typically relies on sufficiently large and centralized training datasets. In industrial environments, however, data are often distributed across production sites, cannot be freely shared, and contain relatively few examples of rare defects. These constraints make conventional centralized training difficult and simultaneously introduce two challenges: learning across heterogeneous clients and learning useful defect representations from limited data.

Federated Learning (FL) provides a natural mechanism for collaborative training without exchanging raw data, but its performance can degrade when client distributions differ Ye et al. (2023). At the same time, the scarcity and imbalance of defect samples make representation learning particularly difficult Chen et al. (2021). Existing work has considered adaptive optimization in FL Chen et al. (2024); Khan et al. (2025); Srinivas et al. (2025) and reconstruction-based learning under limited data Zhang et al. (2023), but these directions are largely studied independently. In industrial defect

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

detection, however, the two problems are tightly coupled: unstable local optimization and weak representations can jointly limit the quality of the resulting global model.

We introduce SOFIA, a federated defect detection framework that jointly addresses these two sources of difficulty. SOFIA combines client-adaptive hyperparameter optimization based on Horse Self-adaptive Optimization with object-oriented inpainting-based pretraining. The former adapts local training configurations to client-specific data characteristics, while the latter generates semantically localized synthetic defects by modifying object-relevant regions rather than relying on purely random masking. The resulting representations and client-specific hyperparameters are then integrated during federated detector training. Figure 1 provides an overview of the proposed pipeline. Although motivated by pharmaceutical inspection, we evaluate the approach on both pill and metal-surface defect detection to study its behavior across different industrial domains.

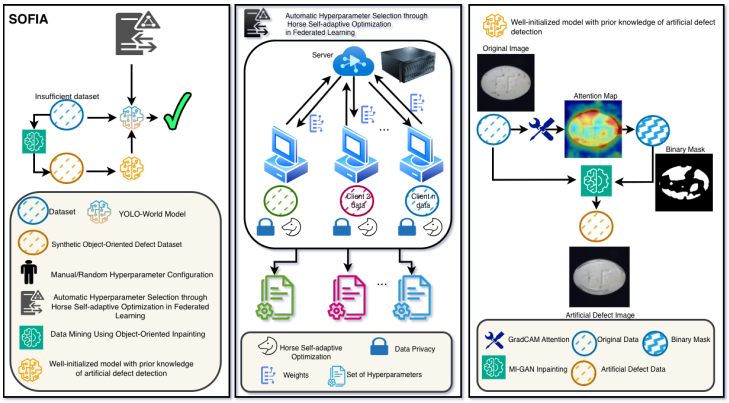

Figure 1: Defect detection under limited and distributed data. Centralized approaches suffer from data scarcity and poor generalization, while SOFIA improves robustness through federated optimization and object-oriented inpainting.

Our experiments show that the two components play complementary roles. Adaptive optimization improves consistency across federated clients, whereas object-oriented inpainting provides its largest benefit when combined with the federated optimization framework. We additionally analyze the effect of communication rounds and federation size: performance exhibits diminishing returns after a small number of rounds, while experiments with increasing numbers of participating clients show improved detection performance and reduced variability as the federation grows. These analyses complement the comparison against centralized and federated baselines and provide evidence about when and why the proposed components are effective. Our main contributions are: 1) We introduce SOFIA, a unified framework that combines client-adaptive federated optimization with object-oriented inpainting-based pretraining for visual defect detection under distributed and limited data. 2) We develop a client-level Self-adaptive optimization strategy for adapting local training hyperparameters and an object-aware inpainting mechanism for generating semantically localized synthetic defects. 3) We conduct an extensive evaluation on pharmaceutical and metal-surface defect detection, including comparisons with centralized and federated baselines, component ablations, communication-round analysis, and experiments with increasing numbers of participating clients.

2 Related Work

Federated learning and adaptive optimization. Federated Learning (FL) enables collaborative training across distributed sites without exchanging raw data and has been applied to medical imaging and other vision tasks where data sharing is restricted Ng et al. (2021); Nandal et al. (2025); Hernandez-Ramos et al. (2025); Li et al. (2025). However, heterogeneous client distributions can degrade convergence and generalization, particularly when standard aggregation schemes such as FedAvg are used Guo et al. (2022). Adaptive strategies based on client weighting, learning-rate adjustment, and personalized aggregation have been proposed to mitigate these effects Giuseppi et al. (2021); Gao et al. (2023). Swarm-intelligence methods, including Particle Swarm Optimization and

Grey Wolf Optimization, have also been explored for client selection and hyperparameter tuning in federated settings Khan & Goodridge (2024).

Inpainting and representation learning for defect detection. Limited labeled data has motivated self-supervised and reconstruction-based approaches for visual inspection. Masked reconstruction and inpainting methods learn representations by predicting missing or corrupted image regions Huang et al. (2023); He et al. (2022); Zhou et al. (2023b); Tran et al. (2021) and have been successfully applied to visual anomaly detection Zavrtanik et al. (2021); Nakanishi et al. (2022); Du et al. (2024). Existing approaches typically use random, iterative, or anomaly-guided masking strategies to improve reconstruction of normal structure and discrimination of anomalous regions. However, they generally do not explicitly exploit the structured, object-centric regions that characterize many industrial inspection tasks.

Joint treatment of heterogeneity and limited data. Existing work largely treats federated optimization under heterogeneous data and representation learning under limited supervision as separate problems. In industrial defect detection, these challenges arise simultaneously. Our work investigates their joint treatment by combining client-adaptive federated optimization with object-oriented inpainting-based pretraining.

3 THE SOFIA FRAMEWORK.

SOFIA shares only model weights and client-specific hyperparameters; raw images and synthetic defect images are generated and used locally at each client and are never transmitted. This follows the standard federated learning protocol. As illustrated in Figure 1, SOFIA consists of three phases that combine client-adaptive federated optimization with object-oriented inpainting-based pretraining.

In Phase 1, training data are distributed across heterogeneous clients participating in a federated learning setup. To mitigate the heterogeneity of local data distributions, each client performs an independent hyperparameter search. Specifically, at each communication round, each client independently applies our proposed Horse Self-adaptive Optimization (HSO), a simplified variant of the Wild Horse Optimizer (WHO) Naruei & Keynia (2022), to identify a local training configuration adapted to its data distribution.

HSO simplifies the original WHO in three ways: (i) it removes the recombination/mating step entirely; (ii) it selects a single member $\mathbf{h} \in \mathcal{P}_k$ as the current solution and refines it across the inner iterations, rather than updating the entire population simultaneously; and (iii) it avoids the overhead of hierarchical leadership and inter-herd communication. These simplifications reduce the per-client computational cost and make HSO suitable for resource-constrained federated clients.

HSO differs from independent random search in three key respects: 1) HSO selects a current solution $\mathbf{h} \in \mathcal{P}_k$ and refines it iteratively, accepting a candidate $\mathbf{h}'$ only when it improves fitness ($f(\mathbf{h}') \geq f(\mathbf{h})$). The refined solution is written back into $\mathcal{P}_k$, so the population is progressively improved across iterations. Random search, in contrast, samples configurations independently and retains no information from previous evaluations. 2) HSO generates candidates by perturbing the current solution, $\mathbf{h}' \leftarrow \mathcal{N}(\mathbf{h})$, thereby concentrating the search around promising regions of $\mathcal{H}$. Random search samples uniformly over the entire space without such a guiding mechanism. 3) HSO maintains a population $\mathcal{P}_k$ of candidate configurations that is progressively refined across iterations, and selects the final configuration $\mathbf{h}_k^*$ via $\arg\max_{\mathbf{h} \in \mathcal{P}_k} f(\mathbf{h})$. This combines diverse initialization with iterative refinement and population-level final selection. The resulting client-specific hyperparameters are retained for the final federated training phase. The search space is defined as $\mathcal{H} = \{\texttt{epochs}, \texttt{batch}, \texttt{lr0}, \texttt{momentum}, \texttt{weight_decay}, \texttt{warmup_epochs}\}$. For a given configuration $\mathbf{h} \in \mathcal{H}$, the objective function is defined as:

$$f(\mathbf{h}) = \text{mAP} + \text{mAP}_{50} + \text{Precision} + \text{Recall}, \tag{1}$$

and optimized through iterative neighborhood exploration $\mathbf{h}' = \mathcal{N}(\mathbf{h})$, where $\mathcal{N}(\cdot)$ denotes a stochastic perturbation operator. After convergence, each client obtains its own optimal hyperparameter set $\mathbf{h}^*$, which reflects its local data distribution. These optimized hyperparameters are preserved and later utilized in Phase 3 to guide the final training process.

We adopt FedAvg as a standard and communication-efficient aggregation rule, allowing us to isolate the effects of client-adaptive HPO and OOI without introducing additional aggregation mechanisms. Prior studies establish it as a simple, communication-efficient, and robust baseline for decentralized training Guo et al. (2026) with resilience to client dropout Adel et al. (2026) and practical deployability Tudor et al. (2026).

Algorithm 1 Federated Learning with Horse Self-adaptive Optimization (FL-HSO) (Phase 1)

Input: Clients $\mathcal{K} = \{1, \ldots, K\}$, hyperparameter space $\mathcal{H}$, rounds R
Initialize: Global model $w^{(0)}$
for $r = 1$ to R **do**
  for each client $k \in \mathcal{K}$ **in parallel do**
    Initialize local hyperparameter population $\mathcal{P}_k \subset \mathcal{H}$ and evaluate fitness $f(\mathbf{h})$ for each $\mathbf{h} \in \mathcal{P}_k$
    for $t = 1$ to T **do**
      $\mathbf{h}' \leftarrow \mathcal{N}(\mathbf{h})$ {Generate neighbor}
      Evaluate $f(\mathbf{h}')$
      $\mathbf{h} \leftarrow \mathbf{h}'$ if $f(\mathbf{h}') \geq f(\mathbf{h})$ {Update population}
    end for
    $\mathbf{h}_k^* \leftarrow \text{argmax}_{\mathbf{h} \in \mathcal{P}_k} f(\mathbf{h})$ {Select best hyperparameters}
    $w_k^{(r)} \leftarrow \text{LocalTrain}(w^{(r-1)}, \mathbf{h}_k^*)$; send $w_k^{(r)}$ to server
  end for
  $w^{(r)} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_k^{(r)}$ {Aggregate global model}
  broadcast $w^{(r)}$ to all clients
end for
Output: Final global model $w^{(R)}$ and optimal hyperparameters $\{\mathbf{h}_k^*\}_{k=1}^{K}$

In Phase 2, SOFIA uses object-oriented inpainting to generate a synthetic defect dataset for pretraining. As shown in Figure 1, starting from original defect-free images, we compute attention maps to identify semantically meaningful regions. These maps are converted into binary masks that guide the inpainting process to introduce artificial defects. Using MI-GAN inpainting with GradCAM++ attention, the model produces defect patterns that alter local textures while preserving global structure. The binary masks define where defects are placed, and the inpainting fills those regions with synthetic defect content. By masking and reconstructing image regions, the model learns structural and contextual patterns, improving its ability to capture subtle defects and irregularities. MI-GAN is pretrained on Places2 Zhou et al. (2017) and FFHQ Karras et al. (2019) datasets, which cover natural scenes and human faces. These domains differ substantially from industrial surface imagery. When applied to defect-free industrial images, the pretrained model produces inpainted regions that are inconsistent with the local surface context. Rather than synthesizing realistic defects, Object-Oriented Inpainting leverages this domain gap to generate localized anomalies: GradCAM++ derived masks restrict inpainting to object-relevant regions, and the resulting mismatch provides a proxy signal for anomalous patterns. We acknowledge that these synthetic anomalies do not necessarily resemble real defects; they serve as irregularities that encourage the detector to learn more discriminative object-centric representations. We adopt an inpainting-based learning strategy inspired by MI-GAN Sargsyan et al. (2023) to enhance the model's ability to capture structural and contextual features for defect detection. Given an original image $I_{\text{orig}} \in \mathbb{R}^{H \times W \times 3}$ and a binary mask $M \in \{0, 1\}^{H \times W}$ indicating known (1) and missing (0) regions, the corrupted input is constructed as:

$$I_n = [I_{\text{orig}} \odot M,\ M], \tag{2}$$

where $\odot$ denotes element-wise multiplication. The generator G_θ produces a completed image I_c, which is then combined with the original image to obtain the final reconstruction:

$$I_{\text{comp}} = I_{\text{orig}} \odot M + I_c \odot (1 - M). \tag{3}$$

The MI-GAN architecture consists of two complementary branches: (i) a U-Net-like main branch that extracts hierarchical features using depthwise separable convolutions, and (ii) a painting branch that progressively reconstructs missing regions in a coarse-to-fine manner. Specifically, intermediate outputs $\{x_i\}$ are generated at multiple scales and aggregated as:

$$\hat{x}_i = \text{Up}(\hat{x}_{i+1}) + x_i, \quad i = 5, \ldots, 0, \tag{4}$$

$$I_c = \hat{x}_0, \tag{5}$$

where $\mathrm{Up}(\cdot)$ denotes bilinear upsampling. The training objective combines adversarial learning and knowledge distillation. The adversarial loss is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}_{(x,m)} \left[\log(1 + e^{-D(G_\theta(x,m),m)}) \right], \tag{6}$$

where D is the discriminator. To further enhance reconstruction quality, a knowledge distillation loss is introduced between the student (MI-GAN) and a teacher network:

$$\mathcal{L}_{KD} = \sum_i \left\| (x_i - x_i^T) \odot (1 - M_i) \right\|, \tag{7}$$

where x_i^T denotes the teacher output and M_i is the resized mask at scale i. The final objective is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{KD}. \tag{8}$$

This pretraining strategy enables the model to learn meaningful visual representations by reconstructing missing regions, which is particularly beneficial for detecting subtle and irregular defects. The pretrained weights obtained in this phase are subsequently utilized in the next phase to improve detection performance under limited and heterogeneous data conditions.

In Phase 3, the pretrained model and the client-specific optimal hyperparameters obtained from Phase 1 are jointly utilized to train the detection model within the federated framework. The detection backbone is based on YOLO-World Cheng et al. (2024), which reformulates object detection as a region–text matching problem $\Omega = \{B_i, t_i\}_{i=1}^{N}$. Given an image I and text prompts T, the model predicts bounding boxes $\{\hat{B}_k\}$ and object embeddings $\{e_k\}$, while text embeddings are denoted as $W = \{w_j\}$. The similarity between object regions and textual representations is computed as:

$$s_{k,j} = \alpha \cdot \frac{e_k}{\|e_k\|_2} \cdot \frac{w_j^\top}{\|w_j\|_2} + \beta, \tag{9}$$

while cross-modal feature fusion is enhanced via $X_l' = X_l \cdot \delta\left(\max_j(X_l w_j^\top)\right)$, where $\delta(\cdot)$ denotes the sigmoid function. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda(\mathcal{L}_{iou} + \mathcal{L}_{dfl}). \tag{10}$$

This framework enables adaptive, privacy-preserving, and high-performance defect detection across heterogeneous client environments by combining federated optimization, synthetic defect awareness pretraining, and open-vocabulary detection.

4 RESULTS.

In this section, we evaluate the proposed framework on multiple industrial defect detection datasets, reporting quantitative results, qualitative comparisons, and ablation analyses.

4.1 EXPERIMENTAL SETTINGS.

We evaluate SOFIA[1] on two datasets: (i) the Pill and Capsule Defect Detection dataset from Roboflow Asma, and (ii) the NEU Surface Defect Database He et al. (2020). The Pill dataset contains 2,823 training, 121 validation, and 122 test images. NEU-DET contains 1,800 grayscale images across six defect categories, with 300 samples per category. We follow the standard 70/20/10 split used in prior work and, in the federated setting, partition the training data evenly across clients.

Federated hyperparameter optimization is computationally expensive, particularly when jointly varying image resolution, federation size, and communication rounds Shenaj et al. (2023). Since all experiments were conducted on a single NVIDIA A4500 GPU, using one configuration for every analysis was not computationally feasible. We therefore adapt the experimental configuration

[1]The source code is available at `https://anonymous.4open.science/r/tayninh`

to the objective of each analysis, while reporting all differences explicitly in Table 1. The primary controlled comparisons use 3 clients and 3 communication rounds. The NEU-DET comparison with prior work uses 640×640 images to match the corresponding evaluation setting, whereas the communication-round and scalability analyses use 224×224 images to make experiments over 1–5 rounds and 5–10 clients computationally feasible. Component ablations are performed on the Pill dataset at 224×224 for the same reason.

Table 1: Summary of experimental settings.

Experiment	Dataset	Configuration	Purpose
Table 2	Pill	224×224; 3 clients, 3 rounds	Main comparison
Figure 2a	NEU-DET	640×640; Client 3	SOTA comparison
Figure 2b	NEU-DET	640×640; discrete/continuous HPO	HPO-space ablation
Figure 3 / Table 3	Pill	224×224; test set	Qualitative / OOI ablation
Figure 4	Pill	224×224; test set	HSO ablation
Table 4 / Figures 5–6	NEU-DET	224×224; 1–5 rounds; 5–10 clients	Round/client scaling

We evaluate both discrete and continuous HPO spaces, following representative discrete and continuous formulations from prior work Anand Kumar & Sountharrajan (2025); Lee et al. (2025); Gill & Rajendran (2026). The complete search ranges are reported in Appendix A. Figure 2b shows that the discrete space achieves better overall performance while also reducing search cost and improving reproducibility; we therefore adopt it for the remaining experiments. The 3-client, 3-round configuration is used for controlled baseline comparisons, while the extended NEU-DET experiments explicitly study behavior under larger federations and longer communication horizons. More details about the used hyperparameters can be found in Appendix A.1.

4.2 Performance Evaluation.

We evaluate the detection performance of the proposed method across multiple clients and datasets. Table 2 reports mAP50 across clients and classes. SOFIA achieves the highest performance in every reported setting, with overall mAP50 reaching 0.834, 0.872, and 0.786 for Clients 1–3, respectively. The gains across all three data partitions indicate consistent performance under distributed training.

Table 2: Comparison of detection performance across different clients and classes.

Method	Client 1		Client 2		Client 3	
	Damage	All	Damage	All	Damage	All
Wang et al. (2024b)	0.819	0.757	0.780	0.740	0.764	0.744
Wang et al. (2024a)	0.808	0.751	0.818	0.764	0.788	0.732
Khanam & Hussain (2024)	0.845	0.812	0.836	0.789	0.817	0.752
Tian et al. (2025)	0.828	0.787	0.840	0.782	0.779	0.727
Zhao et al. (2024)	0.654	0.669	0.771	0.745	0.808	0.754
Ours	**0.890**	**0.834**	**0.921**	**0.872**	**0.834**	**0.786**

These results suggest that the proposed framework provides stable detection performance across distributed training settings. In contrast, some baselines, such as Zhao et al. (2024), exhibit more variable performance across clients, indicating reduced consistency under the same conditions. Overall, the results demonstrate that SOFIA improves detection accuracy while maintaining stable performance across different data partitions.

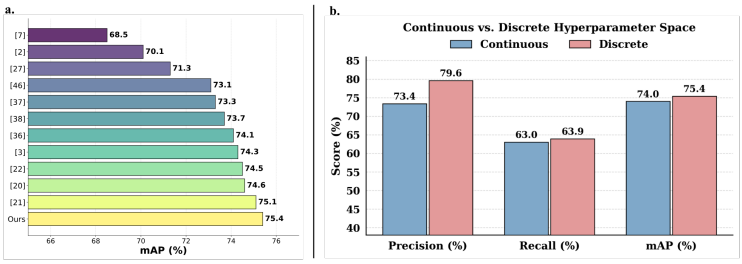

Figure 2: SOFIA's performance on NEU-DET dataset. **(a)** mAP (%) comparison with prior methods. **(b)** Discrete versus continuous HPO spaces across Precision, Recall, and mAP evaluation metrics.

Figure 2a compares SOFIA with recent detectors on NEU-DET. Despite using only one-third of the training data, SOFIA reaches 75.4% mAP, slightly exceeding the strongest competing result

(75.1%). Figure 2b further shows that the discrete HPO space consistently outperforms the continuous alternative, motivating its use in subsequent experiments.

These results indicate that the proposed framework achieves competitive detection performance while requiring substantially less training data, making it well-suited for data-constrained industrial scenarios. These findings highlight the benefit of combining adaptive optimization with object-oriented representation learning for improving defect detection under limited data.

4.3 Qualitative Analysis.

Figure 3 presents a qualitative comparison between the proposed method and competing approaches on pill defect detection, a critical scenario where accurate defect identification directly impacts patient safety and manufacturing compliance. The results show that SOFIA produces more accurate and consistent predictions, particularly in challenging scenarios involving small objects, low contrast, and cluttered backgrounds.

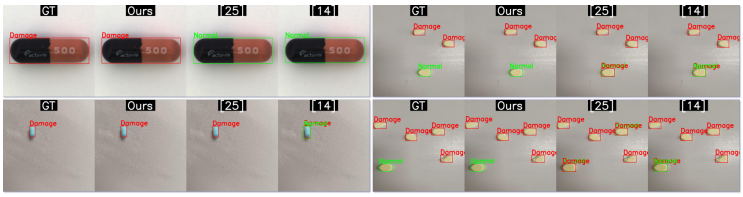

Figure 3: Qualitative Comparison of Pill Defect Detection Performance

Compared to Khanam & Hussain (2024) and Tian et al. (2025), the proposed method better distinguishes between Normal and Damage classes, aligning more closely with ground truth annotations. In several cases where competing models misclassify damaged items or incorrectly label normal samples, SOFIA maintains correct predictions, indicating improved feature representation and decision boundaries. Furthermore, SOFIA produces more precise and stable bounding box localizations, especially in scenes with multiple instances, where competing methods often generate noisy or overlapping detections. These observations suggest more reliable predictions and support the applicability of the approach to industrial inspection tasks.

4.4 Ablation Study.

Table 3: Isolating OOI's contribution to detection performance across Normal, Damage, and All categories.

Method	Normal	Damage	All
CL	0.799	0.888	0.843
CL + OOI	0.794	0.885	0.839
FL + HSO	0.782	0.901	0.842
FL + HSO + OOI (Ours)	**0.822**	**0.921**	**0.872**

In this section, we analyze the contribution of key components of the proposed framework, focusing on Federated Learning (FL), Horse Self-adaptive Optimization (HSO), and Object-Oriented Inpainting (OOI). In our setting, the Centralized Learning (CL) baseline refers to models trained independently on each client's local data, without any form of federated aggregation.

Table 3 isolates the contribution of OOI. Neither OOI with centralized training (0.839 vs. 0.843) nor FL+HSO alone (0.842) improves the overall baseline. Their combination, however, increases mAP50 to 0.872 and improves both classes, indicating a complementary interaction between the two components.

Figure 4 further isolates HSO. FL alone provides moderate improvements over CL but remains sensitive to data partitioning. In comparison, FL+HSO improves performance consistently across clients and metrics. For the Damage class, mAP50 increases from approximately 0.82 to 0.871 and recall from 0.86 to 0.909, with similar improvements observed under mAP95. Similar improvements are observed across other classes and are maintained under stricter localization criteria (mAP95), indicating improved bounding box precision.

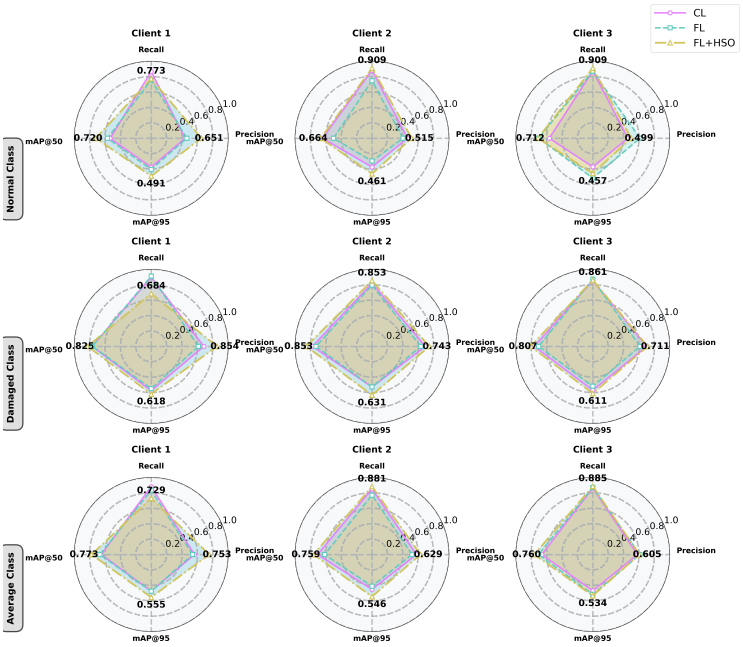

Figure 4: HSO ablation on the Pill dataset across clients and defect classes. Each radar plot compares CL, FL, and FL+HSO using Precision, Recall, mAP50, and mAP95.

Overall, the ablation study shows that HSO improves performance consistency across clients, while OOI enhances representation learning, with their combination yielding the strongest performance.

Ablation on Communication Rounds. In the optimal-parameter search phase, we ablate the number of communication rounds to determine an appropriate stopping point (See Fig. 5). The first graph reports mean $\pm 95\%$ CI over rounds for 5 to 10 clients. Round 3 already reaches near-saturated performance, while rounds 4 and 5 yield only marginal changes despite additional communication and computation. We therefore select 3 communication rounds as the default setting for the hyperparameter-finding phase. Aggregate performance values across all rounds are reported in Table 9 in Appendix B.1.

Ablation on Client Scalability. With communication fixed to round 3, the second analysis evaluates performance as a function of federation size (see Fig. 6). Across 5–10 clients, both mAP50 and mAP95 show an overall upward trend, while the confidence intervals narrow as more clients participate. These results show improved performance and reduced variability within the evaluated range, indicating improved generalization as federation size increases. They also support the viability of expanding the federated setup and suggest strong potential for scaling to larger federations.

5 Conclusion.

In this work, we introduced SOFIA, a unified framework for privacy-preserving visual defect detection that combines federated learning, client self-adaptive hyperparameter optimization, and object-oriented inpainting-based pretraining. Experiments on pharmaceutical and metal surface datasets show that SOFIA improves detection performance across distributed clients while remaining effective under limited training data.

Our analyses further show that the proposed components play complementary roles. HSO improves performance consistency across clients, while object-oriented inpainting provides its strongest benefit when integrated with the federated optimization framework. Moreover, the communication-round analysis reveals diminishing returns beyond a small number of rounds, while experiments with 5–

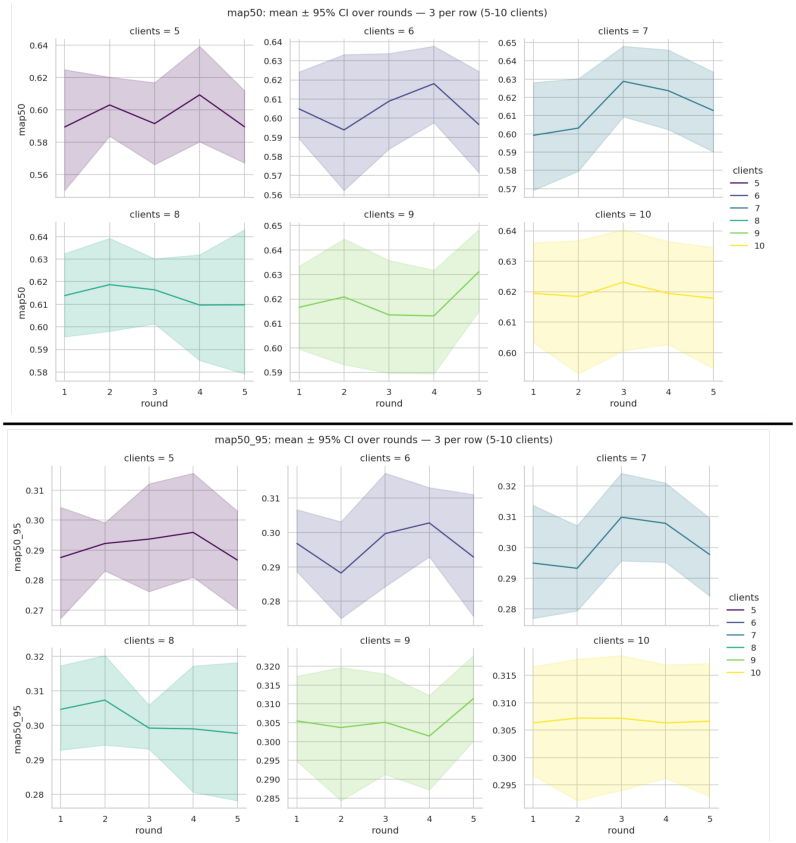

Figure 5: Effect of communication rounds on mAP50 (top) and mAP95 (mAP50-95, bottom) for federation sizes from 5 to 10 clients. Shaded regions denote 95% confidence intervals.

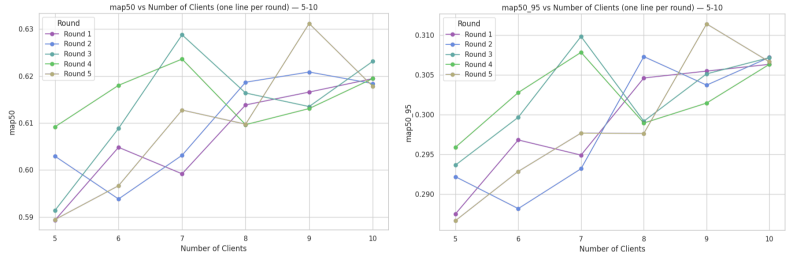

Figure 6: Client-scalability analysis showing mAP50 (top) and mAP95 (mAP50-95, bottom) as the federation size increases from 5 to 10 clients.

10 clients show increasing detection performance and reduced variability as federation size grows, providing encouraging evidence of scalability.

The evaluation nevertheless relies on simulated federated partitions of two benchmark datasets rather than real multi-site industrial deployments, and the hyperparameter optimization stage remains computationally demanding. Future work will therefore investigate SOFIA under broader forms of client and system heterogeneity, reduce the computational cost of the optimization process, and explore more advanced generative approaches for defect synthesis.

GENERATIVE AI DISCLOSURE

Generative AI tools were used during the preparation of this manuscript to assist with drafting and revising text, improving readability and structure, and obtaining feedback on the presentation of the research methodology, experimental setup, and results. Generative AI was not used to design the research methodology or experiments, generate datasets or experimental results, interpret the experimental findings, implement the proposed methods, or formulate or prove mathematical claims. All AI-assisted content was critically reviewed, verified, and revised by the authors. The authors take full responsibility for the final content of this work, including any text or other material produced with the assistance of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or the collection of sensitive personal data. The experiments are conducted on existing datasets for visual defect detection and are intended to study machine learning methods for industrial quality-control applications. While the proposed federated learning setting reduces the need to exchange raw client data, we do not claim that it provides formal privacy guarantees against all possible attacks; additional mechanisms such as differential privacy or secure aggregation may be required in deployments with stronger privacy requirements. The authors have considered the potential implications of the proposed methodology and are not aware of additional ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary to facilitate reproduction of our results throughout the paper and supplementary material. The proposed framework and its main components are described in detail in the methodology section, while the experimental setup specifies the datasets, client configurations, training procedure, evaluation protocol, and baselines used in our experiments. Additional implementation details, dataset statistics, and the complete client-specific hyperparameter configurations selected during both the artificial defect generation and final training stages are reported in the appendix. An anonymous repository containing the source code is linked in the main paper to further support reproducibility. The supplementary material also includes additional experimental results and analyses supporting the reported findings.

REFERENCES

Amr Adel, Bhagwan Das, and Tony Jan. Federated predictive maintenance with nasa bp930 satellite battery in industry 5.0: A. adel et al. *Journal of Intelligent Manufacturing*, pp. 1–19, 2026.

P Anand Kumar and S Sountharrajan. Insurance claims estimation and fraud detection with optimized deep learning techniques. *Scientific Reports*, 15(1):27296, 2025.

Asma. Pill Detection Dataset. `https://universe.roboflow.com/asma-pbbz2/pill-detection-gggra`. Available at Roboflow Universe, accessed March 17, 2026.

Lin Chen, Xiaofeng Ding, Zhifeng Bao, Pan Zhou, and Hai Jin. Differentially private federated learning on non-iid data: Convergence analysis and adaptive optimization. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4567–4581, 2024.

Yajun Chen, Yuanyuan Ding, Fan Zhao, Erhu Zhang, Zhangnan Wu, and Linhao Shao. Surface defect detection methods for industrial products: A review. *Applied Sciences*, 11(16):7657, 2021.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911, June 2024.

Xiao Du, Bing Li, Zhuo Zhao, Bingke Jiang, Yupeng Shi, Liuyi Jin, and Xiao Jin. Anomaly-prior guided inpainting for industrial visual anomaly detection. *Optics & Laser Technology*, 170: 110296, 2024.

Yulian Gao, Gehao Lu, Jimei Gao, and Jinggang Li. A high-performance federated learning aggregation algorithm based on learning rate adjustment and client sampling. *Mathematics*, 11 (20), 2023. ISSN 2227-7390. doi: 10.3390/math11204344. URL https://www.mdpi.com/2227-7390/11/20/4344.

Muhammad Uzair Gill and Parvathy Rajendran. Hyperparameter optimization of yolo using differential evolution, multi-fidelity optimization, and bayesian optimization. *Scientific Reports*, 2026.

Alessandro Giuseppi, Lucrezia Della Torre, Danilo Menegatti, and Antonio Pietrabissa. Adafed: Performance-based adaptive federated learning. In *Proceedings of the 5th International Conference on Advances in Artificial Intelligence*, pp. 38–43, 2021.

Enwei Guo, Xiumin Wang, and Weiwei Wu. Adaptive aggregation weight assignment for federated learning: A deep reinforcement learning approach. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 1610–1612, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Wenyou Guo, Ting Qu, Yongheng Zhang, Hainan Huang, and George Q Huang. A hierarchical federated learning approach based on cloud–fog–edge computing architecture for distributed smart manufacturing systems. *Robotics and Computer-Integrated Manufacturing*, 101:103312, 2026.

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*, pp. 16000–16009, 2022.

Yu He, Kechen Song, Qinggang Meng, and Yunhui Yan. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1493–1504, 2020. doi: 10.1109/TIM.2019.2915404.

Jose Luis Hernandez-Ramos, Georgios Karopoulos, Efstratios Chatzoglou, Vasileios Kouliaridis, Enrique Marmol, Aurora Gonzalez-Vidal, and Georgios Kambourakis. Intrusion detection based on federated learning: a systematic review. *ACM Computing Surveys*, 57(12):1–65, 2025.

Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised anomaly detection and localization. *IEEE Transactions on Multimedia*, 25:4426–4438, 2023. doi: 10.1109/TMM.2022.3175611.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Koffka Khan and Wayne Goodridge. Swarm intelligence-driven client selection for federated learning in cybersecurity applications. *arXiv preprint arXiv:2411.18877*, 2024.

Naeem Khan, Shibli Nisar, Muhammad Asghar Khan, Yasar Abbas Ur Rehman, Fazal Noor, and Gordana Barb. Optimizing federated learning with aggregation strategies: A comprehensive survey. *IEEE Open Journal of the Computer Society*, 2025.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Yong-Suk Lee, Maheshkumar Prakash Patil, Jeong Gyu Kim, Yong Bae Seo, Dong-Hyun Ahn, and Gun-Do Kim. Hyperparameter optimization of apple leaf dataset for the disease recognition based on the yolov8. *Journal of Agriculture and Food Research*, 21:101840, 2025.

Xingyu Li, Wenzhe Zhang, Linfeng Liu, and Jia Xu. Exploring the robustness: Hierarchical federated learning framework for object detection of uav cluster. *IEEE Transactions on Mobile Computing*, 2025.

Ravikant Mordia and Arvind Kumar Verma. Visual techniques for defects detection in steel products: A comparative study. *Engineering Failure Analysis*, 134:106047, 2022.

Hitoshi Nakanishi, Masahiro Suzuki, and Yutaka Matsuo. Fixing the train-test objective discrepancy: Iterative image inpainting for unsupervised anomaly detection. *Journal of Information Processing*, 30:495–504, 2022.

Priyanka Nandal, Navdeep Bohra, Prerna Mann, and Nripendra Narayan Das. Yolov11 with transformer attention for real-time monitoring of ships: A federated learning approach for maritime surveillance. *Results in Engineering*, pp. 106297, 2025.

Iraj Naruei and Farshid Keynia. Wild horse optimizer: A new meta-heuristic algorithm for solving engineering optimization problems. *Engineering with computers*, 38(Suppl 4):3025–3056, 2022.

Dianwen Ng, Xiang Lan, Melissa Min-Szu Yao, Wing P Chan, and Mengling Feng. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quantitative Imaging in Medicine and Surgery*, 11(2):852, 2021.

Qi Qiao, Huiying Hu, Azlin Ahmad, and Ke Wang. A review of metal surface defect detection technologies in industrial applications. *IEEE Access*, 2025.

Zhonghe Ren, Fengzhou Fang, Ning Yan, and You Wu. State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 9(2):661–691, 2022.

Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7301–7311, 2023. doi: 10.1109/ICCV51070.2023.00674.

Hongbing Shang, Chuang Sun, Jinxin Liu, Xuefeng Chen, and Ruqiang Yan. Defect-aware transformer network for intelligent visual surface defect detection. *Advanced Engineering Informatics*, 55:101882, 2023.

Donald Shenaj, Giulia Rizzoli, and Pietro Zanuttigh. Federated learning in computer vision. *Ieee Access*, 11:94863–94884, 2023.

T. Srihari, A. Murugesan, Neeraj Kumar, T Eswaran, T. V. Hyma Lakshmi, and K.T. Shivaram. Defect detection in electronics manufacturing via deep learning-based visual inspection. In *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, pp. 1–4, 2025. doi: 10.1109/ICCTDC64446.2025.11159033.

T Aditya Sai Srinivas, Ahmed Kareem Shakir, G Saritha, et al. Adaptive federated learning through particle swarm optimization for dynamic client weighting. In *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1–6. IEEE, 2025.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Multi-task learning for medical image inpainting based on organ boundary awareness. *Applied Sciences*, 11(9):4247, 2021.

George-Alexandru Tudor, Radu-Ioan Ciobanu, and Ciprian Dobre. Distributed collective learning on mobile devices: A real-hardware fedavg evaluation. In *2026 22nd International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pp. 828–833. IEEE, 2026.

Ajantha Vijayakumar, Joseph Abraham Sundar Koilraj, and Muthaiah Rajappa. Pharmanet deep: real-time pharmaceutical defect detection using defect-guided feature fusion and uncertainty-driven inspection. *International Journal of Computational Intelligence Systems*, 18(1):250, 2025.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024a.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pp. 1–21. Springer, 2024b.

Haixia Xu, Yuting Xu, and Kaiyu Hu. A vision-based inspection system for pharmaceutical production line. *The Journal of Supercomputing*, 81(4):625, 2025.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

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

Xiaobo Zhang, Donghai Zhai, Tianrui Li, Yuxin Zhou, and Yang Lin. Image inpainting based on deep learning: A review. *Information Fusion*, 90:74–94, 2023.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Hantao Zhou, Rui Yang, Runze Hu, Chang Shu, Xiaochu Tang, and Xiu Li. Etdnet: Efficient transformer-based detection network for surface defect detection. *IEEE transactions on instrumentation and measurement*, 72:1–14, 2023a.

Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pp. 1–6. IEEE, 2023b.

A Additional Experimental Details.

This appendix provides additional details to ensure reproducibility of the proposed framework, including hyperparameter configurations, system setup, and extended training analysis across datasets and clients.

A.1 Hyperparameter Configuration

This section provides the complete hyperparameter settings used in our experiments, including the search spaces considered by FL-HSO and the client-specific configurations used in the artificial-defect and final real-data training phases.

Hyperparameter search space. Hyperparameter optimization methods in deep learning commonly consider either discrete or continuous search spaces. Following representative formulations from prior work, we evaluate both alternatives Anand Kumar & Sountharrajan (2025); Lee et al. (2025); Gill & Rajendran (2026). For the discrete formulation, we use `epochs` $\in \{1, 5, 10\}$, `batch` $\in \{2, 3, 4\}$, `lr0` $\in \{0.001, 0.005, 0.01\}$, `momentum` $\in \{0.900, 0.937, 0.950\}$, `weight_decay` $\in \{0.0001, 0.0005, 0.001\}$, and `warmup_epochs` $\in \{2, 3, 4\}$.

For the continuous formulation, `lr0` $\in [10^{-4}, 10^{-2}]$, `momentum` $\in [0.85, 0.95]$, and `weight_decay` $\in [10^{-5}, 10^{-3}]$, while the integer-valued parameters are sampled from `epochs` $\in \{1, 5, 10\}$, `batch` $\in \{2, 3, 4, 5\}$, and `warmup_epochs` $\in \{2, 3, 4\}$.

As shown in Figure 2b, the discrete formulation provides better overall performance in our ablation while also reducing search cost through a finite set of candidate configurations. It additionally improves reproducibility because the same candidate configuration can be recovered exactly across independent runs. We therefore adopt the discrete search space for the remaining FL-HSO experiments.

Client-specific configurations. Tables 4–7 report the exact client-specific configurations used for the Pill and NEU-DET datasets. Because the experiments cover different datasets, image resolutions, training stages, federation sizes, and computational budgets, the final training configurations are not identical across all experiments. We report these differences explicitly to make the experimental protocol fully reproducible and to avoid conflating the hyperparameter-search setting with the subsequent training budget.

The artificial-defect phase uses the client-specific configurations obtained during hyperparameter optimization. In the subsequent real-data training phase, the training configuration is adapted to the corresponding dataset and training budget; the exact values used for each client are reported below.

Parameter	Client 1	Client 2	Client 3
Epochs	10	10	10
Batch Size	4	3	3
Learning Rate	0.01	0.01	0.01
Momentum	0.9	0.937	0.937
Weight Decay	0.0005	0.0005	0.0005
Warmup Epochs	2	4	4
Train Images	941	941	941
Train Instances	2959	2849	2872
Validation Images	40	40	40
Validation Instances	123	136	44

Table 4: Artificial Defect Detection Phase: Hyperparameters, Dataset Configuration, and Image/Instance Info for Pill Tablet Dataset

Parameter	Client 1	Client 2	Client 3
Epochs	100	100	100
Batch Size	4	3	3
Learning Rate	0.01	0.01	0.01
Momentum	0.9	0.937	0.937
Weight Decay	0.0005	0.0005	0.0005
Warmup Epochs	2	4	4
Train Images	941	941	941
Train Instances	2959	2849	2872
Validation Images	40	40	40
Validation Instances	123	136	44

Table 5: In the Wild Defect Detection Phase: Hyperparameters, Dataset Configuration, and Image/Instance Info for Pill Tablet Dataset

Parameter	Client 1	Client 2	Client 3
Epochs	10	10	10
Batch Size	3	4	4
Learning Rate	0.01	0.01	0.01
Momentum	0.937	0.9	0.9
Weight Decay	0.0005	0.0005	0.0005
Warmup Epochs	4	2	2
Train Images	420	420	420
Train Instances	952	989	981
Validation Images	120	120	120
Validation Instances	275	288	282

Table 6: Artificial Defect Detection Phase: Hyperparameters, Dataset Configuration, and Image/Instance Info for NEU-DET Dataset

A.2 System Setup

All experiments were conducted under the software and hardware configuration reported in Table 8.

Parameter	Client 1	Client 2	Client 3
Epochs	300	300	300
Batch Size	9	12	12
Learning Rate	0.03	0.03	0.03
Momentum	0.937	0.9	0.9
Weight Decay	0.0005	0.0005	0.0005
Warmup Epochs	4	2	2
Train Images	420	420	420
Train Instances	952	989	981
Validation Images	120	120	120
Validation Instances	275	288	282

Table 7: In the Wild Defect Detection Phase: Hyperparameters, Dataset Configuration, and Image/Instance Info for NEU-DET Dataset

Software/Hardware	Details
Ultralytics Version	8.3.163
Python Version	3.10.18
Torch Version	2.6.0+cu124
GPU	NVIDIA RTX A4500
GPU Memory	20149 MiB

Table 8: System Configuration

B Training Dynamics

In this section, we provide a detailed view of the training dynamics for each client and dataset to illustrate convergence behavior and stability across training phases. These plots allow inspection of optimization consistency and confirm the absence of significant training instability across clients.

B.1 Communication-Round Analysis

Table 9 reports the aggregate performance across communication rounds, with mean and standard deviation computed over all client counts and participating clients. The results complement the analysis in Section 4.4 by providing the exact values underlying the observed saturation behavior.

B.2 Pill Dataset

Figures 7–9 show the training progression for each client on the Pill dataset.

B.3 NEU-DET Dataset

Figures 10–12 show the training progression for each client on the NEU-DET dataset.

C Dataset Analysis

In this section, we analyze the distribution of training labels to provide additional insight into class balance and data characteristics for each dataset. Figures 13 and 14 illustrate the label distribution during training for the Pill and NEU-DET datasets, respectively. These statistics provide additional insight into class balance and data characteristics across datasets, which are relevant for interpreting the performance differences observed in the main results.

Round	Precision	Recall	mAP_{50}	mAP_{95}
1	0.694 ± 0.052	0.523 ± 0.026	0.609 ± 0.037	0.301 ± 0.020
2	0.685 ± 0.049	0.527 ± 0.028	0.612 ± 0.036	0.301 ± 0.019
3	0.690 ± 0.051	0.528 ± 0.027	0.612 ± 0.038	0.301 ± 0.021
4	0.686 ± 0.050	0.529 ± 0.029	0.613 ± 0.037	0.302 ± 0.020
5	0.687 ± 0.053	0.530 ± 0.028	0.613 ± 0.039	0.302 ± 0.021

Table 9: Performance across federated rounds. Each cell reports mean $\pm$ std computed over all clients and all client counts .

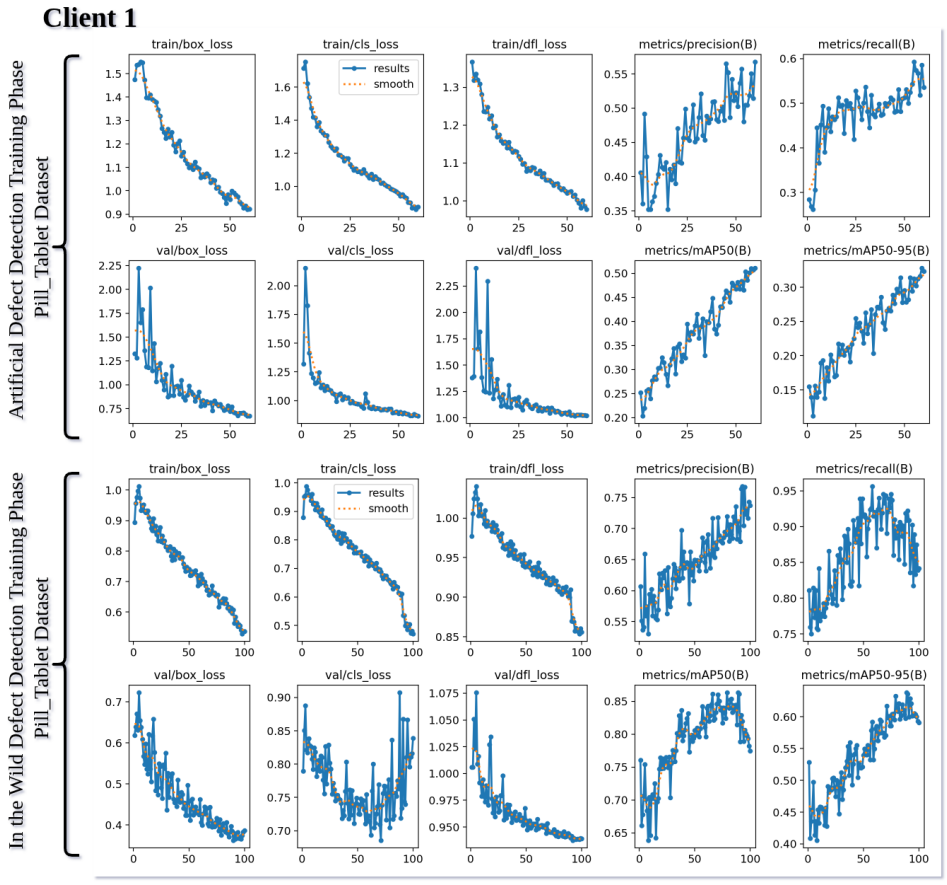

Figure 7: Progression of Client 1 Training in the Pill Tablet Dataset

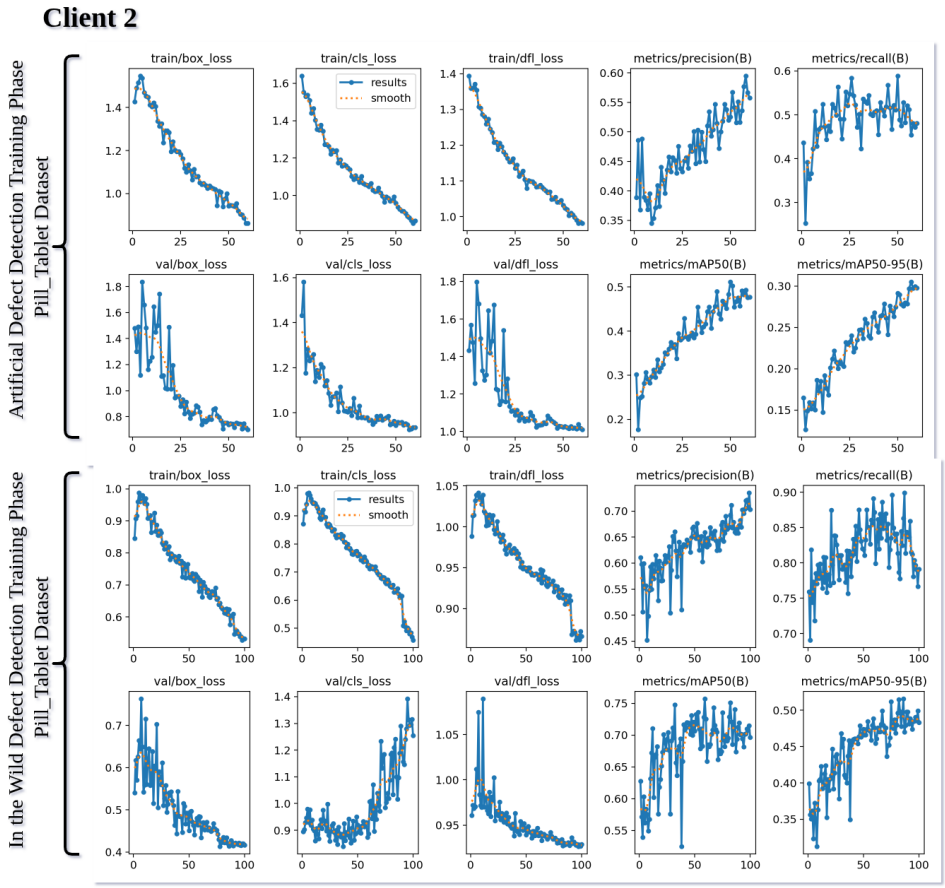

Figure 8: Progression of Client 2 Training in the Pill Tablet Dataset

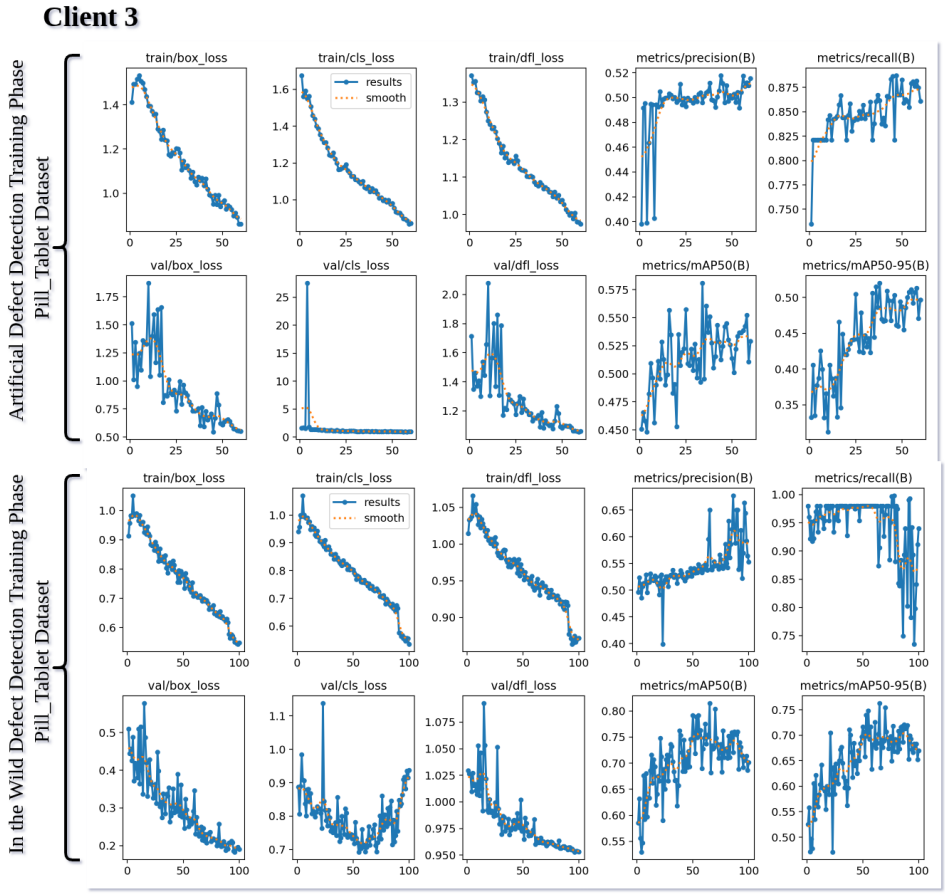

Figure 9: Progression of Client 3 Training in the Pill Tablet Dataset

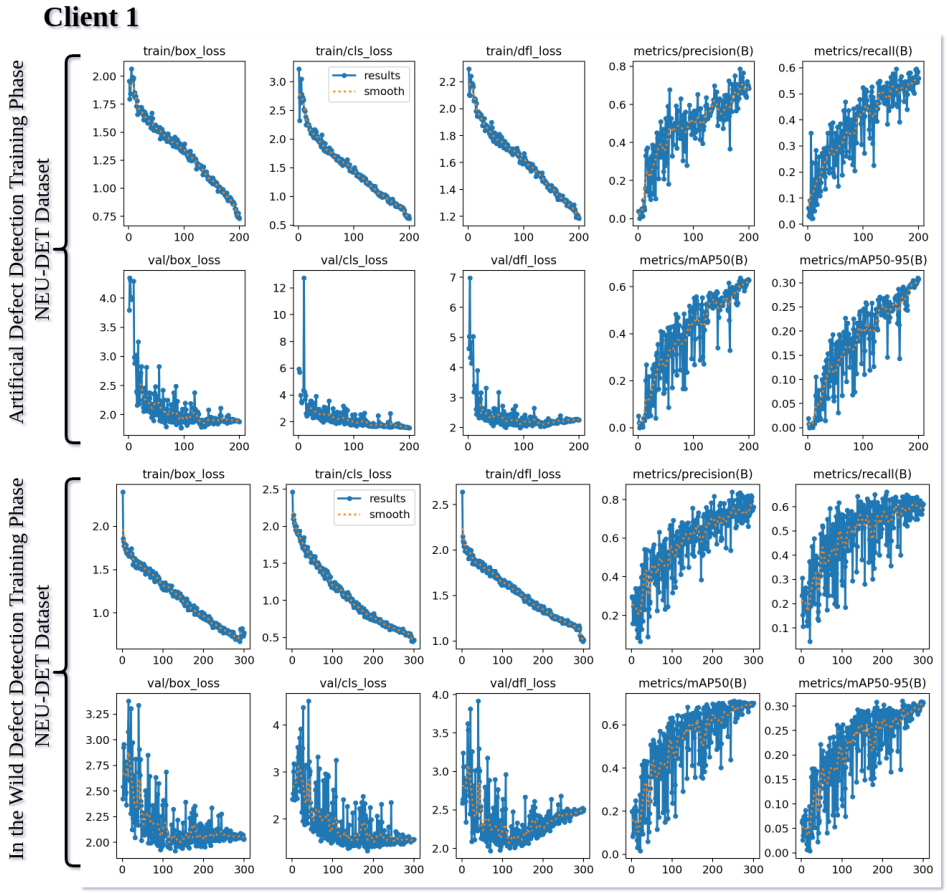

Figure 10: Progression of Client 1 Training in the NEU-DET Dataset

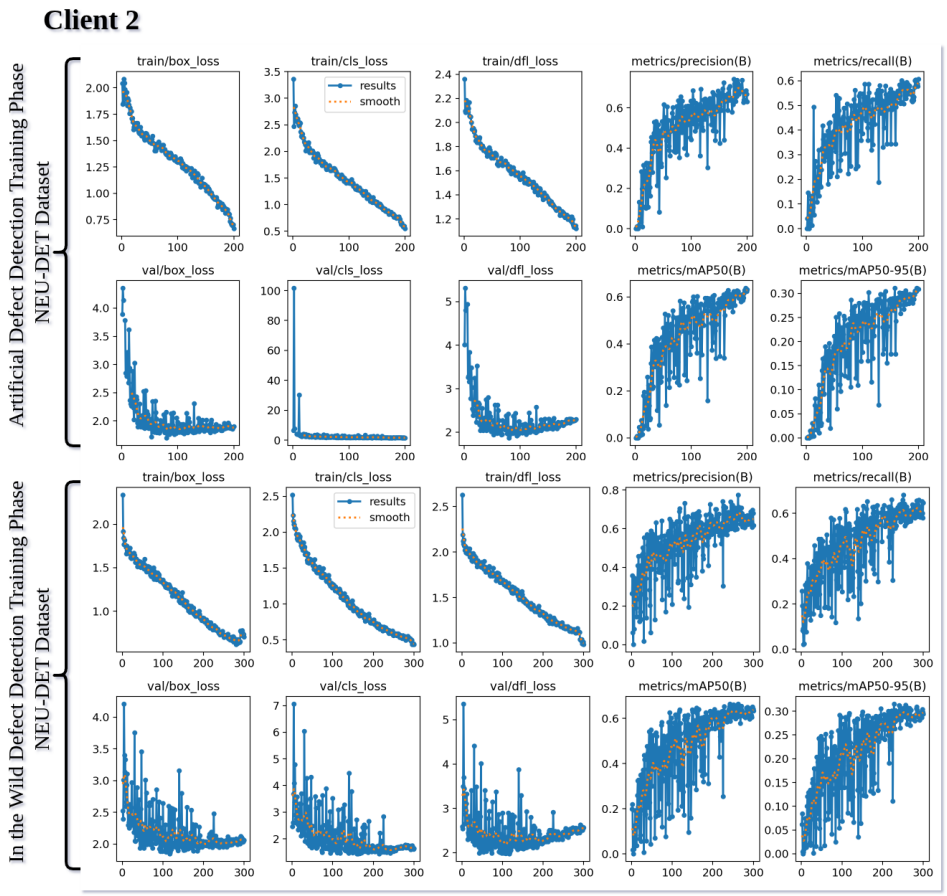

Figure 11: Progression of Client 2 Training in the NEU-DET Dataset

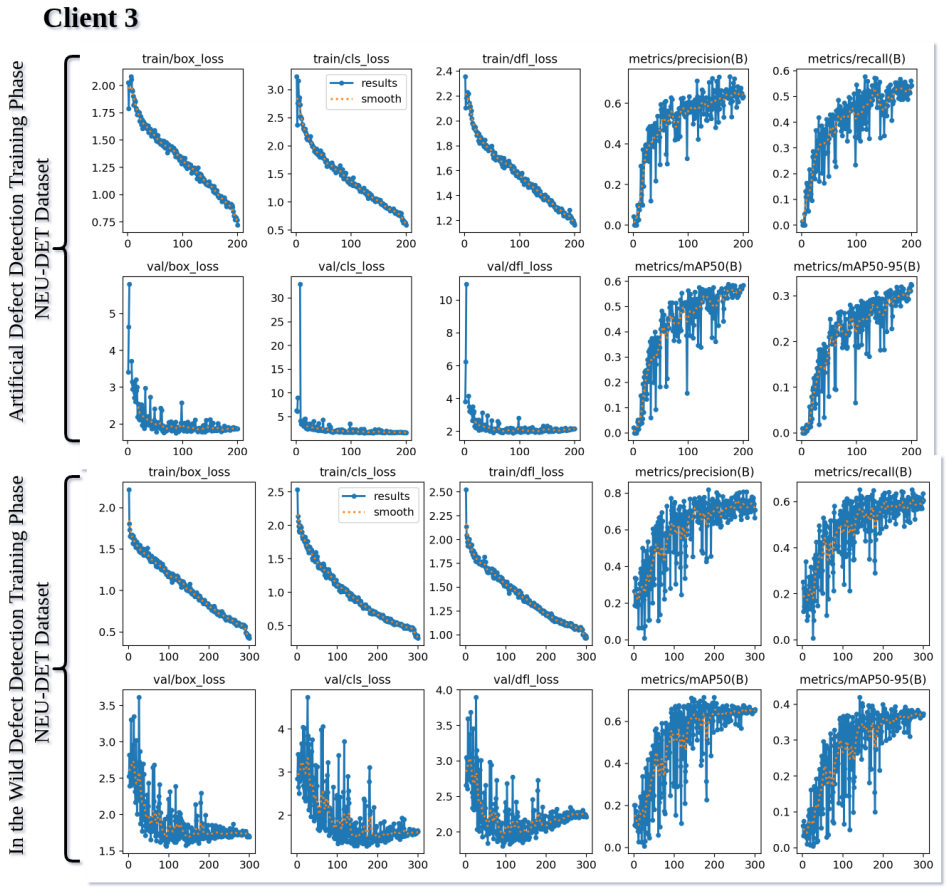

Figure 12: Progression of Client 3 Training in the NEU-DET Dataset

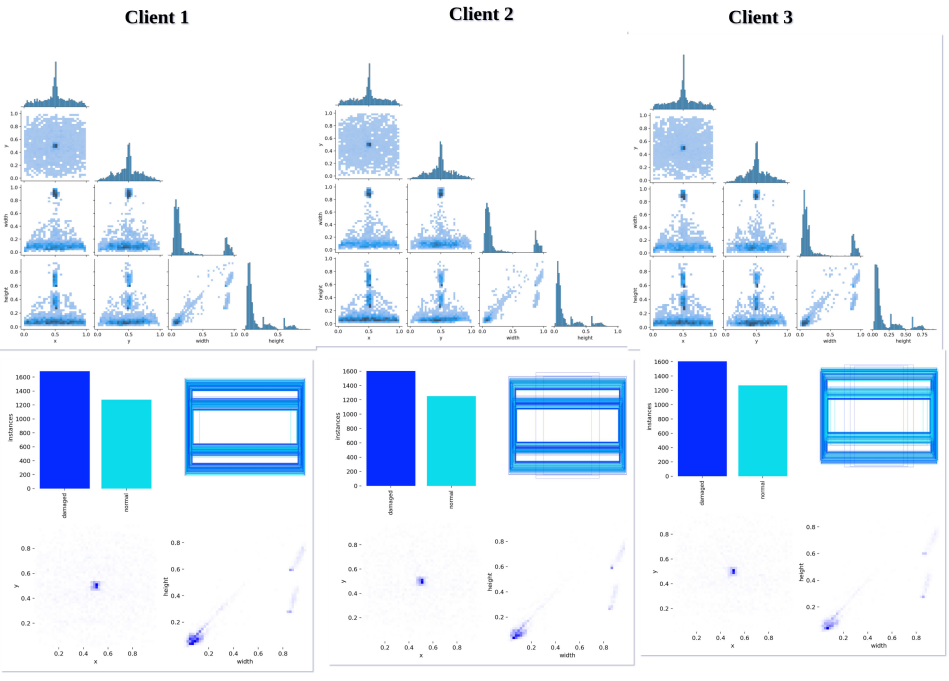

Figure 13: Training Label Distribution in the Pill Tablet Dataset

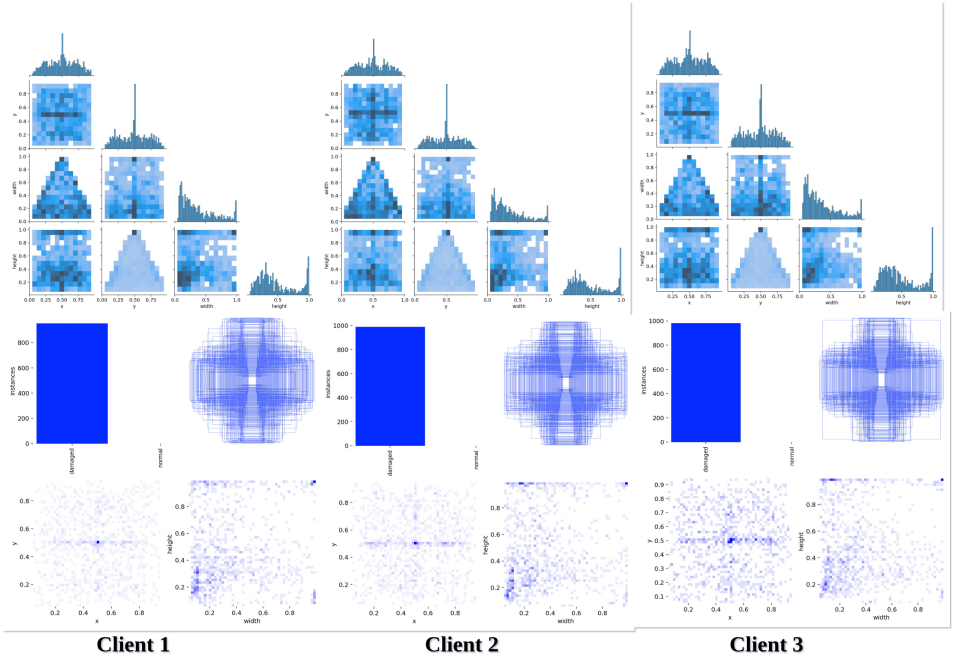

Figure 14: Training Label Distribution in the NEU-DET Dataset