electronics

MDPI

Article

Decoupled Candidate Detection and Multi-Evidence Verification for Tiny Surface Defect Inspection

Zhicheng Cheng, Qi Long, Mei Wang * and Jigang Huang *

School of Mechanical Engineering, Sichuan University, Chengdu 610065, China; 2024223020019@stu.scu.edu.cn (Z.C.); 2021141410032@stu.scu.edu.cn (Q.L.)
* Correspondence: wangmei@scu.edu.cn (M.W.); jigang.huang@scu.edu.cn (J.H.)

Abstract

Tiny surface defects are difficult to inspect because of their small scale, weak contrast, diverse morphology, and similarity to normal textures. To address the conflict between early candidate retention and false-positive suppression, this study proposes a coarse-to-fine inspection framework that separates candidate detection from final verification. You Only Look Once (YOLO) 26s first generates defect candidates, which are locally reconstructed and analyzed by a region-of-interest (ROI)-guided Multi-Task SAM (ROI-MT-SAM) for joint segmentation, semantic prediction, and mask-quality estimation. Multi-Evidence Verification (MEV) then integrates complementary confidence and spatial cues to refine the final decision, while residual-transfer augmentation improves defect diversity under limited-data conditions. Experiments show that ROI-MT-SAM improves mean Intersection over Union (mIoU) from 0.3918 to 0.8869 compared with whole-image SAM, with particularly substantial gains for small defects. MEV further reduces false positives from 245 to 46 and achieves a precision of 0.9565 and an F1-score of 0.9241. These results demonstrate the potential of combining candidate-centered representation with multi-evidence verification for tiny-defect inspection of automotive interior surfaces under the evaluated conditions.

Keywords: surface defect inspection; small defect detection; segment anything model; multi-evidence verification; residual-transfer augmentation

1. Introduction

Surface appearance quality is a critical requirement in automotive interior manufacturing, as visible defects directly affect perceived craftsmanship, product consistency, and final acceptance. Automatic inspection of such defects remains challenging because they are often extremely small, weakly contrasted, morphologically diverse, and easily confused with normal material textures [1]. Point-like defects may occupy only a minute fraction of a high-resolution image, whereas scratches and fiber-like defects often exhibit thin or elongated structures [2]. Consequently, the discriminative information of a defect can be severely weakened by image resizing and hierarchical feature downsampling [3]. These difficulties are compounded by the scarcity and class imbalance of annotated defect samples [4]. Reliable inspection therefore requires not only accurate localization but also effective preservation of weak defect evidence throughout the inspection pipeline [5].

Deep-learning-based methods have advanced industrial surface inspection through multi-scale feature extraction, defect localization, and anomaly representation learning [6]. Global-context-guided feature aggregation has further improved the representation of weak defects and suppression of background interference [7]. Recent detection–segmentation

Check for updates

Academic Editor: Honarvar Shakibaei Asli

Received: 9 October 2026
Revised: 3 November 2026
Accepted: 4 November 2026
Published: 5 November 2026

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

pipelines have integrated YOLO-based localization with SAM-based mask generation [8], while dual-stage SAM frameworks have explored boundary refinement and false-positive suppression [9]. Other studies have improved SAM adaptation to industrial defects through hierarchical feature modulation [10] and parameter-efficient fine-tuning [11]. Despite these advances, reliable inspection of tiny, low-contrast defects remains challenging because weak defect responses can be confused with background textures. Aggressive confidence filtering risks rejecting genuine defects, whereas permissive detection retains false-positive candidates. Moreover, whole-image resizing may suppress fine-scale defect information required for subsequent segmentation and verification.

To address these limitations, this study decouples recall-oriented candidate generation from evidence-based final defect acceptance. Although recent industrial segmentation methods have explored complementary semantic and texture representations, their primary objective remains accurate defect delineation rather than explicit candidate-level acceptance [12]. The proposed framework explicitly exploits the asymmetric consequences of candidate-stage errors: discarded true defects cannot be recovered downstream, whereas retained false positives can be reassessed using additional evidence. Recent two-stage verification research also supports separating initial defect predictions from subsequent evidence-based reassessment [13]. Accordingly, YOLO26s generates defect candidates, ROI-guided Multi-Task SAM (ROI-MT-SAM) provides fine-grained geometric and semantic information through pre-encoding ROI reconstruction, and Multi-Evidence Verification (MEV) integrates complementary cues for final candidate acceptance. The resulting framework is designed to improve candidate discrimination while explicitly accounting for the trade-off between false-positive suppression and defect recall.

Secondary verification requires fine-grained information that is unavailable from detector confidence alone [14]. Promptable segmentation models provide a promising mechanism for extracting spatial structure from detector-generated candidates [15], but direct whole-image SAM inference is poorly matched to tiny industrial defects. When a high-resolution image is resized to the fixed input scale of the SAM encoder, a small defect may occupy only a negligible portion of the encoded feature space [16]. Weak boundaries and subtle texture deviations can therefore be lost before mask decoding [17], and such information cannot be recovered by an accurate box prompt after encoding [18]. The key challenge is therefore not merely how SAM is prompted but at what spatial scale the defect is represented before encoding. Candidate-centered reconstruction directly addresses this scale mismatch by enlarging the relative representation of the defect while retaining sufficient surrounding context [19,20].

To this end, we develop an ROI-guided Multi-Task SAM (ROI-MT-SAM). Each retained candidate is expanded into a context-preserving ROI before SAM encoding, increasing the effective spatial representation of tiny defects while avoiding unnecessary whole-image computation. Using a shared SAM representation, ROI-MT-SAM jointly predicts a fine-grained segmentation mask, semantic category confidence, and mask-quality estimate. These outputs characterize complementary properties of the same candidate—its spatial structure, semantic consistency, and segmentation reliability—and provide the evidence required for subsequent verification. Thus, SAM is extended from pure mask generation to candidate-level multi-task interpretation.

Limited defective data introduces a separate but important challenge [21]. Conventional geometric and photometric augmentation increases apparent sample diversity but largely preserves the original pairing between a defect and its local background [22]. Direct RGB copy-and-paste provides greater spatial flexibility, yet may transfer source-background texture together with the defect, producing artificial boundaries or local inconsistencies [23]. This is particularly problematic for automotive interior defects, which often appear as local

perturbations embedded in material-specific textures rather than as independent foreground objects [24]. We therefore introduce a residual-transfer augmentation strategy that transfers defect-induced appearance variations, rather than complete RGB patches, to texture-compatible target regions using morphology-aware transformations and local intensity adaptation.

Based on these considerations, a recall-preserving coarse-to-fine framework is constructed for automotive interior surface-defect inspection. YOLO26s first operates as a class-agnostic candidate generator to retain potentially defective regions. The retained candidates are then reconstructed locally and analyzed by ROI-MT-SAM to recover fine-scale geometric, semantic, and mask-reliability information. Finally, MEV integrates detector confidence, semantic confidence, mask quality, and spatial consistency to reject unsupported candidates. In this way, false-positive suppression is deliberately postponed until richer evidence becomes available, while irreversible early misses are minimized.

The main contributions of this work are summarized as follows:

Recall-oriented candidate generation and multi-evidence verification. A decoupled inspection framework is developed to separate candidate retention from final defect acceptance. The proposed MEV mechanism integrates detection confidence, semantic confidence, mask-quality estimates, and spatial consistency to suppress false positives while explicitly accounting for the associated reduction in recall.

An ROI-guided Multi-Task SAM is developed for fine-grained analysis of extremely small surface defects. Candidate-centered reconstruction is performed before SAM encoding to mitigate scale-induced information loss, while the shared representation jointly supports segmentation, semantic prediction, and mask-quality estimation.

A residual-transfer augmentation strategy is introduced for limited and imbalanced defect data. Defect-induced appearance residuals are transferred to texture-compatible regions using morphology-aware transformations and local intensity adaptation, increasing defect diversity while reducing the background inconsistency associated with direct RGB patch transfer.

2. Dataset Construction and Residual-Transfer Augmentation

2.1. Dataset Construction

A dedicated automotive interior surface-defect dataset was constructed from real production surfaces. It contains 152 independent source images covering nine defect categories: bump, dent, scratch, color variation, fiber, collision mark, pin hole, fluff, and black spot. As illustrated in Figure 1, the defects exhibit diverse morphologies, including point-like, elongated, and irregular structures, posing challenges for localization and segmentation.

All defect instances were annotated with polygonal contours to preserve pixel-level geometry. The annotations were subsequently converted into bounding boxes for candidate detection and binary masks for fine-grained segmentation, ensuring consistent ground truth across the two tasks.

To prevent augmentation-induced data leakage, the dataset was partitioned at the source-image level before augmentation. All augmented samples derived from the same source image were assigned to the same subset, avoiding cross-subset contamination and potentially optimistic performance estimates.

Specifically, the 152 source images were divided into 50 training, 50 validation, and 52 test images. Augmentation was performed only within the predefined training and validation subsets, producing 1600 training images and 400 validation images, containing 4576 and 1131 annotated defect instances, respectively. The 52 test images, containing 98 original defect instances, were kept completely independent of augmentation, hyper-

parameter selection, and model optimization. This partition was fixed throughout all subsequent experiments.

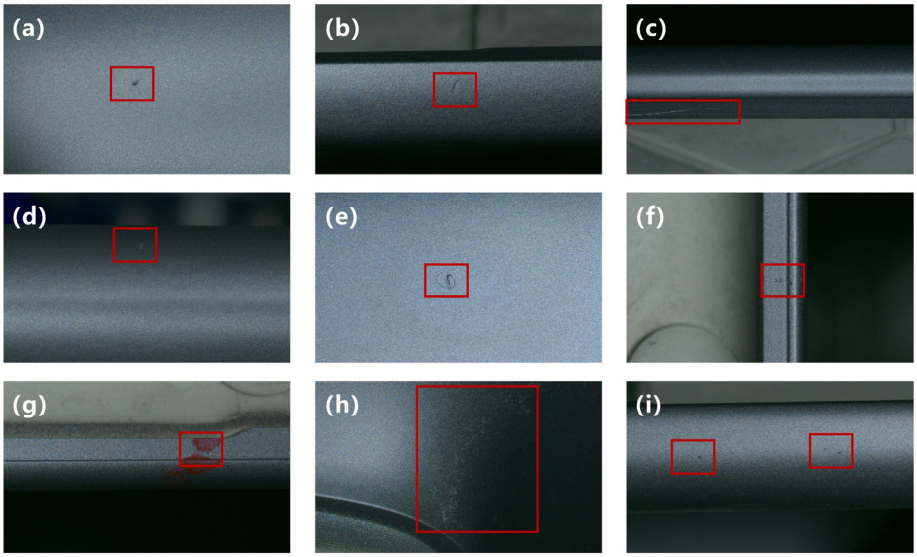


Figure 1. Representative examples of the nine automotive interior surface-defect categories: (**a**) bump, (**b**) dent, (**c**) scratch, (**d**) color variation, (**e**) fiber, (**f**) collision mark, (**g**) pin hole, (**h**) fluff, and (**i**) black spot.

Table 1 summarizes the class-wise distributions. Because one image may contain multiple instances or defect categories, the class-wise image counts are not mutually exclusive. The instance distribution is imbalanced, with fiber, collision mark, and scratch occurring more frequently than pin hole and dent, motivating the targeted augmentation described in Section 2.2.

Table 1. Dataset composition and class-wise distribution of defect instances.

Defect Category	Train Images	Train Instances	Val Images	Val Instances	Test Images	Test Instances
Bump	316	408	80	108	6	8
Dent	195	231	49	55	2	3
Scratch	389	641	98	144	7	15
Color variation	420	519	106	135	9	11
Fiber	658	1058	167	254	16	26
Collision mark	423	703	107	175	8	17
Pin hole	152	152	38	38	1	1
Fluff	361	408	91	103	7	8
Black spot	368	456	92	119	7	9

2.2. Data Augmentation

The limited number and imbalanced distribution of real defects constrain representation learning [25]. Conventional geometric and photometric augmentation increases sample variability but largely preserves the original defect–background pairing [21]. Direct red-green-blue (RGB) copy-and-paste provides greater spatial diversity [26], yet is poorly suited to automotive interior surfaces because many defects appear as local appearance perturbations embedded in material-specific textures, rather than as independent foreground objects [27]. Transferring a complete RGB patch may therefore introduce source-background texture, artificial boundaries, or illumination discontinuities.

To preserve target-material consistency, we introduce a residual-transfer augmentation strategy. Instead of transferring the complete source patch, the method isolates the defect-induced appearance variation and injects only this residual into a texture-compatible target region.

Let I_s denote a source image, M_s the binary defect mask, and $\hat{I}_s^b$ the estimated defect-free background within the annotated region. The defect residual is defined as

$$R_s = M_s^{\text{sup}} \odot \left(I_s - \hat{I}_s^b\right) \tag{1}$$

where $\odot$ denotes element-wise multiplication. Thus, R_s captures the local intensity and texture deviation associated with the defect while suppressing the surrounding source-background appearance. M_s^{sup} denotes the dilated support mask derived from the annotated defect mask M_s.

A texture-compatible target region is selected based on local background similarity. The residual and its associated masks are then jointly transformed through morphology-aware scaling, rotation, flipping, elastic deformation, and mask perturbation. The augmented image is generated as

$$I_{aug} = clip(I_t + \lambda_r \tau(R_s)) \tag{2}$$

where I_{aug} denotes the synthesized augmented image, I_t is the target image, $\tau(\cdot)$ represents the geometric transformation and spatial placement applied to the defect residual, λ_r controls the residual intensity, and clip$(\cdot)$ constrains the synthesized pixel values to the valid image range. All spatial transformations are applied consistently to the residual and its associated masks to maintain alignment.

The defect-free background is estimated using Telea inpainting with a 5-pixel radius following 7 × 7 mask dilation. Residuals are extracted within a 5 × 5 dilated support region and clipped to $[-70, 70]$. Texture-compatible target locations are selected using a 12-dimensional color–texture descriptor, with up to 70 candidate positions evaluated and one sampled from the six highest-ranked matches. Category-dependent transformations are applied to the residuals and associated masks. Residual intensity and local contrast are adjusted using factors in $[0.75, 1.20]$ and $[0.65, 1.35]$, respectively, followed by spatial feathering. Additional implementation details, including texture-matching parameters and morphology-dependent transformations, are provided in Supplementary Table S1.

Figure 2a illustrates the augmentation process, while Figure 2b,c shows the class distributions of the independent test set and the augmented training–validation set. The latter contains 5707 annotated instances across 2000 augmented images, whereas the independent test set contains 98 instances from 52 unaugmented source images.

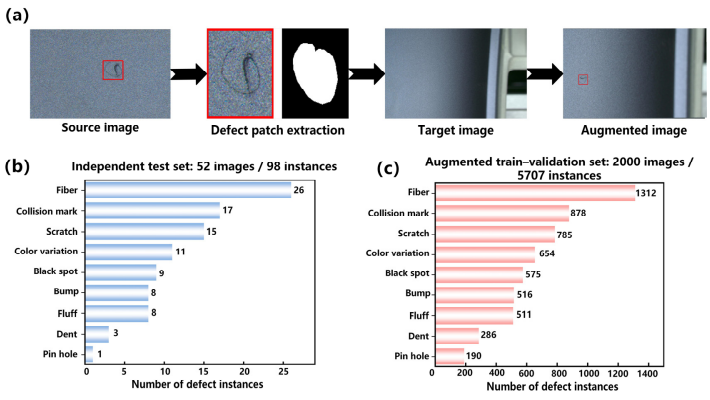


Figure 2. Residual-transfer augmentation and dataset distributions. (**a**) Schematic of the proposed residual-transfer augmentation procedure. (**b**) Class distribution of the independent test set. (**c**) Class distribution of the augmented training and validation sets.

The augmented images are derivatives of the original source images and should not be treated as statistically independent observations.

3. Method

3.1. Overall Framework

As illustrated in Figure 3, the proposed framework comprises three stages: recall-oriented candidate generation, ROI-guided multi-task analysis, and multi-evidence verification.

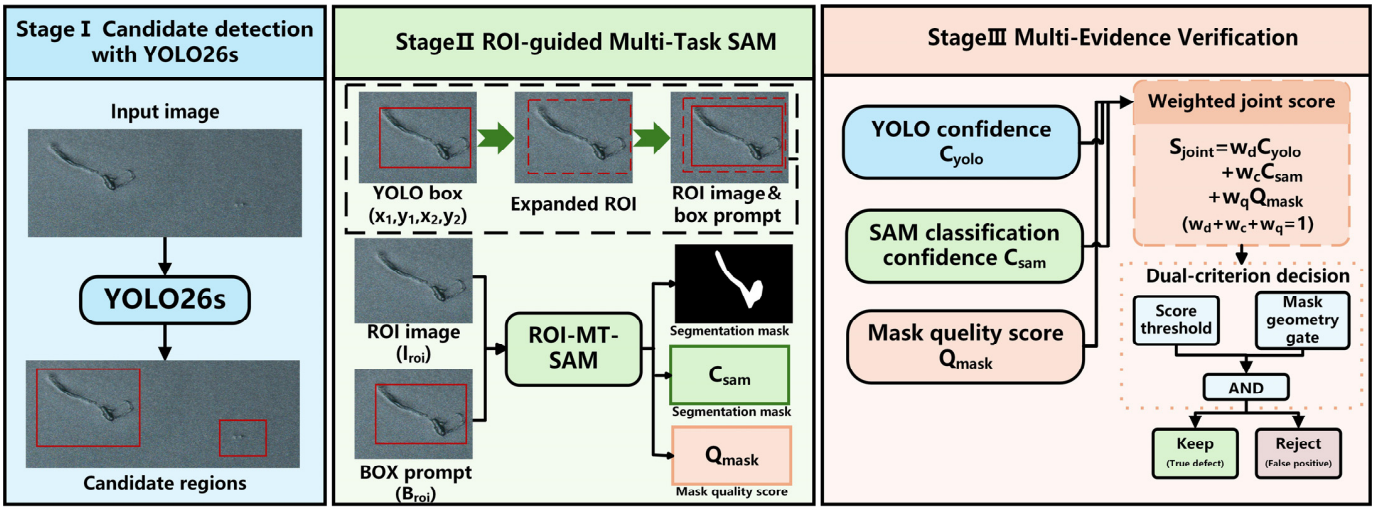


Figure 3. Overall architecture of the proposed coarse-to-fine framework with recall-oriented candidate generation and multi-evidence verification for automotive interior surface-defect inspection.

In Stage I, YOLO26s operates as a class-agnostic candidate generator rather than the final defect classifier. The nine defect categories are merged into a single defect class during training, allowing the detector to focus on preserving potentially defective regions and reducing irreversible early misses.

In Stage II, each retained candidate is expanded into a context-preserving ROI and processed by ROI-MT-SAM. Candidate-centered reconstruction increases the relative spatial extent of tiny defects at the encoder input. The model jointly predicts a fine-grained segmentation mask, semantic confidence, and mask-quality estimate.

In Stage III, MEV integrates detector confidence with the semantic and mask-quality cues produced by ROI-MT-SAM. A spatial-consistency gate further rejects candidates whose predicted masks violate predefined geometric constraints.

3.2. High-Recall Candidate Detection

The first stage aims to retain a sufficiently complete set of potentially defective regions for subsequent analysis. Because fine-grained category recognition is performed by ROI-MT-SAM, the nine original categories are merged into a single defect class during YOLO26s training. The detector therefore concentrates on distinguishing abnormal regions from normal backgrounds rather than performing nine-class recognition [28].

After non-maximum suppression (NMS), candidates satisfying the retention threshold τ_d are preserved:

$$C(\tau_d) = \left\{ \left(B_i, C_i^{\text{det}} \right) \middle| C_i^{\text{det}} \geq \tau_d,\ i = 1, \ldots, N \right\} \tag{3}$$

where $C(\tau_d)$ denotes the set of retained candidates under the confidence threshold τ_d, N is the number of candidates remaining after non-maximum suppression, i indexes an individual candidate, B_i denotes its predicted bounding box, and C_i^{det} denotes the corresponding

detector confidence. The threshold τ_d is used for candidate retention rather than final defect acceptance. A low candidate-retention threshold is used to reduce premature rejection of weak defects, while false-positive filtering is deferred to MEV. YOLO26s retains its original architecture and is used exclusively for recall-oriented candidate generation.

3.3. ROI-Guided Multi-Task SAM

3.3.1. Candidate-Centered ROI Reconstruction

Direct whole-image SAM inference is poorly suited to extremely small defects because global resizing substantially reduces their relative spatial representation before feature encoding [29]. To mitigate this scale mismatch, a candidate-centered ROI reconstruction strategy is introduced before SAM inference, as illustrated in Figure 4a.

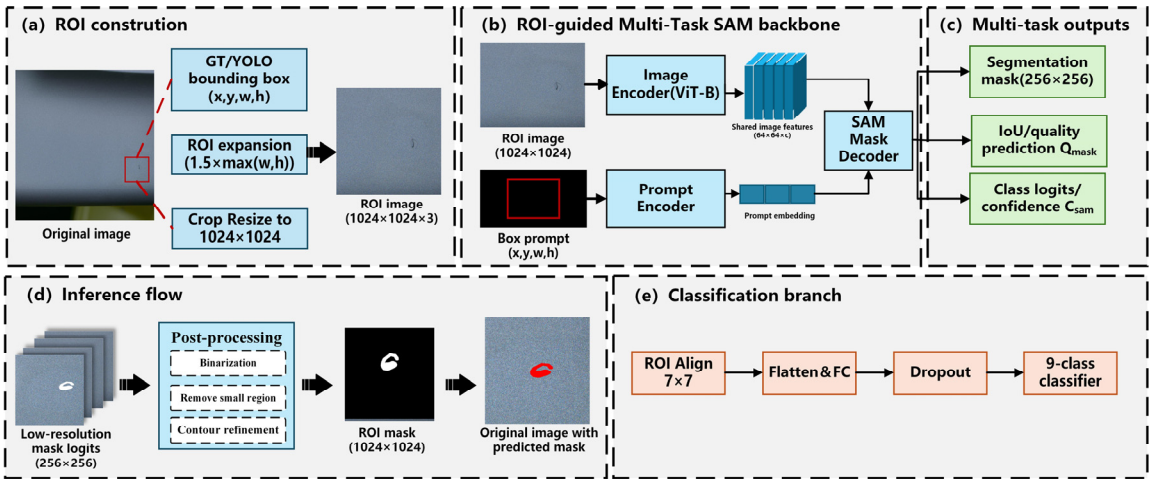


Figure 4. Architecture of the proposed ROI-guided Multi-Task SAM (ROI-MT-SAM). (**a**) Candidate-centered ROI reconstruction; (**b**) shared SAM feature encoding; (**c**) multi-task outputs including segmentation, mask-quality estimation, and defect classification; (**d**) inference and mask restoration; and (**e**) auxiliary classification branch.

For candidate $B_i = (x_i, y_i, w_i, h_i)$, let $l_i = \max(w_i, h_i)$. The side length of the square ROI is defined as

$$L_i = \max(\alpha l_i, L_{\min}) \tag{4}$$

where α controls the amount of contextual expansion and $L_{\min}$ specifies the minimum allowable ROI size. In the adopted configuration, $\alpha = 1.5$ and $L_{\min} = 256$ pixels.

The ROI is centered on the detected candidate, adjusted at image boundaries, cropped from the original-resolution image, and resized to 1024 × 1024 before SAM encoding. The detector box is transformed into the ROI coordinate system and used as the box prompt.

The ROI is intentionally larger than the candidate itself. An overly tight crop may remove useful surrounding texture, whereas excessive expansion reduces the magnification benefit [30]. The reconstruction therefore balances defect magnification and contextual preservation, which is critical for weak and extremely small defects [31].

3.3.2. Shared Feature Encoding and Multi-Task Prediction

ROI-MT-SAM is built on SAM ViT-B. The reconstructed ROI is encoded by the SAM image encoder, while the transformed candidate box is processed by the prompt encoder. The resulting embeddings are passed to the mask decoder to obtain the defect segmentation mask. The SAM-native IoU prediction head provides the mask-quality estimate $\left(Q_i^{mask}\right)$, avoiding the need for an additional quality-assessment network [32].

To provide semantic evidence for candidate verification, a lightweight classification branch is attached to the shared SAM image representation [33]. ROIAlign extracts the candidate-level feature region, which is followed by feature pooling, a fully connected layer, dropout, and a nine-class classifier. The semantic confidence is defined as

$$C_i^{sem} = \max_k p_{i,k},\ p_i = Softmax(z_i) \tag{5}$$

where z_i denotes the classification-logit vector for candidate i, and $p_i = Softmax(z_i)$ represents its nine-class probability distribution. The probability assigned to category k is denoted by $p_{i,k}$, and the semantic confidence is defined as $C_i^{sem} = \max_k p_{i,k}$. During inference, the predicted mask is restored to the ROI resolution and mapped back to its original image coordinates.

3.3.3. Multi-Task Optimization

ROI-MT-SAM is optimized using a multi-task objective that jointly supervises defect segmentation, semantic classification, and mask-quality estimation:

$$L = L_{BCE} + L_{Dice} + \lambda_c L_{cls} + \lambda_q L_{qual} \tag{6}$$

where L denotes the total training loss; L_{BCE} and L_{Dice} supervise pixel-level defect segmentation; L_{cls} denotes the nine-class semantic classification loss; and L_{qual} supervises the predicted mask-quality score. The weighting coefficients λ_c and λ_q control the relative contributions of the classification and mask-quality objectives, respectively.

The Dice loss is defined as

$$L_{Dice} = 1 - \frac{2\sum_j p_j g_i + \varepsilon}{\sum_j p_j + \sum_j g_j + \varepsilon} \tag{7}$$

where j indexes the pixels within the predicted mask, p_j denotes the predicted foreground probability at pixel j, g_j denotes the corresponding ground-truth mask value, and ε is a small constant introduced for numerical stability.

The mask-quality branch is supervised using the IoU between the binarized predicted mask and its corresponding ground-truth mask.

$$q_i^* = IoU(\hat{M}_i^b, M_i^*), L_{qual} = (Q_i^{mask} - q_i^*)^2 \tag{8}$$

where q_i^* denotes the target mask-quality score for candidate i, $\widehat{M}_i^b$ is the binarized predicted mask, M_i^* is the corresponding ground-truth mask, and Q_i^{mask} is the mask-quality score predicted by ROI-MT-SAM. L_{qual} denotes the mean-squared error used to supervise the mask-quality prediction. The quality branch is therefore trained to estimate the spatial reliability of the predicted mask rather than an independent image-level confidence.

During training, the pretrained SAM image encoder and prompt encoder remain frozen, while the mask decoder and the semantic classification branch are optimized. This configuration adapts the task-specific prediction components while retaining the pretrained encoders.

3.4. Multi-Evidence Verification

The recall-preserving detector intentionally retains uncertain candidates to reduce irreversible early misses, but this also introduces false-positive responses caused by material textures, reflections, edges, and other local variations [34]. Rather than suppressing these candidates by further increasing the YOLO threshold, the final stage performs secondary verification using complementary evidence from ROI-MT-SAM [35].

For each candidate i, three confidence cues are considered: the YOLO detection confidence C_i^{det}, the semantic confidence C_i^{sem} predicted by ROI-MT-SAM, and the mask-quality estimate Q_i^{mask}. They are combined into a multi-evidence verification score:

$$S_i^{MEV} = w_d C_i^{\text{det}} + w_c C_i^{sem} + w_q Q_i^{mask},\ w_d + w_c + w_q = 1 \tag{9}$$

where S_i^{MEV} denotes the fused verification score of candidate i, and w_d, w_c, and w_q are the weighting coefficients assigned to the detector confidence, semantic confidence, and mask-quality estimate, respectively. The weights are constrained to be non-negative and to sum to one.

The three scores represent distinct aspects of each candidate: detector confidence reflects the initial detection response, semantic confidence quantifies the certainty of the predicted defect category, and mask quality estimates segmentation overlap. Their weighted combination provides multiple criteria for candidate verification [36].

A lightweight spatial-consistency gate is further applied to remove degenerate or spatially inconsistent mask predictions. Two quantities are examined: the mask fill ratio, defined as the fraction of the YOLO box occupied by the predicted mask, and the mask inside fraction, defined as the proportion of the predicted mask located inside the YOLO box. In the adopted configuration, a candidate passes the gate when the mask fill ratio lies between 0.001 and 1.0 and at least 10% of the predicted mask falls inside the detector box. These deliberately permissive constraints remove clearly unsupported masks without imposing strong shape priors on defects with heterogeneous morphologies. Let $G_i^{\mathrm{sp}} \in \{0, 1\}$ denote whether candidate i satisfies the spatial-consistency gate. The final decision is then defined as

$$\hat{y}_i = \begin{cases} 1, S_i^{MEV} \geq \mathrm{T} \text{ and } G_i^{sp} = 1 \\ 0, \textit{otherwise} \end{cases} \tag{10}$$

where $\hat{y}_i$ denotes the final binary decision for candidate i, with $\hat{y}_i = 1$ indicating a retained defect and $\hat{y}_i = 0$ indicating a rejected candidate; S_i^{MEV} is the multi-evidence verification score defined in Equation (9); T denotes the final verification threshold; and G_i^{sp} indicates whether the predicted mask satisfies the spatial-consistency criterion.

The fusion weights and verification threshold were selected on the fixed validation set. An initial grid search minimized false positives subject to a maximum recall reduction of 0.02 relative to the YOLO26s baseline. Because no evaluated configuration satisfied this constraint, the selection criterion was changed to maximizing F1-score. The resulting configuration was $w_d = 0.30$, $w_c = 0.10$, $w_q = 0.60$, and $T = 0.66$. The search covered nonnegative weights in increments of 0.10 and thresholds from 0.05 to 0.95 in increments of 0.01. Fusion-weight and threshold sensitivity analyses are reported in Supplementary Tables S3 and S5, respectively.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Implementation Details

Unless otherwise specified, quantitative experiments were conducted on the fixed validation set containing 400 images and 1131 annotated defect instances. The 52 original test images were kept independent of model training and parameter selection.

YOLO26s was trained for 60 epochs with an input resolution of 960 × 960 and a batch size of 4. AdamW was used with an initial learning rate of 3×10^{-4}, a final learning-rate factor of 0.01, and a weight decay of 1×10^{-3}. The box-regression, classification, and distribution-focal loss weights were set to 8.5, 0.75, and 1.60, respectively. Mosaic augmentation was applied with a probability of 0.40, whereas MixUp was disabled. The random seed was fixed to 2024. Consistent with its role as a category-agnostic candidate generator, the nine defect categories were merged into a single defect class during YOLO training.

ROI-MT-SAM was built on SAM ViT-B. The pretrained image encoder and prompt encoder were frozen, while the mask decoder and auxiliary classification branch were optimized for 30 epochs. Training used a batch size of 2 with a gradient-accumulation factor of 4. The learning rates of the mask decoder and classification branch were 1×10^{-5}

and 1×10^{-4}, respectively, with a weight decay of 1×10^{-4}. Label smoothing of 0.05 and bounding-box jitter of 0.10 were applied during training. The ROI expansion factor and minimum ROI size were set to 1.5 and 256 pixels, respectively, based on the ablation results in Section 4.3.2.

For candidate generation, the YOLO confidence threshold was set to 0.06 and the NMS IoU threshold to 0.55. Predicted boxes were matched to ground-truth instances using an IoU threshold of 0.50. These settings were kept fixed in the subsequent ROI-MT-SAM and MEV experiments unless otherwise stated.

All experiments were conducted under Windows 10 using Python 3.9.25, PyTorch 2.5.1, CUDA 12.1, and cuDNN 9.1. The experimental workstation was equipped with an Intel Xeon Gold 5317 CPU @ 3.00 GHz, 64 GB of system memory, and an NVIDIA GeForce RTX 4070 Ti GPU with 12 GB of memory.

4.1.2. Evaluation Metrics

Different metrics were used according to the role of each stage. Candidate-generation performance was evaluated using precision, recall, F1-score, mAP_{50}, and $mAP_{50:95}$, with recall emphasized because missed candidates cannot be recovered downstream.

Segmentation performance was evaluated using mIoU, mean Dice coefficient (mDice), pixel-level precision, and pixel-level recall. Scale-specific mIoU was additionally reported for small, medium, and large defects. Semantic classification was evaluated using classification accuracy and Macro-F1.

For MEV, true positives (TP), false positives (FP), false negatives (FN), precision, recall, F1-score, and the reduction in false positives relative to the original YOLO candidate set were reported. Miss rate was defined as $1 - R$. For MEV, TP, FP, FN, precision, recall, F1-score, and false-positive reduction were reported. The miss rate was defined as $1 - recall$.

4.2. Evaluation of High-Recall Candidate Detection

Figure 5 presents the YOLO26s training curves over 60 epochs. The box-regression and classification losses decrease progressively, while precision, recall, mAP_{50}, and $mAP_{50:95}$ approach stable values in later epochs, without a pronounced divergence between the training and validation losses.

YOLO26s was compared with YOLOv8s, YOLOv10s, YOLO11s, RT-DETR-R18, and two representative two-stage detectors, Faster R-CNN-R50-FPN and Faster R-CNN-R101-FPN, under the same data partition and evaluation protocol. The quantitative results are summarized in Table 2.

Table 2. Quantitative comparison of candidate-detection models.

Model	Precision	Recall	mAP0.5	mAP0.5:0.95	Params (M)	FLOPs (G)	FPS
YOLOv8s	0.7842	0.9065	0.9284	0.6837	11.1	28.6	115
YOLOv10s	0.7968	0.9172	0.9416	0.7025	8.0	24.4	122
YOLO11s	0.8053	0.9261	0.9508	0.7154	9.4	22.5	128
RT-DETR-R18	0.8006	0.9138	0.9447	0.7083	20.3	55.9	75
Faster R-CNN-R50-FPN	0.7981	0.9185	0.9437	0.7041	41.3	280	24
Faster R-CNN-R101-FPN	0.8147	0.9314	0.9562	0.7264	60.8	390	17
YOLO26s	0.8125	0.9390	0.9623	0.7363	9.9	26.1	69

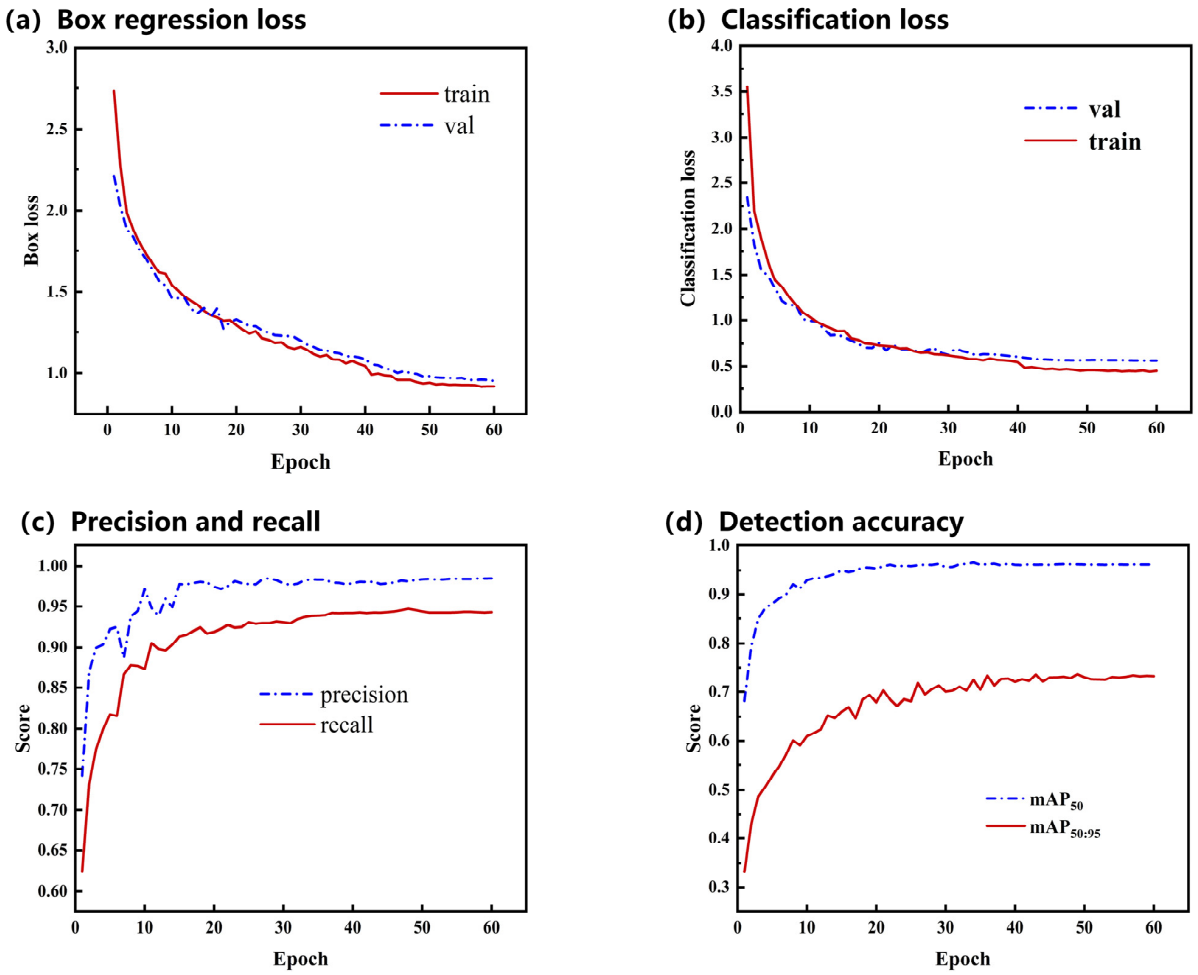


Figure 5. Training curves of YOLO26s over 60 epochs, including box-regression loss, classification loss, precision, recall, mAP_{50}, and $mAP_{50:95}$.

As shown in Table 2, YOLO26s achieves the highest recall (0.9390) and mAP_{50} (0.9623) among the evaluated detectors. Its recall exceeds that of Faster R-CNN-R50-FPN and Faster R-CNN-R101-FPN by 0.0205 and 0.0076, respectively. Although its inference speed (69 FPS) is lower than that of the other YOLO variants, its superior candidate recall supports its selection as the front-end detector, where missed defects cannot be recovered by downstream verification. On the fixed validation set, YOLO26s detects 1062 of 1131 annotated defects, with 245 false positives. These candidate-level predictions form the baseline for subsequent MEV evaluation.

4.3. ROI-MT-SAM Evaluation

4.3.1. Effect of Candidate-Centered ROI Reconstruction

The effect of candidate-centered reconstruction was assessed by comparing whole-image SAM and ROI-MT-SAM on the fixed validation set. Table 3 reports their segmentation and semantic classification performance.

Table 3. Comparison of segmentation and semantic classification performance between whole-image SAM and ROI-MT-SAM.

Model	Seg.mIoU	Seg.mDice	Seg.Precision	Seg.Recall	Cls.Acc	Cls.MacroF1
Whole-image SAM	0.3918	0.5174	0.6753	0.4610	0.7401	0.6717
ROI-MT-SAM	0.8869	0.9384	0.9353	0.9453	0.6940	0.6488

As shown in Table 3, ROI-MT-SAM improves mIoU from 0.3918 to 0.8869 and mDice from 0.5174 to 0.9384. Pixel-level precision and recall also increase, indicating more accurate defect delineation under the evaluated protocol. In contrast, classification accuracy decreases from 0.7401 to 0.6940, while Macro-F1 declines from 0.6717 to 0.6488. These results reveal a trade-off between improved local segmentation and reduced semantic discrimination. The paired source-image-level bootstrap analysis further supports the observed mIoU improvement in the 95% CI for the difference: [0.4213, 0.5625]; Supplementary Table S2.

Figure 6 illustrates representative cases in which ROI-MT-SAM produces more localized defect masks than whole-image SAM, consistent with the quantitative segmentation results.

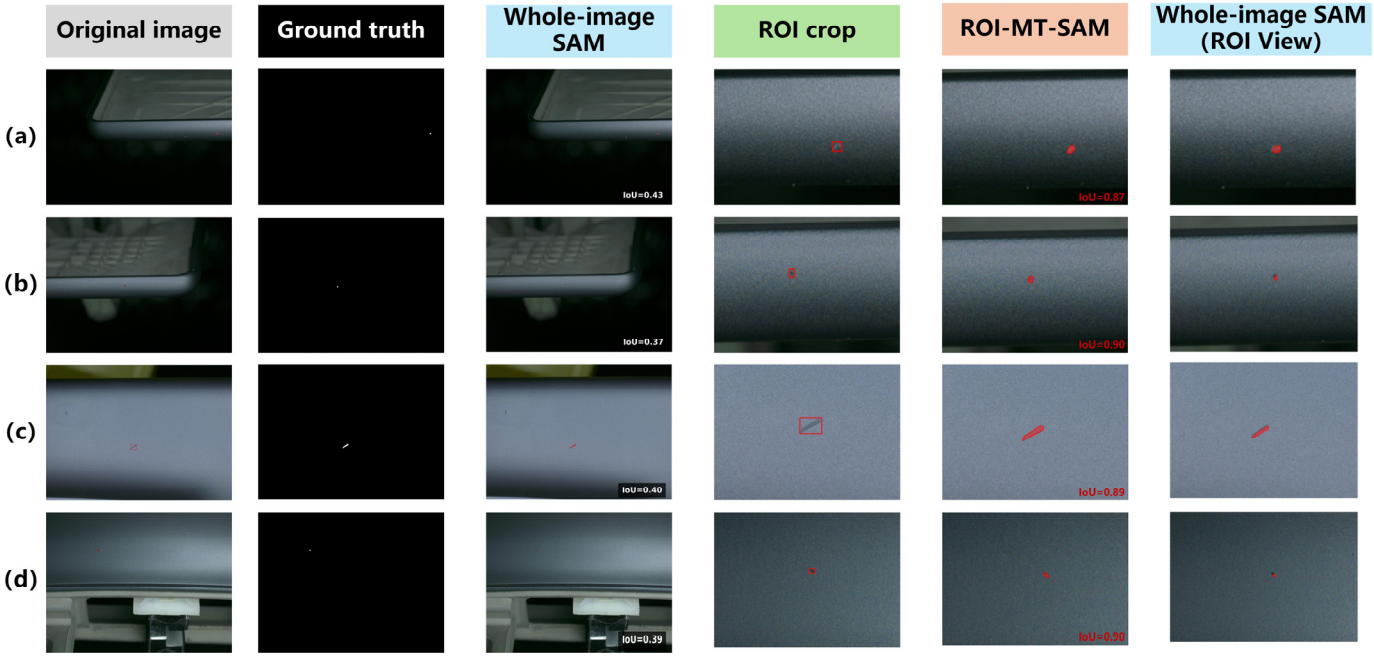


Figure 6. Qualitative comparison of whole-image SAM and ROI-MT-SAM. Columns show the original image, ground-truth mask, whole-image SAM prediction, ROI crop, ROI-MT-SAM prediction, and whole-image SAM prediction in the corresponding ROI. The ROI views use identical spatial scales for direct comparison. (**a**–**d**) Four representative tiny-defect inspection cases.

The improvement in segmentation is accompanied by reduced semantic classification performance, suggesting that finer local representation does not necessarily improve category discrimination. ROI cropping may exclude broader contextual cues, while the frozen SAM encoder may limit adaptation to subtle inter-class differences. The limited representation of rare categories may further contribute to classification errors. Representative discrepancies are shown in Supplementary Figure S1; however, the relative contributions of these factors remain undetermined.

Figure 7 presents ROI-MT-SAM segmentation examples across all nine defect categories, including point-like, elongated, and irregular defects.

4.3.2. Component Ablation

A controlled ablation study was conducted to assess the effects of ROI expansion, minimum ROI size, and mask-quality supervision. All other settings were kept unchanged, and the results are summarized in Table 4.

Table 4 shows that ROI expansion affects segmentation accuracy. Increasing the expansion factor from 1.0 to 1.5 improves mIoU from 0.8586 to 0.8869, whereas further expansion to 2.0 reduces it to 0.8748. This non-monotonic trend is consistent with a trade-off between

defect magnification and contextual coverage. Increasing the minimum ROI size from 256 to 384 pixels also reduces overall mIoU from 0.887 to 0.867, despite a slight improvement for small defects. These results support the selected configuration of $\alpha = 1.5$ and $L_{\min} = 256$ within the evaluated parameter range.

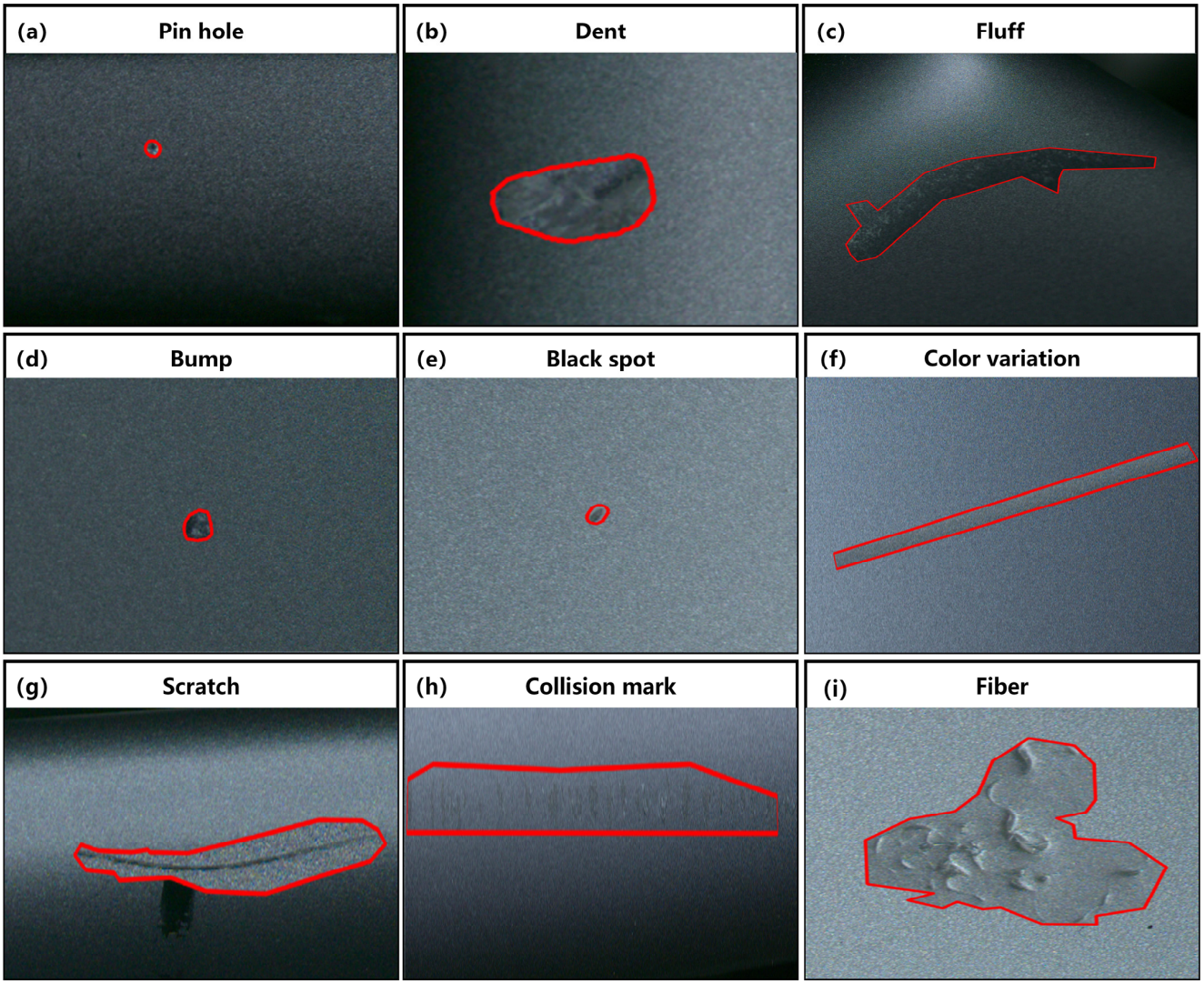


Figure 7. Qualitative segmentation results of ROI-MT-SAM across nine automotive interior defect categories: (**a**) pin hole, (**b**) dent, (**c**) fluff, (**d**) bump, (**e**) black spot, (**f**) color variation, (**g**) scratch, (**h**) collision mark, and (**i**) fiber. Red contours indicate the predicted segmentation boundaries.

Table 4. Ablation study of ROI-MT-SAM.

Model	ROI Scale	Min-ROI	Quality Supervision	mIoU	mDice	Small-mIoU	Medium-mIoU	Large-mIoU
A	1	256	√	0.8586	0.9240	0.8412	0.8681	0.8865
B	1.5	256	×	0.8837	0.9355	0.8694	0.8923	0.8849
C	1.5	256	√	0.8869	0.9384	0.8722	0.8948	0.8879
D	2	256	√	0.8748	0.9298	0.8589	0.8865	0.8931
E	1.5	384	√	0.8671	0.9241	0.8787	0.8781	0.8918

Comparing Models B and C, adding mask-quality supervision increases mIoU from 0.883 to 0.887 and mDice from 0.935 to 0.938. Model C yields the highest overall mIoU among the evaluated configurations and is adopted for subsequent experiments.

4.3.3. Effect of Residual-Transfer Augmentation

Residual-transfer augmentation was compared with no augmentation, conventional geometric–photometric augmentation, and naive RGB copy-and-paste under identical data partitions, model architecture, training schedules, and evaluation settings. Table 5 summarizes the segmentation results.

Table 5. Ablation study of different data augmentation strategies.

Augmentation Strategy	mIoU	Precision	Recall	F1-Score
No Augmentation	0.8214	0.8841	0.9075	0.8956
Geometric–Photometric	0.8491	0.9050	0.9242	0.9145
Naive RGB Copy–Paste	0.8627	0.9192	0.9318	0.9255
Residual-Transfer (Ours)	0.8869	0.9353	0.9453	0.9403

As shown in Table 5, residual-transfer augmentation achieves the highest mIoU (0.8869) and F1-score (0.9403) among the evaluated strategies. Compared with geometric–photometric augmentation and naive RGB copy-and-paste, it improves mIoU by 0.0378 and 0.0242, respectively. These improvements suggest that transferring localized defect residuals is more effective than the evaluated alternatives for segmentation learning under the present data conditions. Paired bootstrap confidence intervals for the corresponding mIoU differences are reported in Supplementary Table S2.

4.4. Multi-Evidence Verification

4.4.1. Contribution of Multi-Evidence Verification

This experiment evaluates whether MEV improves candidate discrimination beyond YOLO26s confidence alone. All strategies are compared on the fixed validation set using the same initial candidate pool and evaluation protocol.

All strategies were evaluated on the same fixed validation set. The YOLO baseline denotes the original candidate set without secondary verification.

As shown in Table 6, Full MEV reduces false positives from 245 to 46 (81.22%) relative to the YOLO26s baseline, while recall decreases from 0.9390 to 0.8939. At the matched recall of 0.8939, MEV further reduces false positives from 109 to 46 (57.80%) compared with YOLO confidence-only thresholding. This matched-recall improvement indicates that the combined verification cues provide additional discriminatory value beyond detector confidence alone. Statistical uncertainty in the corresponding F1-score improvements is assessed using paired source-image-level bootstrap analysis (Supplementary Table S2).

Table 6. Comparison of candidate-verification strategies.

Strategy	TP	FP	FN	Precision	Recall	F1	FP Reduction
YOLO baseline	1062	245	69	0.8125	0.9390	0.8712	-
YOLO threshold only	1011	109	120	0.9027	0.8939	0.8983	0.5551
YOLO + SAM Quality	1011	93	120	0.9158	0.8939	0.9047	0.6204
YOLO + SAM Class	1011	68	120	0.9370	0.8939	0.9149	0.7224
Full MEV	1011	46	120	0.9565	0.8939	0.9241	0.8122

Among the individual evidence configurations, YOLO + SAM-Class yields fewer false positives than YOLO + SAM-Quality (68 versus 93). Combining semantic confidence, mask-quality estimation, and spatial consistency further reduces false positives to 46 and achieves the highest F1-score (0.9241). These results support the complementary contribution of the verification cues under the evaluated conditions. Combining detector confidence, semantic confidence, mask quality, and spatial consistency in the full MEV reduces the false-positive count to 46. Relative to the original candidate set, this corresponds to an 81.22% reduction in false positives, while precision and F1 increase to 0.9565 and 0.9241, respectively. The contribution of the spatial-consistency gate is further isolated in Supplementary Table S4. Under fixed fusion weights and thresholds, the gate reduces false positives from 61 to 46, while true positives decrease from 1017 to 1011.

The final MEV configuration was selected by maximizing validation-set F1-score after the initial recall constraint proved infeasible, as described in Section 3.4. The selected weights were $w_d = 0.30$, $w_c = 0.10$, and $w_q = 0.60$, with a verification threshold of $T = 0.66$. The sensitivity of this operating point to fusion weights and decision thresholds is examined in Supplementary Tables S3 and S5.

MEV therefore does not eliminate the intrinsic trade-off between false-positive suppression and missed-defect risk; instead, it provides a more informative mechanism for controlling this trade-off using complementary evidence. Figure 8 shows representative candidate-level outcomes.

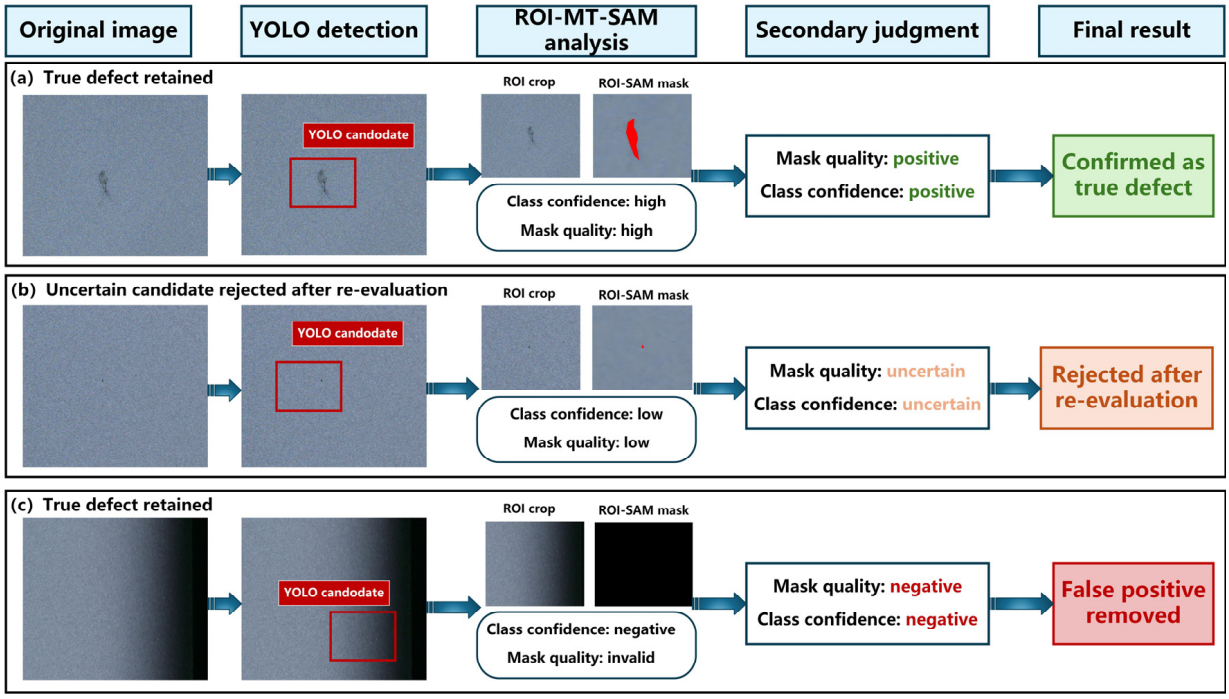


Figure 8. Representative Joint Decision cases. (**a**) A true defect retained after ROI-MT-SAM verification; (**b**) an uncertain candidate rejected after re-evaluation; and (**c**) a false-positive candidate removed by the secondary decision.

Although MEV eliminates 199 false positives relative to the YOLO26s baseline, it introduces 51 additional false negatives on the fixed validation set. This trade-off is consequential in industrial inspection, where missed defects may carry greater costs than unnecessary rejection. Therefore, the selected F1-oriented operating point should not be interpreted as universally optimal for deployment.

4.4.2. Influence of the Decision Threshold

Figure 9a shows the effect of the MEV decision threshold on the miss rate, false-positive fraction, and F1-score. Increasing the threshold suppresses false positives but increases missed detections. The highest F1-score (0.9241) among the evaluated thresholds occurs at T = 0.66, beyond which further false-positive reduction is accompanied by a greater loss of recall. The corresponding detection counts and performance metrics across the evaluated thresholds are provided in Supplementary Table S5.

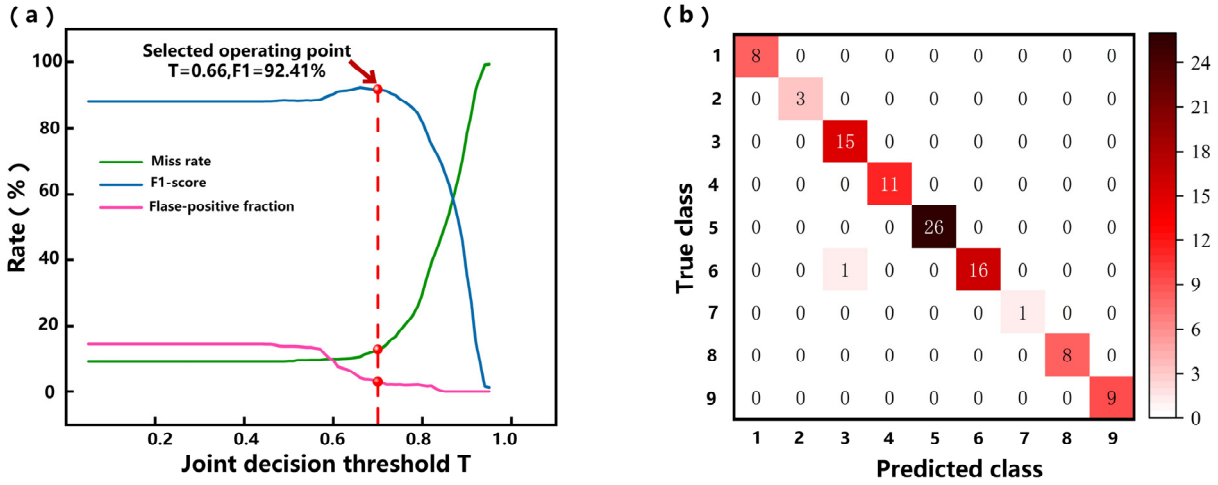


Figure 9. Decision behavior and independent-test semantic evaluation. (**a**) Effect of the MEV threshold on miss rate, F1-score, and false-positive fraction; (**b**) confusion matrix of the semantic classification branch on the independent test set comprising 52 original images and 98 defect instances. Class labels 1–9 denote bump, dent, scratch, color variation, fiber, collision mark, pin hole, fluff, and black spot, respectively.

Figure 9b presents the semantic classification confusion matrix for the 98 annotated defect instances in the independent test set. This matrix evaluates category prediction separately from end-to-end defect detection and therefore does not account for missed detections or false-positive candidates.

Class-wise detection and segmentation results on the independent test set are reported in Supplementary Table S7. Across 52 original images containing 98 annotated defects, the complete framework achieves a precision of 0.9355, recall of 0.8878, and F1-score of 0.9100. These values are lower than those obtained on the augmented validation set (0.9565, 0.8939, and 0.9241, respectively). Class-wise performance varies, with particularly limited evidence for rare categories such as dent and pin hole. Given the small number of independent test instances, these findings provide preliminary evidence of performance on unseen source images rather than conclusive cross-domain generalization.

Supplementary Figure S2 compares the precision–recall curves of YOLO26s and MEV across decision thresholds. At a matched recall of 0.8939, MEV achieves higher precision than YOLO confidence-only thresholding (0.9565 versus 0.9027), consistent with the operating-point comparison in Table 6.

Supplementary Table S6 reports class-wise false-negative counts before and after MEV, with the total increasing from 69 to 120. The additional missed detections are unevenly distributed across defect categories, indicating that a single global threshold may not adequately reflect category-specific inspection requirements. Risk-aware or category-dependent verification warrants further investigation.

4.5. Computational Efficiency

The computational complexity and inference latency of the proposed framework are summarized in Table 7. Latency was measured separately for YOLO26s and ROI-MT-SAM on the hardware specified in Section 4.1.1, with an average of 3.27 candidate ROIs processed per image.

Table 7. Model complexity and stage-wise inference latency.

Method	Params (M)	Model Size (MB)	Avg. ROIs	Latency (ms/Image)
YOLO26s	9.95	19.40	-	14.49
ROI-MT-SAM	93.77	385.15	3.27	505.96
Full framework	103.72	404.55	3.27	520.45

As shown in Table 7, YOLO26s requires 14.49 ms per image, whereas ROI-MT-SAM accounts for 505.96 ms, approximately 97.2% of the combined stage-wise latency. The estimated total of 520.45 ms per image is the sum of the two measured stage averages rather than a directly measured end-to-end latency and excludes data transfer, preprocessing, post-processing, and MEV computation overhead. ROI-level SAM inference therefore constitutes the primary computational bottleneck.

The current computational cost limits the applicability of the framework to high-throughput inspection. As an illustrative upper-bound estimate, assuming ideal parallel processing of the average 3.27 candidate ROIs with negligible batching overhead, the ROI-MT-SAM latency would decrease from 505.96 to approximately 154.73 ms per image. The corresponding combined stage-wise latency would be approximately 169.22 ms per image, implying a theoretical speed-up of 3.08×. This estimate assumes perfect parallelization and does not represent measured batched inference performance. Potential improvements include batched ROI processing, lightweight SAM architectures, and mixed-precision inference, although their effectiveness requires experimental validation. Under the evaluated configuration, the framework is better suited to offline or selectively triggered inspection. Future work will assess latency reduction alongside its effects on segmentation accuracy and defect recall.

5. Discussion and Conclusions

This study investigates a coarse-to-fine framework that separates recall-oriented candidate generation from evidence-based final verification for automotive interior surface-defect inspection. Its central premise is that prematurely rejected defects cannot be recovered downstream, whereas retained false-positive candidates can be further evaluated using complementary geometric and semantic evidence.

Candidate-centered ROI reconstruction yields the largest observed improvement in segmentation performance, increasing mIoU from 0.3918 to 0.8869 relative to whole-image SAM. This result is consistent with the hypothesis that pre-encoding ROI reconstruction mitigates the loss of fine-scale defect information during global image resizing. The ablation results further indicate that segmentation performance depends on the balance between defect magnification and contextual preservation, rather than on ROI enlargement alone.

The segmentation gains are accompanied by reduced semantic classification performance, highlighting the distinct requirements of geometric delineation and category recognition. While local ROI features support precise boundary extraction, semantic discrimination may depend on broader material and structural context. This finding motivates future integration of local defect features with global contextual information.

MEV demonstrates that complementary verification cues can improve candidate discrimination beyond detector confidence alone. At a matched recall of 0.8939, MEV reduces false positives from 109 to 46 compared with YOLO confidence-only thresholding, indicating that the improvement is not solely attributable to more restrictive candidate rejection. Nevertheless, the reduction in recall relative to the original YOLO baseline highlights the need for application-specific operating thresholds that account for the consequences of missed defects.

Several limitations remain. First, the dataset comprises 152 independent source images, including 52 test images with 98 annotated defects. Although augmentation increases training diversity, it does not provide additional independent observations. The limited test size, class imbalance, and single-domain data collection constrain the generalizability of the findings. Evaluation on larger datasets covering diverse materials, imaging conditions, and production environments is therefore necessary. Second, the semantic classification branch relies primarily on local ROI features, potentially limiting its ability to distinguish defects requiring broader contextual information. Future work will investigate global–local feature integration to improve semantic discrimination without compromising fine-grained segmentation. Third, the computational cost of ROI-MT-SAM limits the framework's suitability for high-throughput inspection. Future work will investigate lightweight segmentation architectures and batched ROI processing, with attention to the trade-off between inference efficiency and inspection accuracy.

Overall, this study demonstrates the potential of decoupling candidate generation from evidence-based verification for tiny surface-defect inspection. The results highlight the benefits of candidate-centered segmentation and complementary verification cues, while revealing trade-offs involving semantic discrimination, defect recall, and computational efficiency. Further validation on independent industrial datasets is needed to establish the robustness and practical applicability of the framework.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15215142/s1, Figure S1: Representative semantic classification discrepancies between whole-image SAM and ROI-MT-SAM; Figure S2: Precision–recall comparison between YOLO26s and MEV; Table S1: Implementation parameters of residual-transfer augmentation; Table S2: Paired source-image-level bootstrap analysis of performance differences; Table S3: Sensitivity analysis of MEV fusion weights; Table S4: Ablation study of the spatial-consistency gate in the proposed MEV framework; Table S5: Sensitivity analysis of the MEV decision threshold under fixed fusion weights; Table S6: Class-wise false-negative analysis before and after MEV on the validation set; Table S7: Class-wise detection and segmentation performance of the proposed framework on the independent test set.

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

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The dataset is not publicly available due to confidentiality restrictions associated with industrial production data.

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

References

1. Ameri, R.; Hsu, C.-C.; Band, S.S. A systematic review of deep learning approaches for surface defect detection in industrial applications. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107717. [CrossRef]
2. Liu, P.; Yuan, X.; Han, Q.; Xing, B.; Hu, X.; Zhang, J. Micro-defect Varifocal Network: Channel attention and spatial feature fusion for turbine blade surface micro-defect detection. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108075. [CrossRef]
3. Yang, B.; Liu, Z.; Duan, G.; Tan, J. Residual shape adaptive dense-nested Unet: Redesign the long lateral skip connections for metal surface tiny defect inspection. *Pattern Recognit.* **2024**, *147*, 110073. [CrossRef]
4. Zhang, Z.; Zhou, Y.; Ma, J. Defect-LoRA: Controllable defect data augmentation based on low-rank adaptation for surface defect recognition under limited data. *Expert Syst. Appl.* **2025**, *268*, 126251. [CrossRef]
5. Cheng, Z.; Sun, H.; Cao, Y.; Cao, W.; Wang, J.; Yuan, G.; Zheng, J. Pyramid cross attention network for pixel-wise surface defect detection. *NDT E Int.* **2024**, *143*, 103053. [CrossRef]
6. Su, J.; Luo, Q.; Yang, C.; Gui, W.; Silven, O.; Liu, L. PMSA-DyTr: Prior-Modulated and Semantic-Aligned Dynamic Transformer for Strip Steel Defect Detection. *IEEE Trans. Ind. Inform.* **2024**, *20*, 6684–6695. [CrossRef]
7. Yan, F.; Jiang, X.; Lu, Y.; Cui, L.; Cao, J.; Xu, M. Global context guided refinement and aggregation network for lightweight surface defect detection. *Pattern Recognit.* **2025**, *173*, 112893. [CrossRef]
8. Kuttuva, G.V.; Prawin, J. Structural defect segmentation using a semi-supervised algorithm integrating YOLO and the segment anything model. *Autom. Constr.* **2026**, *182*, 106709. [CrossRef]
9. Lin, S.; Wang, C.; Li, J.; He, Z. STREAM: Spatial-tuned refinement with enhanced adaptive masking for few-shot surface defect segmentation. *Expert Syst. Appl.* **2026**, *333*, 133750. [CrossRef]
10. Yan, F.; Zhang, D.; Wu, Z.; Du, Q. Improved Droop Control Strategy Based on Reactive Power Sharing Error Rate. In Proceedings of the IEEE 16th International Symposium on Power Electronics for Distributed Generation Systems, Nanjing, China, 22–25 June 2025; pp. 1–5.
11. Su, J.; Luo, Q.; Gui, W.; Yang, C. Few-Shot Parameter Efficient Finetuning for SAM in Salient Steel Surface Defect Detection. *IEEE Trans. Ind. Inform.* **2025**, *21*, 7742–7753. [CrossRef]
12. Zhang, S.; Jiang, L.; Ge, L.; Wu, C.; Wang, Y.; Lu, K.; Zhang, C. SGTP-Net: Semantic Guidance and Texture Priors-Based Dual-Branch Segmentation Network for Surface Defect Detection. *IEEE Trans. Autom. Sci. Eng.* **2025**, *23*, 417–432. [CrossRef]
13. Liu, W.; Hu, J.; Qi, J. Coarse-to-fine vision-based welding spot anomaly detection in production lines of body-in-white. *J. Manuf. Syst.* **2025**, *81*, 144–154. [CrossRef]
14. Wang, Z.; Song, Y. FI2Net: A two-stage network with information fusion and integration mechanisms for surface defect detection. *Meas. Sci. Technol.* **2025**, *36*, 065401. [CrossRef]
15. Bian, J.; Feng, M.; Dong, W.; Wu, Z.; Luo, J.; Wu, F.; Wang, Y.; Shi, G. Locally Aware Visual State Space for Small Defect Segmentation in Complex Component Images. *IEEE Trans. Ind. Inform.* **2025**, *21*, 5965–5976. [CrossRef]
16. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. Segment anything model for medical images? *Med. Image Anal.* **2024**, *92*, 103061. [CrossRef] [PubMed]
17. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.* **2023**, *89*, 102918. [CrossRef] [PubMed]
18. Zhang, Y.; Shen, Z.; Jiao, R. Segment anything model for medical image segmentation: Current applications and future directions. *Comput. Biol. Med.* **2024**, *171*, 108238. [CrossRef] [PubMed]
19. Xu, M.; Wei, J.; Feng, X. Two-stage encoder multi-decoder network with global-local up-sampling for defect segmentation of strip steel surface defects. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109469. [CrossRef]
20. Liu, L.-J.; Sun, S.-Q.; Karimi, H.R. A real-time surface defect detection model based on adaptive feature information selection and fusion. *Inf. Fusion* **2025**, *129*, 104041. [CrossRef]
21. Wei, J.; Shi, Q.; Zhang, Z.; Shen, F.; Luo, H. Adaptive few-shot image augmentation for fine-grained industrial defects based on region-level modeling. *Eng. Appl. Artif. Intell.* **2025**, *152*, 110695. [CrossRef]
22. Wang, X.; Li, W.; Lu, C. A mask guided cross data augmentation method for industrial defect detection. *Future Gener. Comput. Syst.* **2025**, *166*, 107676. [CrossRef]
23. Tai, Y.; Zhang, Z. Data expansion for industrial defect detection with diffusion prior. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111835. [CrossRef]
24. Tian, S.; Ma, H.; Huang, P.; Wang, X.; Li, T.; Huang, R. RADDA-Net: Residual attention-based dual discriminator adversarial network for surface defect detection. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108887. [CrossRef]
25. Yang, M.; Wu, P.; Feng, H. MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105835. [CrossRef]
26. Peng, J.; Shao, H.; Xiao, Y.; Cai, B.; Liu, B. Industrial surface defect detection and localization using multi-scale information focusing and enhancement GANomaly. *Expert Syst. Appl.* **2024**, *238*, 122361. [CrossRef]

27. Yang, M.; Liu, J.; Yang, Z.; Wu, Z. SLSG: Industrial image anomaly detection with improved feature embeddings and one-class classification. *Pattern Recognit*. **2024**, *156*, 110862. [CrossRef]
28. Lv, S.; Liang, T.; Zhang, K.; Jiang, S.; Ouyang, B.; Li, Q.; Li, X. A lightweight hierarchical aggregation task alignment network for industrial surface defect detection. *Expert Syst*. *Appl*. **2025**, *263*, 125727. [CrossRef]
29. Chen, P.; Lyu, W.; Guo, Q.; Deng, Z. Efficient aggregate distribute network for tiny defect detection. *Expert Syst*. *Appl*. **2025**, *280*, 127551. [CrossRef]
30. Liu, X.; Liu, J.; Zhang, H.; Zhang, H. Low-contrast X-ray image defect segmentation via a novel core-profile decomposition network. *Comput*. *Ind*. **2024**, *161*, 104123. [CrossRef]
31. He, Y.; Cai, C.; Chen, G.; Hu, J.; Hu, S.; Fu, J. PFEI-Net: A profound feature exploration and interaction network for ceramic substrate surface defect detection. *Expert Syst*. *Appl*. **2025**, *263*, 125741. [CrossRef]
32. Ali, M.; Wu, T.; Hu, H.; Luo, Q.; Xu, D.; Zheng, W.; Jin, N.; Yang, C.; Yao, J. A review of the Segment Anything Model (SAM) for medical image analysis: Accomplishments and perspectives. *Comput*. *Med*. *Imaging Graph*. **2025**, *119*, 102473. [CrossRef] [PubMed]
33. Jie, L.; Zhang, H. ShadowAdapter: Adapting Segment Anything Model with Auto-Prompt for shadow detection. *Expert Syst*. *Appl*. **2025**, *273*, 126809. [CrossRef]
34. Zhao, J.; Yu, K.; Miao, Y.; Wang, Y.; Ma, Y.; Zhang, J.; Zhao, J.; Qiang, Y.; Pei, B. Frequency domain nuances guided parallel transformer model for industrial anomaly localization. *Eng*. *Appl*. *Artif*. *Intell*. **2025**, *142*, 109852. [CrossRef]
35. Ma, Z.; Li, J.; Wong, W.K. Patch distance based auto-encoder for industrial anomaly detection. *Expert Syst*. *Appl*. **2025**, *270*, 126537. [CrossRef]
36. Li, J.; Wen, S.; Karimi, H.R. Zoom-Anomaly: Multimodal vision-Language fusion industrial anomaly detection with synthetic data. *Inf*. *Fusion* **2025**, *127*, 103910. [CrossRef]

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