UniPCB: A Generation-Assisted Vision-Based Measurement Framework for PCB Defect Inspection

Huan Zhang, Lianghong Tan, Yichu Xu, Zishan Su, Jiangzhong Cao†, Huanqi Wu, Linwei Zhu, and Xu Zhang†

***Abstract*—Automated optical inspection (AOI) is a vision-based instrumentation process that detects and localizes physical defects on printed circuit boards (PCBs) from optically acquired images. Its reliability is often limited by scarce and class-imbalanced defect observations and by insufficient representation of small, low-contrast defects against dense circuit patterns. We propose UniPCB, a generation-assisted PCB inspection framework that combines controlled defect synthesis with task-specific detection. The generation branch derives heterogeneous edge, relative pseudo-depth, and text conditions from PCB images. A ScaleEncoder embeds the conditions at four U-Net resolutions, while a Condition Modulation block performs FiLM-style spatially adaptive fusion. The detection branch introduces an Inverted Residual Shift Attention block for joint global–local modeling and a Cross-level Complementary Fusion block for selective hierarchical feature integration. The generated images augment the detector training data and improve defect coverage without changing the online inspection path. Experiments using the corrected annotations of DsPCBSD+ yield an mAP@0.5 of 98.0% and an mAP@0.5:0.95 of 61.8%. Under the same generation-augmented training setting, these values exceed those of the RT-DETR baseline by 2.1 and 1.7 percentage points, respectively. The generation branch obtains an FID of 129.61 and an SSIM of 0.619. These results indicate the potential of structured synthetic data and PCB-oriented feature modeling for improving vision-based defect inspection. Code is available at https://github.com/House-yuyu/UniPCB.**

***Index Terms*—Automated optical inspection, conditional diffusion model, data augmentation, printed circuit board defect detection, vision-based measurement.**

I. Introduction

PRINTED circuit boards (PCBs) are essential components in modern electronic products, and their manufacturing quality directly affects the reliability of downstream devices. Complex fabrication processes may introduce physical defects such as shorts, opens, mouse bites, spurs, and spurious copper, which can degrade electrical connectivity and cause product failure. Automated optical inspection (AOI) can be formulated as a vision-based instrumentation process in which the camera, illumination, image-acquisition chain, and computational model jointly detect and localize these defects [1], [2]. The resulting defect class, confidence, and bounding-box location constitute the inspection output used for manufacturing quality decisions. Consequently, missed detections, false alarms, and localization errors directly affect the reliability of the inspection process. Traditional reference-image comparison and handcrafted feature matching are sensitive to lighting variations, board-layout changes, and novel defect morphologies, often degrading these outputs in practical inspection scenarios [3].

Deep learning-based detectors improve robustness by learning discriminative representations directly from data. However, their effectiveness depends on large and well-annotated defect datasets, which are difficult to obtain in PCB manufacturing. First, mature production lines usually have high yield rates, making defective samples naturally rare. Second, PCB layouts vary across products, limiting the transferability of defect samples across board types. Third, defect annotation requires domain expertise and is time-consuming, especially when dense and small defects need to be precisely localized. As a result, PCB defect datasets are often scarce and class-imbalanced, causing detectors to underfit rare defects and generalize poorly to unseen board layouts.

To alleviate this data bottleneck, traditional augmentation techniques such as rotation, flipping, and blurring have been widely adopted. Nevertheless, these operations mainly transform existing samples and cannot sufficiently model the diverse defect morphologies and local structural variations observed in real PCB images. Deep generative models provide a more flexible alternative. VAEs [4] and GANs [5] have been explored for defect synthesis, but VAEs often produce blurry details and GANs may suffer from training instability and mode collapse. Diffusion models [6]–[9] offer stronger training stability and better texture modeling ability, making them suitable for fine-grained industrial defect synthesis. However, their training on natural image datasets limits their effectiveness on industrial images such as PCBs [10], and existing PCB-oriented methods typically rely on a single control condition injected at the initial resolution, providing insufficient structural guidance for the dense, fine-grained circuit patterns unique to PCB images.

In addition to data scarcity, the feature representation capability of the detector remains critical. Transformer-based detectors such as DETR [11] and RT-DETR [12] provide strong global modeling ability. Lightweight detection designs likewise seek to balance accuracy and computational cost. Recent studies in surface defect and tiny object detection further indicate that fine-grained feature separation, relation reason-

This work was in part supported by the National Natural Science Foundation of China under Grant 62302105, in part by the Guangdong Provincial Key Laboratory of Intellectual Property & Big Data under Grant 2018B030322016. *(Corresponding author: Jiangzhong Cao, Xu Zhang.)*

Huan Zhang, Lianghong Tan, Jiangzhong Cao, and Huanqi Wu are with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: huanzhang2021@gdut.edu.cn; 656896486@qq.com; cjz510@gdut.edu.cn; 2441025664@qq.com).

Yichu Xu, Zishan Su, and Xu Zhang are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: xuyichu@whu.edu.cn; zishan.su@whu.edu.cn; zhangx0802@whu.edu.cn).

Linwei Zhu is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: lw.zhu@siat.ac.cn).

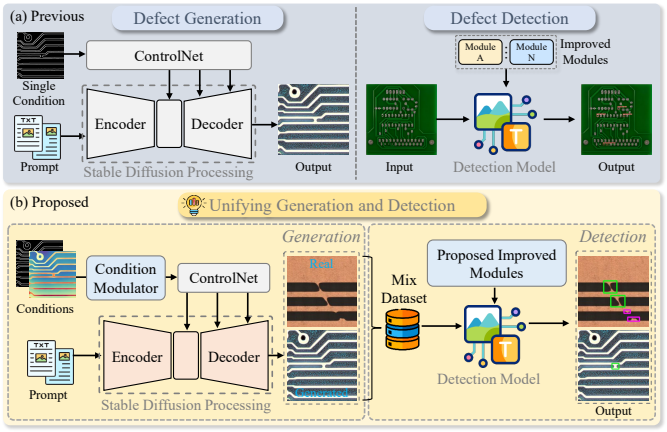


Fig. 1. Comparison between UniPCB and previous approaches for PCB defect detection. (a) Previous methods address the two challenges in isolation: generation methods rely on single-modality conditions with limited structural control, while detection methods improve architectures but remain bottlenecked by scarce training data. (b) UniPCB jointly addresses both within a unified framework: the generation branch synthesizes high-quality defect samples while the detection branch exploits the enriched data through task-specific feature modeling modules.

ing, class knowledge, and regional attention are important for detecting subtle defects under complex backgrounds [13]–[15]. Nevertheless, architecture-level improvements alone may still be constrained by insufficient and imbalanced training data [16].

These two issues often appear together in PCB inspection and jointly limit practical detection performance. On the one hand, a detection model trained on scarce and imbalanced data cannot fully leverage improvements in model architecture, as the data bottleneck fundamentally limits what the model can learn. On the other hand, a generation model producing low-fidelity or poorly controlled defect samples provides little benefit to detection, even when the detector itself is powerful. This coupling implies that addressing either challenge in isolation yields diminishing returns: the two must be resolved jointly within a coherent framework to achieve meaningful gains in PCB defect inspection.

To break this mutual bottleneck, we propose a unified generative–detection framework, termed **UniPCB**, which simultaneously enhances data quality through controlled defect synthesis and improves detection capability through task-specific architectural design. Fig. 1 compares UniPCB with previous approaches. As illustrated in Fig. 1(a), prior work treats the two problems separately: diffusion-based methods reconstruct images to identify pixel-level discrepancies but cannot generate new training data, while object detection methods improve architectures but remain fundamentally bottlenecked by scarce labeled data. As shown in Fig. 1(b), UniPCB unifies both within a single framework. On the generation side, a multi-modal conditional strategy synthesizes diverse and realistic defect samples by injecting edge, depth, and text conditions at multiple scales, ensuring structural alignment and semantic fidelity to real PCB images. On the detection side, the generated data directly serves as augmented training input. To further strengthen the detector's sensitivity to small, low-contrast defects under complex circuit backgrounds, two task-specific modules are introduced: the IRSA Block for joint global–local feature modeling, and the CLCF Block for cross-level selective feature fusion. Through this co-design of generation and detection, each component amplifies the benefit of the other.

In summary, our main contributions are as follows.

❑ We propose UniPCB as a generation-assisted vision-based measurement framework for PCB defect inspection. It couples controlled defect synthesis with a task-specific detector and evaluates how structured synthetic samples and PCB-oriented feature modeling improve the reliability of defect classification and localization.

❑ On the generation side, we design a Multi-modal Condition Generator that constructs parallel edge, relative pseudo-depth, and text conditions, providing heterogeneous and structured control signals to alleviate data scarcity and class imbalance. Building on this, we propose a latent diffusion-based synthesis network with a ScaleEncoder that embeds conditions at four resolutions and a CondMod block that jointly modulates noise and textual features, improving sample fidelity and structural consistency.

❑ On the detection side, we propose an Inverted Residual Shift Attention Block that couples self-attention and shift-wise convolution in cascade within an inverted-residual structure, and a Cross-level Complementary Fusion Block that selectively integrates fine spatial detail with high-level semantics. Together, the two modules improve the classification and localization of small, low-contrast PCB defects.

II. Related Work

A. Vision-Based PCB Inspection as an I&M System

Vision-based measurement uses an imaging sensor and computation to detect, monitor, or quantify a physical phenomenon [1]. In PCB AOI, the phenomena are manufacturing defects, and the reported quantities include their presence, category, confidence, and image-plane location. UniPCB operates in the processing and analysis stage of this measurement chain; its performance is therefore tied to the acquisition conditions, data distribution, and reliability of the inspection decisions [2].

B. PCB Defect Generation

In PCB defect detection, deep learning models require sufficient labeled data, yet real defect samples are scarce, imbalanced, and costly to annotate. Synthetic defect generation has therefore been explored as an effective way to enrich training data and improve detector generalization. Existing methods mainly follow two technical routes: GAN-based synthesis and diffusion-based conditional generation.

Early defect generation methods often adopted GAN-based frameworks to synthesize additional defective samples for data augmentation [17]–[19]. By learning the distribution of defect textures and backgrounds, these methods can increase sample diversity under limited-data settings. Related conditional generation techniques, such as self-attention based generation [20] and spatially-adaptive normalization [21], further show the importance of long-range dependency modeling and spatially-aware modulation for preserving structural consistency. However, GAN-based methods are generally limited

by unstable training, mode collapse, and insufficient control over the spatial relationship between defects and complex PCB backgrounds.

Recently, diffusion models have become a more promising alternative because of their stable training process, strong texture modeling ability, and flexible conditional control [22], [23]. In PCB-related defect generation, Deng et al. [24] introduced feature modulation and normalized flow modules to coordinate local and non-local information for learning high-quality defect distributions. Beyond PCB-specific settings, controllable diffusion methods such as T2I-Adapter [25] and Uni-ControlNet [26] demonstrate that external structural conditions can effectively guide pretrained diffusion models and improve controllability. These studies suggest that diffusion-based generation is well suited for synthesizing fine-grained industrial defect images.

Despite these advances, existing defect generation methods still have two limitations. First, most methods rely on limited or single-form conditions, which are insufficient to describe the dense circuit structures and subtle defect patterns in PCB images. Second, conditional signals are often injected in a coarse or single-scale manner, making it difficult to preserve fine traces and local defect geometry across the denoising process. To address these issues, we introduce a multi-modal condition generator together with multi-scale condition embedding and modulation, enabling more structured and defect-aware PCB sample synthesis for downstream detection.

C. PCB Defect Detection

PCB defect detection differs from natural image detection because PCB defects are usually small, sparse, irregularly shaped, and visually similar to normal circuit patterns. Traditional handcrafted or template-matching methods are sensitive to manufacturing variations and therefore generalize poorly across board layouts and imaging conditions. Deep learning based detectors have improved robustness by learning discriminative representations automatically, and recent PCB defect detection studies [27]–[31] mainly improve performance from two perspectives: adaptive feature modeling and multi-scale feature fusion. Beyond PCB-specific scenarios, recent TCSVT studies on industrial defect and anomaly detection have also explored knowledge distillation for rail surface defects [32], differential distillation for unsupervised anomaly localization [33], multi-perspective multimodal fusion for 3D industrial defects [34], and memory-augmented wavelet representations for anomaly detection [35]. These works further indicate that efficient representation learning and robust localization are central issues in industrial visual inspection.

Adaptive feature modeling [36], [37] aims to enhance defect-relevant responses while suppressing complex circuit background interference. Liu et al. [36] introduced a multi-head nonlocal transformer module to capture long-range dependencies and focus on defect regions. Guo et al. [37] proposed an Image-to-Instance Contextual Enhancement Component that incorporates global contextual information to reweight proposal features. Dai et al. [38] designed Dynamic Head, which dynamically modulates detection features with scale-aware, spatial-aware, and task-aware attention. These methods demonstrate the importance of adaptive representation learning for subtle defect perception.

Another line of work focuses on multi-scale feature fusion [39]–[41], which is critical for localizing tiny defects while preserving semantic discrimination. Yang et al. [39] designed a Coordinate Feature Refinement module to fuse channel- and spatial-level features and improve scale robustness. Cao et al. [40] proposed MRC-DETR, which embeds multi-scale residual coupling into a DETR-based framework for bare-board PCB defect detection. Ji et al. [41] introduced a Slim-Scale Adaptive Fusion structure that aggregates hierarchical features using grouped spatial convolutional pyramids and channel shuffle, improving fusion efficiency and detection accuracy.

Although these PCB-specific and general industrial inspection methods improve feature representation, localization, and deployment efficiency, they mainly operate at the detector or anomaly modeling level and typically assume that the training distribution is sufficiently representative. In practical PCB inspection, however, defect samples remain scarce and highly imbalanced, especially for rare defect categories. As a result, architecture-level improvements alone may still be limited by insufficient training data. This motivates our generation-assisted detection pipeline, which combines structured defect synthesis with task-specific feature modeling to address both data scarcity and representation limitations.

III. METHOD

A. Overview architecture

1) ***Generation Framework****:* As illustrated in Fig. 2, the generation network is built upon Stable Diffusion [8] with ControlNet [42] and extended with three dedicated modules for non-natural industrial images. A real PCB image first passes through the Multi-modal Condition Generator (MmCG), which produces complementary edge, relative pseudo-depth, and text conditions in parallel. These conditions are then embedded into the diffusion U-Net at four resolutions by the ScaleEncoder, and at each resolution a CondMod block jointly modulates the noise feature with the condition and textual embedding to steer the denoising process. The synthesized defect images are merged with real data to form a generation-augmented training set with improved defect coverage for downstream detection. The three modules are detailed in Sec. III-B.

2) ***Detection Framework****:* As shown in Fig. 4, the detection network extends RT-DETR [12] with a ResNet-18 backbone and two task-specific modules. IRSA Blocks replace the original BasicBlocks at every backbone stage to produce more discriminative features, and three CLCF blocks (CLCF_L, CLCF_M, CLCF_S) subsequently fuse the backbone features with the CCFM outputs at pyramid levels S3, S4, and S5. The fused features are concatenated and fed to the IoU-aware query selection and decoder. The two modules are detailed in Sec. III-C.

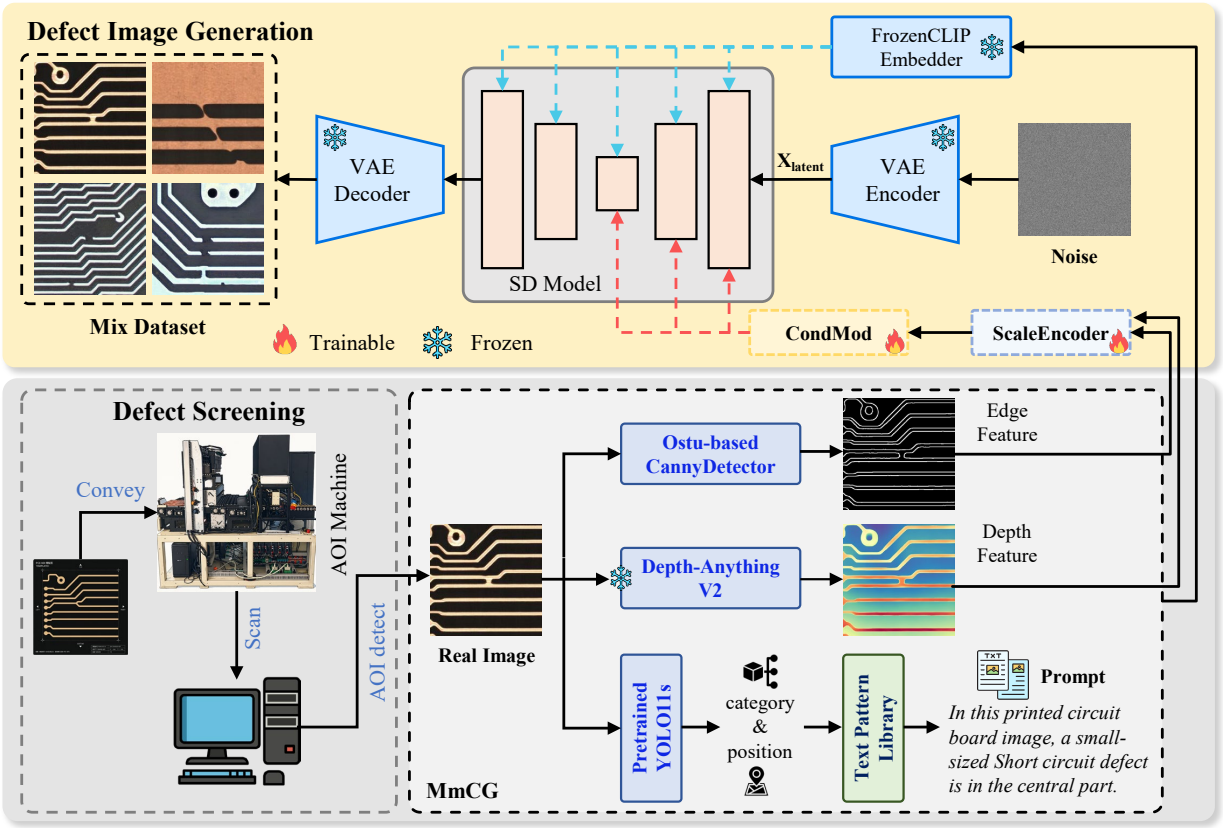


Fig. 2. Overview of the proposed defect generation framework. DsPCBSD+ images are processed by the Multi-modal Condition Generator (MmCG) to extract edge, relative pseudo-depth, and text conditions. ScaleEncoder and CondMod embed these conditions into Stable Diffusion, and the synthesized samples augment the detector training data. The AOI block is a workflow schematic, not a private data source.

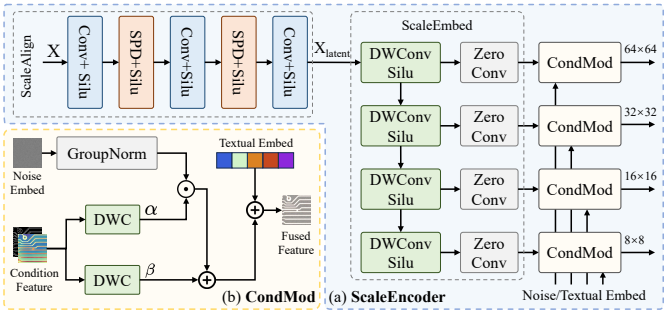


Fig. 3. (a) Structure of the ScaleEncoder, which encodes the condition map via ScaleAlign layer and injects multi-resolution features into the diffusion U-Net via ScaleEmbed layer. (b) Structure of the Condition Modulation (CondMod), which applies FiLM-style spatially-adaptive modulation to fuse condition features and textual embeddings with the noise map at each resolution.

B. Generation Framework

*1) **Latent Diffusion Model**:* We build our generator on the Latent Diffusion Model (LDM) [8], which performs denoising in the latent space of a pretrained autoencoder rather than in pixel space, substantially reducing computational cost. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, the encoder $\mathcal{E}$ compresses it into a latent code $z = \mathcal{E}(x)$, on which forward diffusion and reverse denoising are performed following the standard DDPM formulation [6]:

$$q(z_t \mid z_0) = \mathcal{N}\big(z_t;\, \sqrt{\bar{\alpha}_t}\, z_0,\, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \quad \bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i), \tag{1}$$

where $\{\beta_t\}_{t=1}^{T}$ is a fixed variance schedule. At inference, the denoised latent $\tilde{z}$ is decoded back to image space as $\tilde{x} = \mathcal{D}(\tilde{z})$. We adopt DDIM sampling [7] for efficient skip-step generation. The condition injection and modulation mechanisms specific to our method are detailed in the following subsections.

*2) **Multi-modal Condition Generator**:* Diffusion models require paired (image, condition) data, yet no public image–text PCB dataset exists. To automate dataset construction and provide rich control signals, we design a Multi-modal Condition Generator (MmCG) with three parallel branches, illustrated in Fig. 2.

Edge branch. PCB traces and substrates differ sharply in color and brightness, making edge maps a natural structural prior. We employ the Canny operator [43] with high and low thresholds adaptively derived from the Otsu threshold [44] (scaled by fixed upper and lower factors), yielding image-specific edges without manual tuning.

Relative pseudo-depth branch. Edge maps alone cannot distinguish foreground traces from background substrate, which may cause spatial misalignment. We therefore estimate a depth map using the pretrained Depth-Anything V2 [45]. Because monocular depth estimation does not measure the physical height of a planar PCB, we treat its output as a relative pseudo-depth prior that supplies complementary scene-layout cues.

Text branch. Since no public image–text PCB dataset exists, we construct structured prompts automatically from detection outputs. A pretrained YOLO11 [46] first predicts defect cat-

egories and bounding boxes for images in the DsPCBSD+ generation subset; no separately collected private images are used in the reported experiments. Each defect is then characterized along three axes: *scale*, classified as small, medium, or large via a dual-threshold rule applied to the bounding-box area; *location*, encoded by mapping the normalized box-center coordinates onto a 3×3 spatial grid (top, bottom, left, right, center, and four corners); and *distribution*, described at the instance level when defect count is low (per-defect category, scale, and position) or abstracted to region-level patterns (e.g., scattered or locally clustered) when count is high. The resulting descriptions are slotted into a template library parameterized by {category, scale, location, quantity} to produce structured prompts, which are subsequently encoded by the frozen CLIP text encoder.

Together, the three branches provide structural, spatial, and semantic conditions that jointly guide the downstream diffusion process.

3) ***ScaleEncoder****:* ControlNet [42] injects condition features only at the initial resolution, which is inadequate for PCB images with dense, fine-grained circuit patterns. Following the multi-scale injection principle of Uni-ControlNet [47], we propose a ScaleEncoder that embeds conditions into the diffusion U-Net at four resolutions (64×64, 32×32, 16×16, 8×8).

As shown in Fig. 3(a), ScaleEncoder consists of two components. *ScaleAlign* encodes the raw condition map through three SiLU-activated standard convolutions interleaved with two SPD-Conv downsampling layers [48]. SPD-Conv reorganizes spatial information into the channel dimension, preserving local structural cues that standard strided convolutions tend to lose, a property well suited to thin PCB traces. *ScaleEmbed* then taps intermediate features at four resolutions and refines each via a DWConv–SiLU block followed by a zero convolution [42], whose zero initialization allows the pretrained diffusion backbone to remain undisturbed at the start of training and gradually learn the optimal embedding. The four refined features are fed to the corresponding Condition Modulation.

4) ***Condition Modulation****:* To strengthen condition guidance at each resolution, we introduce a Condition Modulation (CondMod), shown in Fig. 3(b). At every scale, parallel CondMod fuse the noise feature, the condition feature, and the textual embedding.

Following the FiLM-style modulation [49], [50], the noise feature is first normalized via GroupNorm to obtain $\mathcal{N}_{norm}$, and the condition feature is passed through two parallel depth-wise convolutions to predict the scale and shift parameters α and β:

$$\mathbf{X}_{\text{mid}} = \mathcal{N}_{norm} \odot (1 + \alpha) + \beta, \tag{2}$$

where $\odot$ denotes element-wise multiplication. The $(1 + \alpha)$ formulation initializes the block close to an identity mapping, stabilizing training. The modulated feature $\mathbf{X}_{\text{mid}}$ is then added to the textual embedding to yield the block output, achieving spatially-aware conditioning driven by the visual control signal and semantically-aware conditioning driven by the prompt.

C. Detection Framework

1) ***Inverted Residual Shift Attention****:* PCB defects are typically small, low-contrast, and easily masked by complex circuit patterns. Self-attention captures global semantics but is insensitive to fine textures, while convolutions capture local details but lack long-range context; neither alone is sufficient for reliable small-defect detection.

We therefore propose the Inverted Residual Shift Attention (IRSA) block (Fig. 5(a)), which couples self-attention and shift-wise convolution in cascaded within an inverted-residual structure. Given input $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we first apply a convolution–batch normalization–ReLU (CBR) layer to obtain a preliminary feature $\mathbf{X}_{\text{pre}} = \text{CBR}(\mathbf{X})$, and then expand the channel dimension from C to C' via a 1×1 convolution, yielding $\mathbf{X}_{\text{expand}}$. Two cascaded modules are then employed:

$$\mathbf{X}_{att} = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3}$$

$$\mathbf{X}_{swc} = DWConv(Shift_G(\mathbf{X}_{att}))\,, \tag{4}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are linear projections of $\mathbf{X}_{pre}$, and $\mathbf{V}$ is a linear projections of $\mathbf{X}_{expand}$. $Shift_G(\cdot)$ splits channels into G groups and spatially shifts each group along one of G pre-defined directions (up, down, left, right, and four diagonals) [51], followed by a depth-wise convolution that aggregates the shifted context. The branch outputs are merged and projected back to C channels:

$$\mathbf{X}_{out} = Conv_{1\times1}(\mathbf{X}_{att} + \mathbf{X}_{swc}). \tag{5}$$

The final output restores the input and the preliminary feature paths to ensure stable gradient flow:

$$\mathbf{X}_{final} = \mathbf{X} + \mathbf{X}_{pre} + \mathbf{X}_{out}. \tag{6}$$

The cascaded coupling of global attention and shift-wise convolution, together with the multi-level residuals, yields efficient yet discriminative features for small PCB defects.

2) ***Cross-level Complementary Fusion****:* Shallow features retain fine textures crucial for localizing tiny defects but are noise-sensitive, whereas deep features carry strong semantics but lose spatial detail. Naive addition or concatenation tends to amplify noise or suppress subtle defects. To better fuse the two, we propose the Cross-level Complementary Fusion (CLCF), comprising a Dual-path Collaborative Attention (DPCA) that predicts a pixel-wise gate, and a gated weighting scheme with residual compensation (Fig. 6(a)).

DPCA sub-block. Given the concatenated input $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, a 1×1 convolution produces the shared tensor $\mathbf{X}_{qkv}$. Two parallel branches then operate on it. The *local* branch captures fine-grained cues:

$$\mathbf{X}_{\text{loc}} = \text{GConv}(\text{CS}(\text{DWConv}_{3\times3}(\mathbf{X}_{qkv})))\,, \tag{7}$$

where $\text{CS}(\cdot)$ denotes channel shuffle. The *global* branch computes self-attention with a learnable per-head scaling vector $\boldsymbol{\alpha}$, following [52]:

$$\mathbf{X}_{\text{glo}} = \text{Conv}_{1\times1}\left(\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\boldsymbol{\alpha}}\right)\mathbf{V}\right) + \mathbf{X}_{qkv}. \tag{8}$$

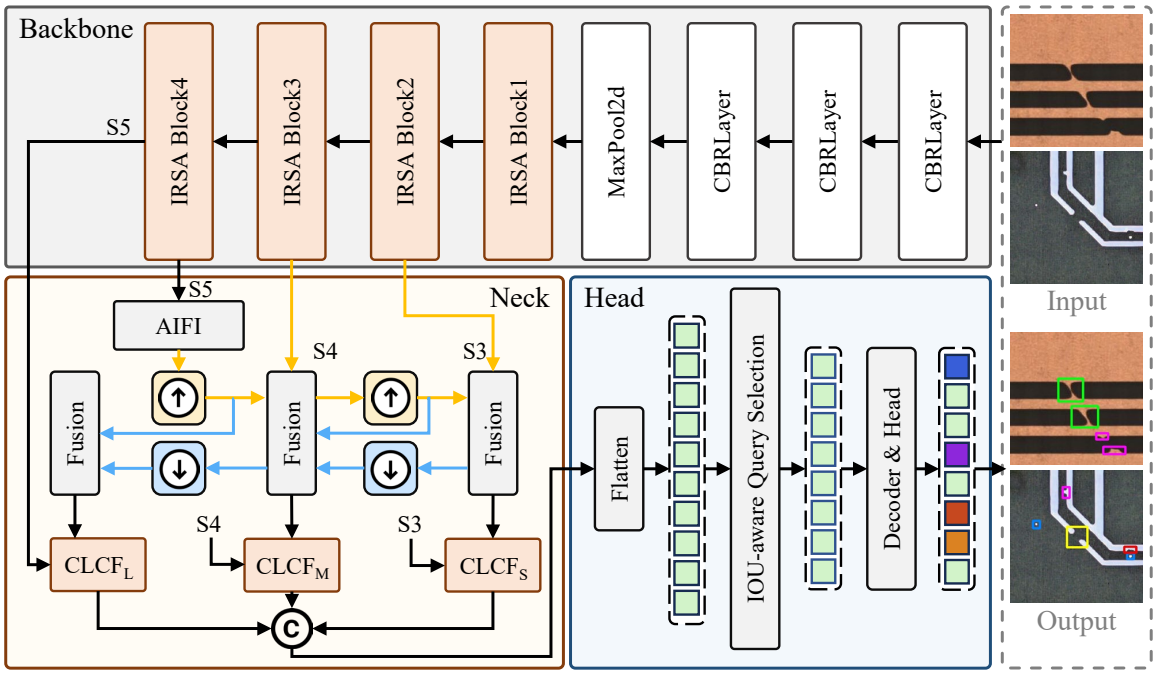


Fig. 4. Overview of the proposed detection framework. IRSA Blocks replace BasicBlocks in all four stages of the ResNet-18 backbone. Yellow and blue blocks in the Neck denote upsampling and downsampling operations, respectively.

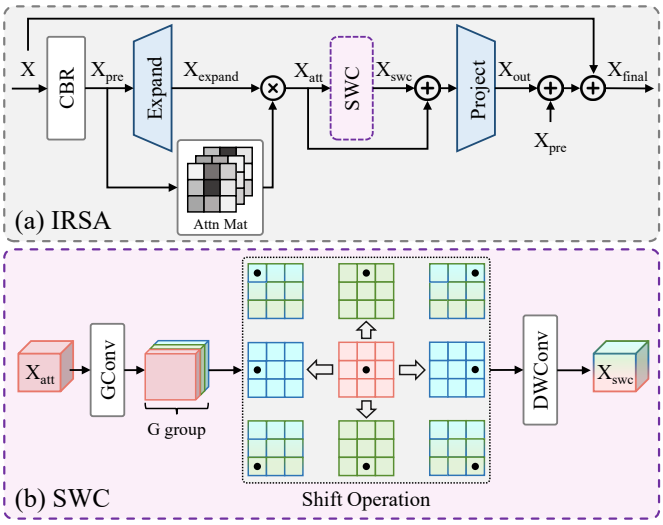


Fig. 5. (a) Overview of the Inverted Residual Shift Attention Block. (b) Structure of the shift-wise convolution (SWC).

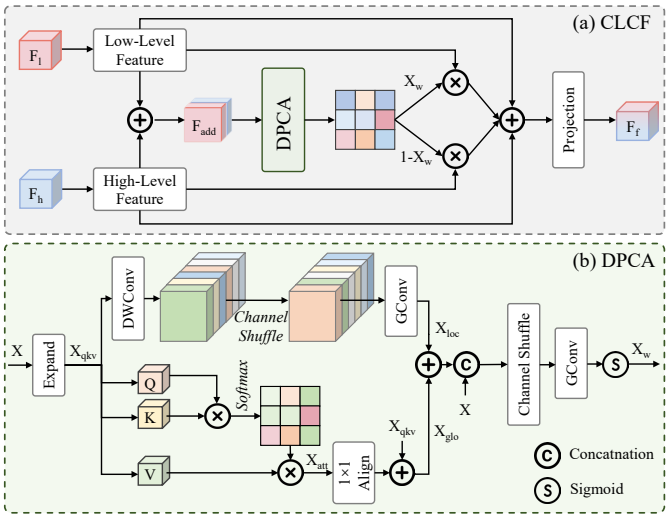


Fig. 6. (a) Overview of the Cross-level Complementary Fusion (CLCF) Block. (b) Structure of the Dual-path Collaborative Attention (DPCA) Block.

The two outputs are concatenated with the input and passed through a grouped convolution with *sigmoid* activation to yield a pixel-wise gate:

$$\mathbf{w} = \sigma(\mathrm{GConv}_{7\times7}(\mathrm{CS}(\mathrm{Concat}[\mathbf{X}_{\mathrm{loc}}, \mathbf{X}_{\mathrm{glo}}, \mathbf{X}]))), \quad (9)$$

Gated fusion. Let $\mathbf{F}_l$ and $\mathbf{F}_h$ denote the low- and high-level inputs aligned to the same resolution. CLCF fuses them via a gate-and-residual scheme:

$$\mathbf{F}_f = \mathrm{Proj}(\mathbf{w} \odot \mathbf{F}_l + (1 - \mathbf{w}) \odot \mathbf{F}_h + (\mathbf{F}_l + \mathbf{F}_h)), \quad (10)$$

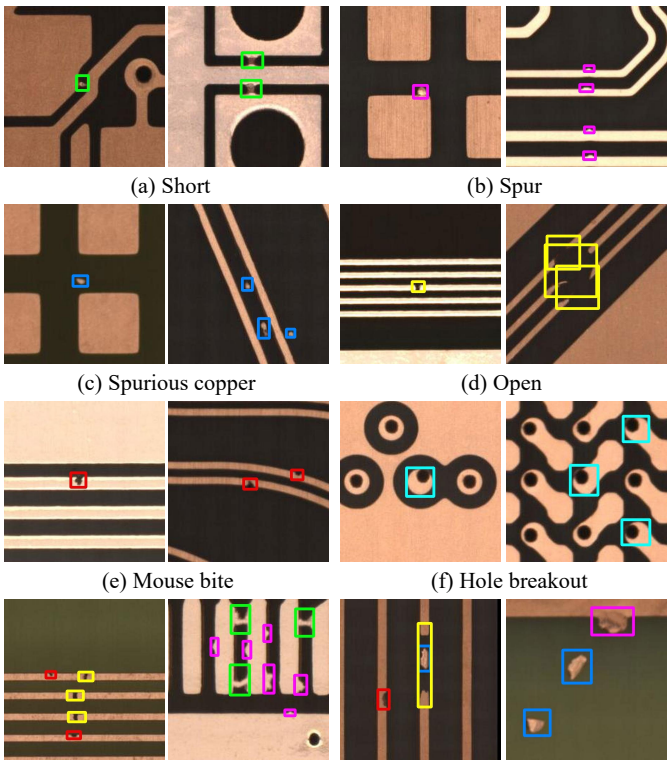
where the residual term $\mathbf{F}_l + \mathbf{F}_h$ prevents information loss from over-gating. The gate $\mathbf{w}$ adaptively selects detail cues from $\mathbf{F}_l$ where defects are subtle and semantic cues from $\mathbf{F}_h$ elsewhere, yielding features well suited for the detection head.

IV. Experiments

A. Datasets

All reported generation and detection experiments use only the public PCB defect dataset DsPCBSD+ [53]; no private or newly collected PCB images are included. Six defect categories that significantly affect PCB quality are selected: short, spur, spurious copper, open, mouse bite, and hole breakout [54], [55]. Annotation errors in the original dataset are corrected before constructing the generation and detection subsets, and representative corrected samples are shown in Fig. 7.

Generation dataset. A subset of 5,252 defect images is reconstructed from DsPCBSD+ to train the diffusion model,



(a) Short (b) Spur (c) Spurious copper (d) Open (e) Mouse bite (f) Hole breakout (g) Multiple classes

Fig. 7. Representative annotated samples from the six defect categories, including single-class images (a–f) and a multi-class example (g). (a) Short; (b) Spur; (c) Spurious copper; (d) Open; (e) Mouse bite; (f) Hole breakout; (g) Multiple classes.

as detailed in Tab. I. The training and validation sets are split at a ratio of 8:2.

Detection dataset. The remaining images are reorganized into a base detection set of 2,027 images (Tab. I, column "Base"), with all defects manually annotated using rectangular bounding boxes and split 8:2 into training and validation sets. To investigate the effect of different data expansion strategies, two extended variants are constructed:

- **Extend I** (3,319 images): the base set expanded via conventional augmentations, including random flipping, rotation, and Gaussian blur.
- **Extend II** (3,382 images): the base set expanded by incorporating defect images synthesized by the trained diffusion model. Annotations are produced semi-automatically: a seed set of real images is manually labeled, an auxiliary annotation model is trained on this seed set to label the synthetic images, and a final manual verification step corrects residual errors.

The two extended datasets are used in Sec. IV-C to provide a controlled comparison between traditional and generation-based augmentation strategies.

B. Experimental Setup

1) ***Implementation Details****:* The experimental environment in this study runs on an Ubuntu system, and the experimental setting employs an Intel Xeon E5-2686 v4 CPU and four 24G NVIDIA TITAN RTX GPUs as the hardware configuration. The generation model is based on Stable Diffusion v1.5 and uses a batch size of 1, 30,000 training steps, and a learning rate of 0.00002, while the detection model adopts a batch size of 4, is trained for 100 epochs, and uses a learning rate of 0.0001. For both the generation and detection tasks, the input image resolution is set to 512×512. All networks are optimized using the AdamW optimizer, where the momentum and weight decay are set to 0.9 and 0.0001, respectively.

2) ***Evaluation Metrics****:* The quality of synthesized PCB defect images is evaluated from three complementary perspectives. Fréchet Inception Distance (FID) [56] measures the distributional similarity between real and generated images in the feature space of a pretrained Inception network, where lower values indicate closer distributions:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right), \quad (11)$$

where μ_r, Σ_r and μ_g, Σ_g are the mean and covariance of the real and generated feature distributions, respectively. Learned Perceptual Image Patch Similarity (LPIPS) [57] quantifies perceptual similarity by comparing deep feature activations across L layers weighted by w_l, with lower values indicating greater perceptual fidelity:

$$\mathrm{LPIPS} = \sum_l w_l \cdot \|\phi_l(x) - \phi_l(y)\|_2^2. \quad (12)$$

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [58] assess pixel-level reconstruction quality and structural consistency, respectively, with higher values indicating better fidelity:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \quad (13)$$

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (14)$$

where μ_x, μ_y are local means, σ_x^2, σ_y^2 are variances, σ_{xy} is the cross-covariance, and C_1, C_2 are stabilizing constants.

Detection metrics. The downstream defect detection performance is evaluated using Precision, Recall, and mean Average Precision at two IoU thresholds (mAP@0.5 and mAP@0.5:0.95). Precision measures the fraction of correct detections among all detections, Recall measures the fraction of ground-truth defects successfully detected, and mAP summarizes the area under the precision–recall curve averaged over N defect categories:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad (15)$$

$$\mathrm{mAP} = \frac{1}{N}\sum_{c=1}^{N} \int_0^1 P(R)\,\mathrm{d}R. \quad (16)$$

C. Comparative Experiment

1) ***Detection Performance Comparison****:* To comprehensively evaluate UniPCB, we compare it against ten representative detectors spanning two architectural families. The Transformer-based group includes DETR [11], Deformable-DETR [59], DAB-DETR [60], DINO [61], RT-DETR [12], D-FINE [62], and DEIM [63]; the CNN-based group includes

TABLE I
DATASET STATISTICS FOR DEFECT GENERATION AND DETECTION EXPERIMENTS. EXTEND **I**: TRADITIONAL AUGMENTATION; EXTEND **II**: GENERATION-BASED AUGMENTATION.

Type	Class	Defect Gen.	Defect Det.		
			Base	Extend **I**	Extend **II**
No. of Images	Short	180	137	217	579
	Spur	1,326	444	705	597
	Spurious copper	826	360	597	613
	Open	770	346	564	768
	Mouse bite	1,084	422	705	592
	Hole breakout	1,066	433	715	540
	Total	**5,252**	**2,027**	**3,319**	**3,382**
No. of Defects	Short	228	161	253	839
	Spur	2,213	546	874	747
	Spurious copper	969	385	641	667
	Open	919	362	594	936
	Mouse bite	1,309	440	732	636
	Hole breakout	2,342	480	796	615
	Total	**7,980**	**2,374**	**3,890**	**4,440**

YOLOv8 [64], YOLOv10 [65], and YOLO11 [46]. RT-DETR serves as the direct baseline, as UniPCB is built upon it.

Quantitative results are reported in Tab. II, and visual comparisons are shown in Fig. 8. UniPCB achieves Precision / Recall / mAP@0.5 / mAP@0.5:0.95 of 0.977 / 0.953 / 0.980 / 0.618, outperforming all compared methods. Relative to the RT-DETR baseline, mAP@0.5 improves by 2.1 percentage points and mAP@0.5:0.95 by 1.7 percentage points, demonstrating that the IRSA and CLCF modules provide consistent gains beyond the base architecture. Among Transformer-based methods, the strongest competitor is DEIM (mAP@0.5: 0.958), which UniPCB surpasses by 2.2 percentage points. Among CNN-based methods, YOLOv8 achieves the best mAP@0.5:0.95 of 0.611, still 0.7 percentage points below UniPCB. The visual results in Fig. 8 further illustrate that UniPCB produces fewer missed detections and tighter bounding boxes on small, low-contrast defects, where most competing methods show visible localization errors or omissions. Because all detectors in this comparison use the same Extend II data, the results isolate the contribution of the detection architecture from differences in training data. Together with the module ablation, these results show that IRSA and CLCF provide architecture-side gains under generation-augmented training.

*2) **Generation Performance Comparison**:* We compare UniPCB against three representative conditional generation baselines: ControlNet [42], which injects a single condition at the initial resolution; Uni-ControlNet [47], which combines edge and depth conditions through a unified control mechanism but still injects them at a single scale; and AnyControl [66], which supports flexible multi-condition combinations but without explicit cross-modal fusion. To further isolate the effect of condition type, we also evaluate a variant UniPCB (+seg) that replaces the text condition with a defect segmentation mask as the third control signal. Results are reported in Tab. III and visualized in Fig. 9.

The results reveal a clear progression. ControlNet, relying solely on edge maps, achieves PSNR/SSIM of 10.23/0.502 but yields a relatively high FID of 131.11, as single-modality conditions cannot fully constrain the structural diversity of PCB layouts. Uni-ControlNet improves PSNR to 10.60 by adding depth as a complementary condition, yet its FID rises to 145.53 and SSIM drops to 0.457, suggesting that naïve concatenation of conditions without dedicated fusion introduces inter-modal interference. AnyControl shows similar limitations despite its flexible condition combinations, achieving the weakest overall performance (FID: 142.01, SSIM: 0.404). UniPCB (+seg) improves PSNR and SSIM over the baselines by adding a segmentation mask, but its FID (142.93) and LPIPS (0.535) remain worse than UniPCB, indicating that a pixel-level mask introduces spatial redundancy with the edge condition rather than complementary information.

UniPCB achieves the best performance across all metrics (FID: 129.61, LPIPS: 0.457, PSNR: 13.45, SSIM: 0.619). The key distinction from prior methods is not the number of conditions but their complementarity and fusion quality: edge, depth, and text provide structural, spatial, and semantic cues that cover orthogonal axes of control, while ScaleEncoder and CondMod ensure that these conditions are effectively embedded across multiple scales rather than injected at a single resolution. The qualitative results in Fig. 9 further confirm that UniPCB preserves the global circuit topology while rendering localized defect patterns with higher fidelity than all compared methods.

*3) **Data Augmentation Comparison**:* To evaluate the downstream benefit of generation-based augmentation, all detection models in Tab. II are retrained on Extend I and Extend II respectively, and their performance is systematically compared in Tab. IV.

Models trained on Extend II consistently outperform those trained on Extend I across almost all metrics. The gains are most pronounced in Recall and mAP@0.5:0.95, where most models improve by 1.5%–2.0%, reflecting that generation-augmented data helps reduce missed detections and improves localization quality at stricter IoU thresholds. mAP@0.5 also shows consistent improvement close to 1%. Precision remains nearly unchanged or shows slight declines for some models; the minor degradation is likely attributable to additional false positives caused by enhanced sensitivity to defect features introduced by the richer training distribution.

These results confirm that the generation-based augmentation provides richer and more informative training samples than conventional augmentation, enabling detection models to better capture discriminative features across all six defect categories. Notably, UniPCB itself benefits from Extend II with improvements of +0.7% / +2.6% / +0.9% / +2.1% in Precision / Recall / mAP@0.5 / mAP@0.5:0.95, demonstrating that the generation and detection components of UniPCB are mutually reinforcing.

D. Ablation Study

*1) **Module contribution**:* Tab. V reports the contribution of each proposed module. Adding IRSA alone (setting (a)) yields gains of +0.9% / +1.3% in mAP@0.5 / mAP@0.5:0.95 over

TABLE II
DETECTION PERFORMANCE OF DIFFERENT MODELS. THE TOP THREE RESULTS ARE MARKED WITH RED, BLUE, AND GREEN, RESPECTIVELY.

Model	Source	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑	FPS↑	FLOPS(G)↓	Param.(M)↓
DETR [11]	20'ECCV	0.896	0.875	0.923	0.582	66.6	60.5	41.5
Deformable-DETR [59]	21'ICLR	0.915	0.905	0.940	0.592	32.5	125.9	40.1
DAB-DETR [60]	22'ICLR	0.872	0.889	0.905	0.570	41.0	65.3	43.7
DINO [61]	23'ICLR	0.922	0.913	0.948	0.598	23.5	183.3	47.5
YOLOv8 [64]	23'Ultralytics	0.954	0.941	0.965	0.611	192.8	23.4	9.83
YOLOv10 [65]	24'NIPS	0.850	0.855	0.911	0.583	168.5	24.5	8.04
RT-DETR [12]	24'CVPR	0.949	0.921	0.959	0.601	97.0	57.2	19.9
YOLO11 [46]	24'Ultralytics	0.956	0.921	0.960	0.607	192.3	21.3	9.42
D-FINE [62]	25'ICLR	0.938	0.927	0.952	0.606	150.6	56.3	19.1
DEIM [63]	25'CVPR	0.942	0.938	0.958	0.610	118.2	92.5	31.3
UniPCB (Detection)	-	0.977	0.953	0.980	0.618	90.9	71.8	19.1

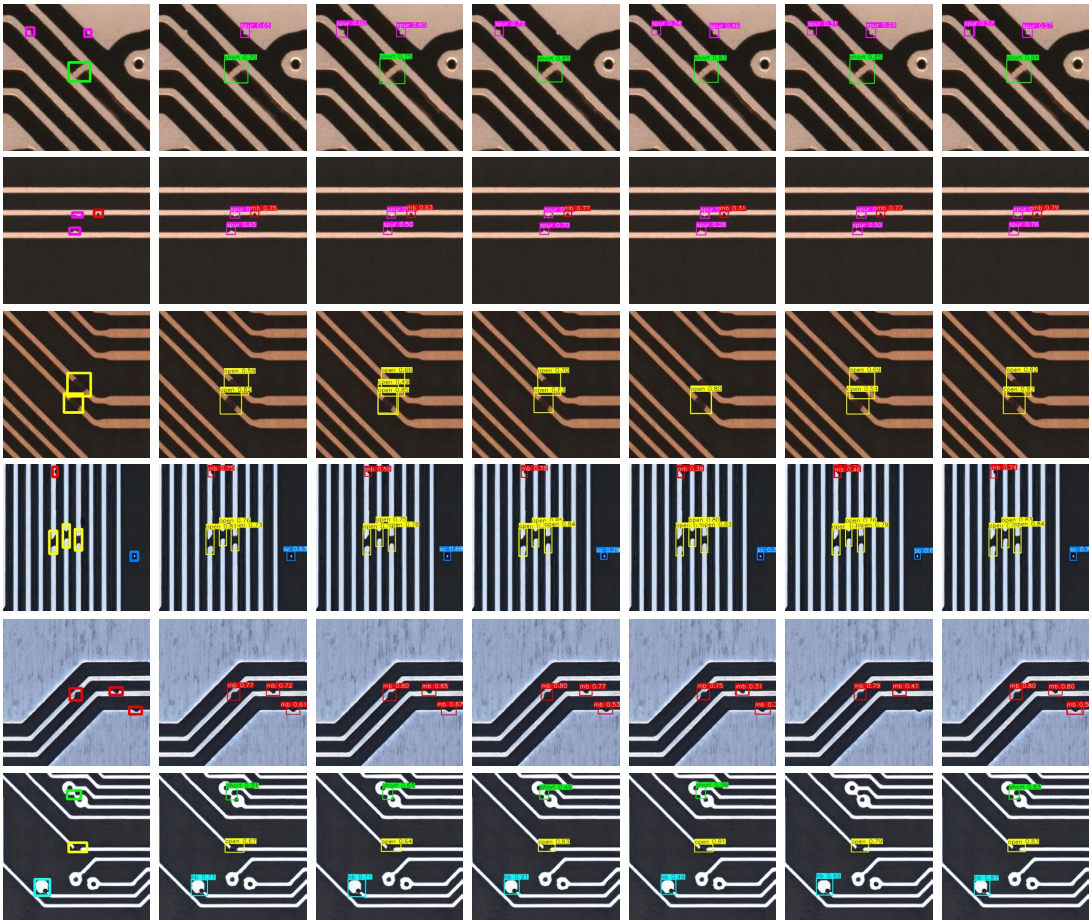


Fig. 8. Visualization of detection results across different models. Zoom-in for best view.

TABLE III
GENERATION PERFORMANCE OF DIFFERENT MODELS.

Model	FID↓	LPIPS↓	PSNR↑	SSIM↑
ControlNet [42]	131.11	0.566	10.23	0.502
Uni-Controlnet [47]	145.53	0.562	10.60	0.457
AnyControl [66]	142.01	0.635	9.02	0.404
UniPCB (+seg)	142.93	0.535	12.47	0.525
UniPCB (ours)	**129.61**	**0.457**	**13.45**	**0.619**

the baseline, with a notable Recall improvement of +1.2%. This suggests that coupling self-attention with shift-wise convolution reduces missed detections by capturing fine-grained local cues that vanilla self-attention overlooks, particularly for small, low-contrast defects embedded in complex circuit patterns. Adding CLCF alone (setting (b)) produces larger Precision gains (+2.0%) alongside comparable mAP improvements, indicating that the pixel-level gating mechanism effectively suppresses false detections caused by background interference during feature fusion. When both modules are combined (setting (c)), the model achieves the best performance across all metrics, with mAP@0.5 and mAP@0.5:0.95 reaching 0.980 and 0.618, improving the baseline by +2.1% and +1.7% respectively. The complementary nature of the two modules is evident: IRSA strengthens single-scale feature discrimination

TABLE IV
EVALUATION OF AUGMENTATION STRATEGIES. SUPERSCRIPTS I AND II DENOTE TRADITIONAL AUGMENTATION AND GENERATIVE AUGMENTATION, RESPECTIVELY. FOR RESULTS OBTAINED WITH GENERATIVE AUGMENTATION, THE PERCENTAGES IN PARENTHESES INDICATE RELATIVE CHANGES COMPARED WITH THE CORRESPONDING TRADITIONAL AUGMENTATION SETTING. RED INDICATES FAVORABLE CHANGES, WHILE GREEN INDICATES UNFAVORABLE CHANGES.

Model	Source	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑
DETR^{I} [11]	20'ECCV	0.892	0.848	0.914	0.556
DETR^{II} [11]		0.896 ↑ 0.4%	0.875 ↑ 2.7%	0.923 ↑ 0.9%	0.582 ↑ 2.6%
Deformable-DETR^{I} [59]	21'ICLR	0.917	0.882	0.932	0.561
Deformable-DETR^{II} [59]		0.915 ↓ 0.2%	0.905 ↑ 2.3%	0.940 ↑ 0.8%	0.592 ↑ 3.1%
DAB-DETR^{I} [60]	22'ICLR	0.869	0.867	0.899	0.545
DAB-DETR^{II} [60]		0.872 ↑ 0.3%	0.889 ↑ 2.2%	0.905 ↑ 0.6%	0.570 ↑ 2.5%
DINO^{I} [61]	23'ICLR	0.928	0.888	0.940	0.574
DINO^{II} [61]		0.922 ↓ 0.6%	0.913 ↑ 2.5%	0.948 ↑ 0.8%	0.598 ↑ 2.4%
YOLOv8^{I} [64]	23'Ultralytics	0.945	0.933	0.956	0.601
YOLOv8^{II} [64]		0.954 ↑ 0.9%	0.941 ↑ 0.8%	0.965 ↑ 0.9%	0.611 ↑ 1.0%
YOLOv10^{I} [65]	24'NIPS	0.853	0.840	0.901	0.564
YOLOv10^{II} [65]		0.850 ↓ 0.3%	0.855 ↑ 1.5%	0.911 ↑ 1.0%	0.583 ↑ 1.9%
RT-DETR^{I} [12]	24'CVPR	0.951	0.895	0.950	0.573
RT-DETR^{II} [12]		0.949 ↓ 0.3%	0.921 ↑ 2.6%	0.959 ↑ 0.9%	0.601 ↑ 2.8%
YOLO11^{I} [46]	24'Ultralytics	0.927	0.899	0.952	0.574
YOLO11^{II} [46]		0.956 ↑ 2.9%	0.921 ↑ 2.2%	0.960 ↑ 0.8%	0.607 ↑ 3.3%
D-FINE^{I} [62]	25'ICLR	0.935	0.910	0.944	0.579
D-FINE^{II} [62]		0.938 ↑ 0.3%	0.927 ↑ 1.7%	0.952 ↑ 0.8%	0.606 ↑ 2.7%
DEIM^{I} [63]	25'CVPR	0.940	0.914	0.948	0.584
DEIM^{II} [63]		0.942 ↑ 0.3%	0.938 ↑ 2.4%	0.958 ↑ 1.0%	0.610 ↑ 2.6%
UniPCB^{I}	-	0.970	0.927	0.971	0.593
UniPCB^{II}		0.977 ↑ 0.7%	0.953 ↑ 2.6%	0.980 ↑ 0.9%	0.618 ↑ 2.1%

TABLE V
ABLATION STUDY ON THE EFFECTIVENESS OF IRSA AND CLCF MODULES.

Index	IRSA	CLCF	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑
Baseline	✗	✗	0.949	0.921	0.959	0.601
a	✓	✗	0.951	0.933	0.968	0.614
b	✗	✓	0.969	0.949	0.969	0.614
c	✓	✓	**0.977**	**0.953**	**0.980**	**0.618**

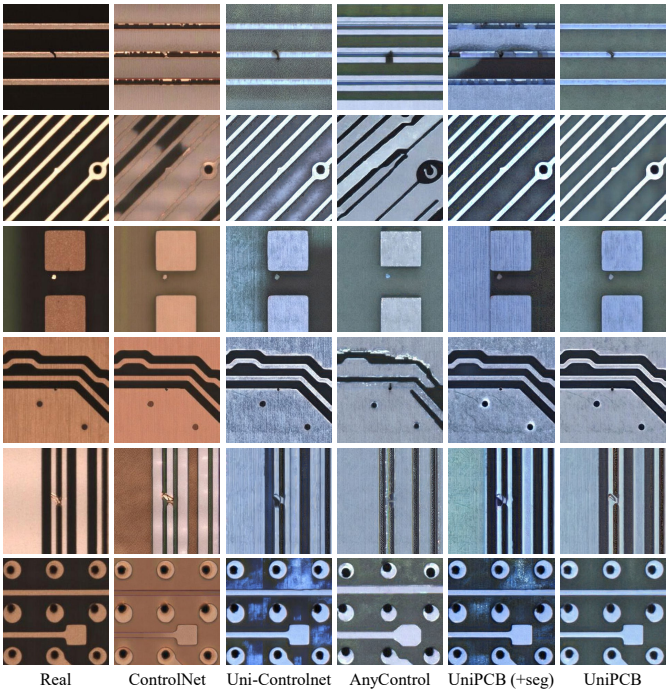


Fig. 9. Visualization of generation results across different models. For better clarity, results are best viewed in enlarged form.

TABLE VI
ABLATION RESULTS OF DIFFERENT ATTENTION MECHANISMS IN CLCF BLOCK.

Method	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑
Baseline	0.949	0.921	0.959	0.601
+SE [67]	0.945	0.932	0.960	0.604
+CA [68]	0.936	0.908	0.945	0.582
+SimAM [69]	0.958	0.918	0.965	0.607
+EMA [70]	0.963	0.913	0.964	0.609
+DPCA (ours)	**0.969**	**0.949**	**0.969**	**0.614**

while CLCF enhances cross-level feature aggregation, and their combination yields gains exceeding either module alone.

*2) **Attention mechanism comparison**:* To validate the design of the DPCA sub-block within CLCF, we replace it with four commonly used attention mechanisms and report results in Tab. VI.

SE increases Recall by 1.1 percentage points through inter-channel reweighting but provides only limited overall gains. CA reduces mAP@0.5:0.95 by 1.9 percentage points relative to the baseline, whereas SimAM and EMA improve that metric by 0.6 and 0.8 percentage points, respectively. DPCA gives the strongest overall result, reaching Precision / Recall /

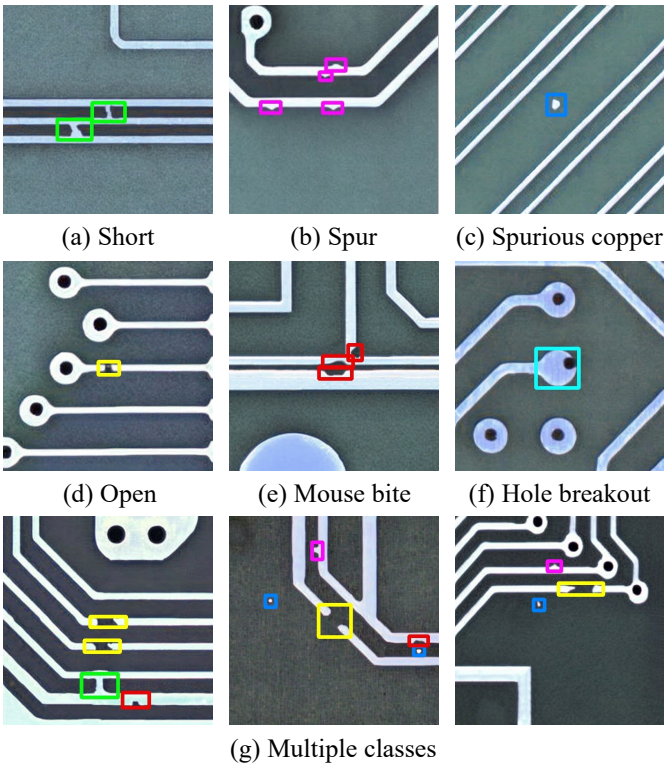


Fig. 10. Examples of synthesized defect images with bounding-box annotations, covering all six categories in Fig. 7. Annotations are generated by an auxiliary model trained on seed annotations and verified manually.

mAP@0.5 / mAP@0.5:0.95 of 0.969 / 0.949 / 0.969 / 0.614. Its grouped-convolution branch preserves fine spatial cues, while its attention branch models longer-range dependencies; their combination produces spatially adaptive weights for defects at different scales and locations.

V. Conclusion and Limitation

This paper presents UniPCB, a generation-assisted vision-based measurement framework for the computational processing stage of PCB AOI. The generation branch derives edge, relative pseudo-depth, and text conditions from public PCB images and injects them into a latent diffusion model through ScaleEncoder and CondMod. The detection branch combines IRSA for local–global feature modeling with CLCF for selective cross-level fusion. Experiments on the corrected DsPCBSD+ annotations show improvements in synthetic-image quality and downstream defect classification and localization over the compared baselines. UniPCB therefore contributes to the analysis stage of an AOI instrument; it does not modify the optical sensor or image-acquisition hardware.

The generator and detector are trained sequentially rather than optimized end to end. Moreover, all reported results are derived from the public DsPCBSD+ dataset and do not constitute validation across private production lines or AOI devices. The relative pseudo-depth condition is a learned image prior rather than a physical height measurement. Future work will evaluate detector-feedback generation and robustness across board layouts, devices, and illumination.

References

[1] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: The rising trend of vision-based measurement," *IEEE Instrum. Meas. Mag.*, vol. 17, no. 3, pp. 41–47, 2014.
[2] M. Khanafer and S. Shirmohammadi, "Applied AI in instrumentation and measurement: The deep learning revolution," *IEEE Instrum. Meas. Mag.*, vol. 23, no. 6, pp. 10–17, 2020.
[3] Y. Zhou, M. Yuan, J. Zhang, G. Ding, and S. Qin, "Review of vision-based defect detection research and its perspectives for printed circuit board," *J. Manuf. Syst.*, vol. 70, pp. 557–578, 2023.
[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.
[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. NeurIPS*, pp. 6840–6851, 2020.
[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*, 2021.
[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, pp. 10 684–10 695.
[9] X. Zhang, J. Huang, and L. Zhang, "Any2rsi: Controllable remote sensing text-to-image generation via any control and enriched description," in *Proc. AAAI*, 2026, pp. 12 852–12 860.
[10] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 10 850–10 869, 2023.
[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.
[12] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proc. CVPR*, 2024, pp. 16 965–16 974.
[13] K. Shen, X. Zhou, and Z. Liu, "Class knowledge-guided lightweight network for salient object detection of strip steel surface defect," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2026.
[14] X. Hou, M. Liu, and S. Du, "Unfolddet: Advancing surface defect detection with dual feature separation and relation reasoning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 4, pp. 5371–5383, 2026.
[15] K. Cheng, H. Cui, H. Abdul Ghafoor, H. Wan, Q. Mao, and Y. Zhan, "Tiny object detection via regional cross self-attention network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 8984–8996, 2024.
[16] D. I. Ural and A. Sezen, "Research on pcb defect detection using artificial intelligence: a systematic mapping study," *Evol. Intell.*, vol. 17, no. 5, pp. 3101–3111, 2024.
[17] Z. Du, L. Gao, and X. Li, "A new contrastive gan with data augmentation for surface defect recognition under limited data," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2022.
[18] Y. Duan, Y. Hong, L. Niu, and L. Zhang, "Few-shot defect image generation via defect-aware feature manipulation," in *Proc. AAAI*, 2023, pp. 571–578.
[19] Y. Liang, S. Feng, Y. Zhang, F. Xue, F. Shen, and J. Guo, "A stable diffusion enhanced yolov5 model for metal stamped part defect detection based on improved network structure," *J. Manuf. Process.*, vol. 111, pp. 21–31, 2024.
[20] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, 2019, pp. 7354–7363.
[21] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. CVPR*, 2019, pp. 2337–2346.
[22] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," in *Proc. ICLR*, 2024.
[23] T. Hu, J. Zhang, R. Yi, Y. Du, X. Chen, L. Liu, Y. Wang, and C. Wang, "Anomalydiffusion: Few-shot anomaly image generation with diffusion model," in *Proc. AAAI*, 2024, pp. 8526–8534.
[24] W. Deng, L. Yan, and C. Wang, "Ddmf: a pcb surface defect detection model based on conditional denoising diffusion and multiscale feature fusion," *J. Supercomput.*, vol. 81, no. 15, pp. 1–33, 2025.
[25] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI*, 2024, pp. 4296–4304.
[26] C. Qin, S. Bai, Y. Shen, L. Chen, B. Ni, J. Liu, Y. Liu, and X. Liu, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," in *Proc. NeurIPS*, 2023.

[27] M. Yuan, Y. Zhou, X. Ren, H. Zhi, J. Zhang, and H. Chen, "Yolo-hmc: An improved method for pcb surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2024.

[28] L. Lei, H.-X. Li, and H.-D. Yang, "Reliable and lightweight adaptive convolution network for pcb surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–8, 2024.

[29] X. Liu, "An adaptive defect-aware attention network for accurate pcb-defect detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2024.

[30] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. CVPR*, 2023, pp. 14 408–14 419.

[31] C. Mo, Z. Hu, J. Wang, and X. Xiao, "Sgt-yolo: A lightweight method for pcb defect detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[32] W. Zhou, J. Hong, W. Yan, and Q. Jiang, "Modal evaluation network via knowledge distillation for no-service rail surface defect detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3930–3942, 2024.

[33] Q. Zhou, S. He, H. Liu, T. Chen, and J. Chen, "Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2176–2189, 2023.

[34] M. Asad, W. Azeem, H. Jiang, H. T. Mustafa, J. Yang, and W. Liu, "2M3DF: Advancing 3D industrial defect detection with multi-perspective multimodal fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 6803–6815, 2025.

[35] K. Wu, L. Zhu, W. Shi, W. Wang, and J. Wu, "Self-attention memory-augmented wavelet-CNN for anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1374–1385, 2023.

[36] T. Liu, G.-Z. Cao, Z. He, and S. Xie, "Refined defect detector with deformable transformer and pyramid feature fusion for pcb detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2023.

[37] F. Guo, Z. Chen, B. Chen, M. Jing, and L. Zuo, "Multi-granularity relation enhancement network for tiny defect detection on printed circuit board," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[38] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proc. CVPR*, 2021, pp. 7373–7382.

[39] J. Yang, Z. Liu, W. Du, and S. Zhang, "A pcb defect detector based on coordinate feature refinement," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[40] J. Cao, H. Wu, X. Zhang, L. Tan, and H. Zhang, "Mrc-detr: An adaptive multi-residual coupled transformer for bare board pcb defect detection," *arXiv preprint arXiv:2507.03386*, 2025.

[41] L. Ji, C. Huang, H. Li, W. Han, and L. Yi, "Ms-detr: a real-time multi-scale detection transformer for pcb defect detection," *Signal Image Video Process.*, vol. 19, no. 3, p. 203, 2025.

[42] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. ICCV*, 2023, pp. 3836–3847.

[43] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 679–698, 2009.

[44] N. Otsu *et al.*, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[45] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Proc. NeurIPS*, 2024, pp. 21 875–21 911.

[46] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[47] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *Proc. NeurIPS*, pp. 11 127–11 150, 2023.

[48] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Proc. ECML-PKDD*, 2022, pp. 443–459.

[49] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[50] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. NeurIPS*, 2021, pp. 8780–8794.

[51] D. Li, L. Li, Z. Chen, and J. Li, "Shiftwiseconv: Small convolutional kernel with large kernel effect," in *Proc. CVPR*, 2025, pp. 25 281–25 291.

[52] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek, and H. Jégou, "XCiT: Cross-covariance image transformers," in *Proc. NeurIPS*, 2021, pp. 20 014–20 027.

[53] S. Lv, B. Ouyang, Z. Deng, T. Liang, S. Jiang, K. Zhang, J. Chen, and Z. Li, "A dataset for deep learning based detection of printed circuit board surface defect," *Sci. Data*, vol. 11, no. 1, p. 811, 2024.

[54] W. Huang, P. Wei, M. Zhang, and H. Liu, "Hripcb: a challenging dataset for pcb defects detection and classification," *The Journal of Engineering*, vol. 2020, no. 13, pp. 303–309, 2020.

[55] S. Tang, F. He, X. Huang, and J. Yang, "Online pcb defect detector on a new pcb defect dataset," *arXiv preprint arXiv:1902.06197*, 2019.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. NeurIPS*, 2017.

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR*, 2018, pp. 586–595.

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[59] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Proc. ICLR*, 2021.

[60] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," in *Proc. ICLR*, 2022.

[61] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *Proc. ICLR*, 2023.

[62] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task of detrs as fine-grained distribution refinement," in *Proc. ICLR*, 2025.

[63] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," in *Proc. CVPR*, 2025, pp. 15 162–15 171.

[64] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[65] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," in *Proc. NeurIPS*, 2024, pp. 107 984–108 011.

[66] Y. Sun, Y. Liu, Y. Tang, W. Pei, and K. Chen, "Anycontrol: create your artwork with versatile control on text-to-image generation," in *Proc. ECCV*, 2024, pp. 92–109.

[67] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[68] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. CVPR*, 2021, pp. 13 713–13 722.

[69] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *Proc. ICML*, 2021, pp. 11 863–11 874.

[70] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. ICASSP*, 2023, pp. 1–5.