Cascaded Semantic-Guided Network for Surface Anomaly Localization in Steel-Cased Batteries

Ying Xu, Dongyan Lin, Tianlei Wang, Yingwen Chen, Yikui Zhai, *Senior Member, IEEE,* Jianhong Zhou, *Member, IEEE,* Pasquale Coscia, *Senior Member, IEEE,* Angelo Genovese, *Senior Member, IEEE,* C. L. Philip Chen, *Life Fellow, IEEE*

***Abstract*—Localization of surface anomalies in steel-cased batteries remains challenging due to subtle defect textures, nonuniform illumination, and complex surface reflections. Existing methods often struggle to separate fine-grained anomalies from normal patterns, leading to incomplete or inaccurate localization. To address these challenges, this paper presents the cascaded semantic-guided network (CSG-Net) framework that performs coarse-to-fine anomaly localization through cross-stage semantic refinement. Within the cascaded architecture, each stage refines anomaly localization using semantic priors from the preceding stage. In the coarse stage, a reconstruction–discrimination subnetwork models normal appearance distributions to generate preliminary anomaly cues. In the fine stage, a semantic-guided enhancement module (SGEM) transfers semantic priors from the coarse stage to adaptively modulate encoder features, thereby strengthening semantic consistency and enhancing discriminative capability. Furthermore, a deep supervision mechanism with weighted multi-scale loss is introduced to improve sensitivity to small defects and stabilize convergence. A complete steel-cased battery surface anomaly detection system is also developed and deployed on an industrial production line to validate engineering feasibility. Experiments on the steel-cased battery surface anomaly dataset (SCBSA), together with evaluations on MVTec AD and VisA, show that CSG-Net delivers strong localization performance across industrial inspection settings. Our code and dataset are available at https://github.com/yikuizhai/CSG-Net.**

***Index Terms*—Steel-cased batteries, surface anomaly detection, cascaded semantic-guided enhancement, industrial inspection system.**

This study was funded by Guangdong Higher Education Innovation and Strengthening School Project (No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2023760300070008390), Guangdong Province Key Disciplines Scientific Research Capacity Enhancement Project (No. 2024ZDJS034), National Natural Science Foundation of China (No. 62273109). This work was supported by the High Performance Computing Center of Wuyi University. (Ying Xu, Dongyan Lin, Tianlei Wang, Yingwen Chen, Yikui Zhai and Jianhong Zhou contributed equally to this work.) (Corresponding author: Yikui Zhai.)

Ying Xu, Dongyan Lin, Tianlei Wang, Yingwen Chen and Yikui Zhai are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, 529020, China (e-mail: xuying117@163.com; iamlindongyan@163.com; tianlei.wang@aliyun.com; yingwenchen2024@163.com; yikuizhai@163.com).

Jianhong Zhou are with the School of Electronics and Information Engineering, South China University Of Technology, Guangzhou, Guangdong 510641, China (e-mail: jackychou_lab@126.com).

Pasquale Coscia, Angelo Genovese are with the Department of Computer Science, Università Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

C. L. Philip Chen is with the Faculty of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510006, China (e-mail: philip.chen@ieee.org).

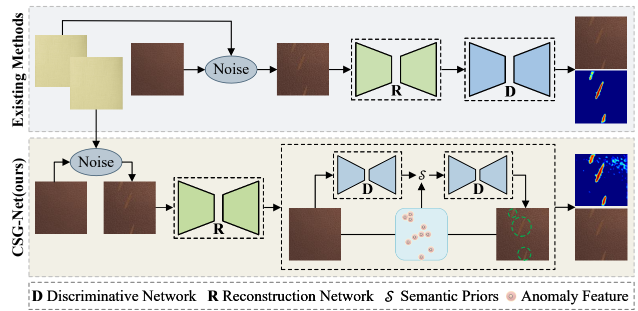


Fig. 1. Overview of conventional reconstruction-based anomaly detection methods. Existing methods mainly rely on reconstruction-based anomaly cues without semantic guidance, leading to limited localization accuracy. In contrast, our CSG-Net framework employs a cascaded coarse-to-fine strategy, where semantic priors from the coarse stage progressively enhance discriminative learning in subsequent stages for improved localization.

I. Introduction

BATTERIES serve as essential energy carriers in modern consumer electronics. Metal-cased batteries are widely used due to their high mechanical strength and resistance to external impacts, which enhance structural stability and operational safety [1]–[3]. During large-scale manufacturing, process variations, equipment wear, and environmental contamination can introduce defects on the steel shell surface, typically classified as mechanical damage (e.g., scratches, dents) or surface contamination (e.g., stains, dust). Even minor defects may compromise shell protection, affect sealing and thermal performance, and degrade product appearance [4]–[6]. More critically, such defects can lead to leakage, overheating, or even explosion. As a result, affected batteries are often scrapped, causing material waste and reduced production efficiency. These considerations underscore the importance of reliable inspection and early detection of surface defects to ensure both safety and manufacturing efficiency.

Recent advances in deep learning have significantly promoted surface defect inspection and industrial automation. Existing multi-scale feature fusion networks, such as PGA-Net [7] and MGNet [8], improve the detection of small surface defects by incorporating global contextual information and attention mechanisms. Nevertheless, these methods are primarily developed under fully supervised settings and rely heavily on sufficient high-quality annotated defect samples [9], which limits their applicability in industrial scenarios with scarce defective samples and costly pixel-level annotation. Accordingly, increasing attention has been directed toward less annotation-dependent paradigms, especially anomaly detection methods that learn the representation of normal samples and identify deviations as anomalies. Semi-supervised and limited-supervision strategies provide a practical alternative when only limited anomalous samples are available [10]. In this context, BGAD [11] enhances anomaly discrimination with a few known anomalies through an explicit boundary-guided contrastive learning mechanism. As illustrated in Fig. 1, existing reconstruction-based anomaly detection methods have achieved notable progress in industrial anomaly detection. DRAEM [12] learns a joint representation of an anomalous image and its anomaly-free reconstruction while modeling a decision boundary between normal and anomalous samples through simulated anomalies. BiSiNet [13] introduces a reference-guided siamese framework that extracts target–reference difference feature pyramids and reconstructs anomaly-related localization cues through a bidirectional feature pyramid decoder. These methods are effective in modeling normal appearance and detecting deviations. However, single-stage reconstruction frameworks still face challenges in precise anomaly localization for industrial surface inspection. When fine-grained defects coexist with appearance variations such as reflections, non-uniform brightness, and boundary-like structures, reconstruction-only anomaly cues may be insufficient for stable and discriminative localization.

To enhance the accuracy of anomaly localization, two-stage non-symmetrical deep convolutional autoencoders proposed by Zhu et al. [14] use a pretrained encoder to generate discriminative representations that are fed to dual decoders, and the combined reconstruction errors serve as an anomaly score. Liu et al. [15] first trained a classification network to extract discriminative representations, upon which a one-class classifier was constructed to distinguish normal and abnormal patterns. Liu et al. [16] refined this design by introducing a quantile-based thresholding mechanism in the decision space. Recent studies highlight that cross-domain feature learning improves generalization and provides a theoretical foundation for more accurate anomaly localization [16], [17]. Existing multi-scale and two-stage frameworks have improved detection performance by enhancing feature fusion and discriminative representation. Existing multi-scale and two-stage designs have shown that cues generated in an earlier processing stage can provide useful information for subsequent localization refinement. This is consistent with the idea of cascaded refinement, where outputs or cues from an earlier stage can be further used by a later stage to improve subsequent predictions [18]. From this perspective, two-stage designs can be regarded as a shallow form of cascaded refinement. However, most existing methods mainly focus on structural fusion, input-level combination, or decision-level refinement, while how coarse-stage anomaly cues are preserved and used for feature refinement in subsequent stages remains less explicitly modeled. The cross-stage semantic coherence discussed in this study refers to the preservation and utilization of early-stage anomaly-related semantic cues during subsequent feature refinement and localization in cascaded anomaly localization. For steel-cased battery surface inspection, background interference may increase the difficulty of distinguishing defect regions from background regions. Therefore, in the absence of explicit cross-stage semantic guidance, such methods may not fully exploit coarse-stage anomaly cues to constrain fine-stage feature refinement, resulting in less concentrated defect responses or false activations in non-defective background regions.

To address the aforementioned issues, we propose a cascaded semantic-guided network (CSG-Net) for surface anomaly localization. CSG-Net is built on a cascaded coarse-to-fine architecture, where the coarse stage provides initial anomaly responses that guide the fine stage to refine subtle defect regions and ambiguous boundaries. This design enables anomaly information to be progressively transferred and refined across stages, improving localization robustness under complex surface variations. To enhance cross-stage interaction, we introduce a semantic-guided enhancement module (SGEM), which converts coarse anomaly responses into semantic priors and injects them into fine-stage multi-scale features. Guided by these priors, the fine stage enhances defect-related responses while suppressing background interference caused by reflections, illumination variations, and boundary-like structures. By integrating cascaded refinement and SGEM-based semantic guidance, CSG-Net achieves accurate anomaly localization in complex industrial scenarios.

The main contributions of this work are summarized as follows:

1) The proposed CSG-Net introduces a cascaded semantic-guided framework for coarse-to-fine surface anomaly localization. It progressively refines anomaly representations through cross-stage semantic guidance, enabling finer-grained localization and more discriminative anomaly representation under industrial appearance variation.
2) The proposed SGEM adaptively modulates cross-scale semantic information from both the coarse and fine stages, preserving fine-grained structural details and enhancing the discriminability between normal and anomalous patterns.
3) A steel-cased battery surface anomaly detection (SCBSAD) system is designed and implemented on a real production line, demonstrating the engineering feasibility and industrial applicability of the proposed framework.
4) Extensive experiments on the SCBSA dataset validate CSG-Net for surface anomaly localization, with extended evaluations on the MVTec AD and VisA datasets demonstrating robust performance across industrial settings.

The remainder of this article is organized as follows. Section II briefly reviews the related work. Section III introduces the SCBSAD system. Section IV presents the proposed CSG-Net. Section V reports extensive experimental results to validate the effectiveness of the proposed method. Section VI further evaluates the generalization performance of the proposed system across different datasets. Finally, Section VII concludes this article.

II. Related Work

In industrial surface anomaly inspection, CNNs have proven capable of autonomously and effectively extracting image features. For instance, they have been applied to metallic textures, electronic components, and battery surfaces [19], [20]. However, defects often exhibit unpredictable forms and severities, making models trained on predefined categories and annotated samples less robust. Training solely on normal samples offers a more scalable and generalized solution.

Unsupervised anomaly detection (UAD) therefore focuses on modeling the distribution of defect-free surfaces, identifying deviations from it as potential anomalies. Recent UAD approaches can be broadly grouped into two main categories: representation-based methods and reconstruction-based methods, to which our method belongs.

Representation-based methods aim to project input images into a feature space via a mapping function, where normal and defective samples can be effectively distinguished [21], [22]. Roth et al. [23] introduced a memory bank of nominal patch features and combined it with pretrained embeddings and outlier detection, enabling anomaly identification by directly comparing test patch features with stored exemplars. Wan et al. [24] proposed bidirectional and multihierarchical feature mapping, which transfers images between pretrained feature spaces to highlight anomalous regions. Tsai et al. [25] modeled the relative feature similarity across multi-scale local regions to enhance the discriminative capability of learned representations. RD4AD [26] introduces a reverse distillation paradigm that reconstructs multi-scale teacher representations from a one-class embedding via a teacher–student architecture, enabling anomaly localization through representation discrepancy. CFLOW [27] constructs a multi-scale feature pyramid from feature maps at different levels to capture diverse receptive fields and performs pixel-wise anomaly localization by estimating the likelihood of each feature vector at its spatial position. Although these representation-based methods effectively exploit pretrained features for unsupervised anomaly detection, they may still be constrained by high memory consumption in patch-level modeling and by limited robustness to intra-class variations or feature distribution shifts.

Reconstruction-based methods are widely adopted in image anomaly detection, as they explicitly model the distribution of normal samples and identify deviations through reconstruction errors, enabling precise localization of anomalous regions [12], [28], [29]. Such methods aim to reconstruct samples lying on the manifold of normal data through generative models like GANs [30], AE [31]. Gong et al. [32] introduced a memory-augmented autoencoder that retrieves prototypical normal patterns for reconstruction, improving anomaly detection by increasing reconstruction errors on abnormal inputs. Hou et al. [33] proposed a divide-and-assemble reconstruction framework with multi-scale block-wise memory to enhance anomaly detection by modulating reconstruction granularity. Mishra et al. [34] combined transformer-based patch embeddings with a Gaussian mixture density network for anomaly localization. Yan et al. [35] proposed SCADN, which leverages multi-scale masked reconstruction to capture semantic context from normal samples for detection. Lv et al. [36] proposed a background reconstruction-based method that integrates an autoencoder-GAN framework with a U-Net inspection network to perform pixel-wise defect detection on textured surfaces. Zavrtanik et al. [37] proposed RIAD, a self-supervised reconstruction-by-inpainting framework that improves anomaly detection by reconstructing images from partially removed regions. Peng et al. [38] combine key-area recognition with local feature reconstruction for bolt anomaly detection, where normal and anomalous regions are first localized and their latent feature embeddings are then reconstructed for anomaly modeling and detection. However, most reconstruction-based approaches still struggle to maintain stable and accurate localization in practical industrial inspection, particularly for fine-grained anomalies, as they rely primarily on single-stage reconstruction and lack explicit semantic guidance across feature hierarchies.

Semantic-conditioned feature modulation methods inject conditioning cues into intermediate representations by predicting modulation parameters and applying them to feature activations, typically through feature-wise affine transformation or conditional normalization. FiLM [39] and conditional batch normalization [40] represent embedding-conditioned schemes, where a conditioning embedding is used to predict channel-wise scaling and bias for affine modulation of convolutional features. However, such channel-wise conditioning is spatially agnostic and therefore limited in providing location-sensitive guidance for fine-grained anomaly localization. SPADE [41] extends this idea by deriving spatially varying affine parameters from a conditioning map, enabling spatially adaptive modulation that better preserves the spatial structure of the guidance signal. In anomaly detection, semantic-guided designs have been introduced to enhance feature representation and facilitate reconstruction or localization, thereby improving the discrimination between normal patterns and anomalous responses [42], [43]. Such semantic cues can also help suppress background interference and emphasize task-relevant responses. Gao et al. [44] incorporated semantic constraints to guide feature learning and improve robustness in anomaly detection. Wang et al. [45] further utilized semantic context to guide representation learning, showing that semantic information is beneficial for distinguishing anomalous patterns from complex backgrounds. Different from these approaches, our method formulates semantic conditioning as a task-coupled intermediate representation derived at the coarse stage, which provides structural guidance for spatially adaptive modulation and supports progressive refinement of anomaly localization.

III. Steel-Cased Battery Surface Anomaly Detection System

A. Components of the SCBSAD System

The proposed SCBSAD system primarily consists of an AOI camera, a data center, and user terminals, as illustrated in Fig. 2. The entire workshop is designed as a fully enclosed environment to ensure inspection stability and effectively avoid external interference. During production, the AOI camera captures high-resolution images of the battery surface. The camera

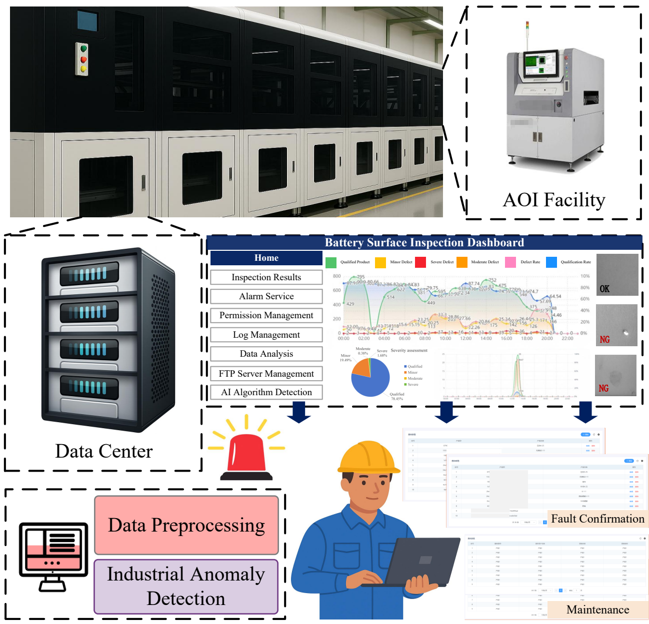


Fig. 2. Overall architecture of the SCBSAD system. The AOI device captures battery surface images and transmits the data to the data center for preprocessing and defect detection. The inspection results are presented through a visualization platform, which supports result management, system maintenance, and statistical analysis. Finally, the inspection outcomes are delivered to human operators for fault confirmation and subsequent maintenance.

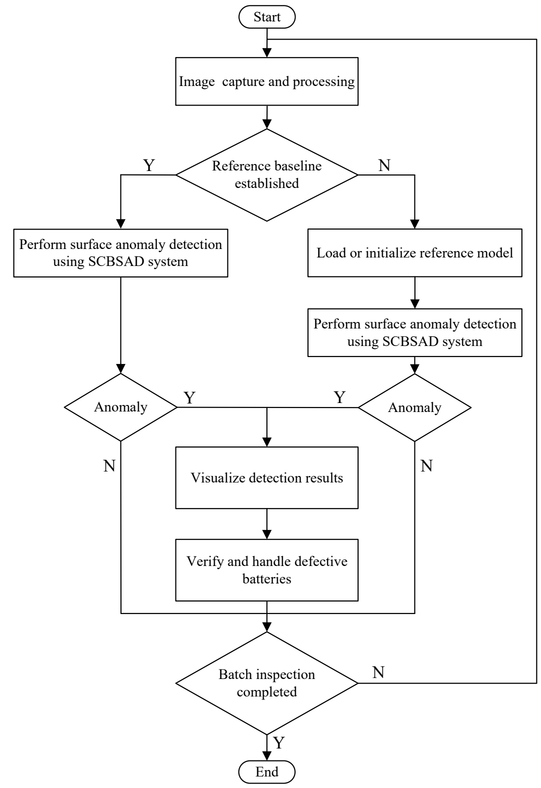


Fig. 3. Operational flowchart of the SCBSAD system for a single production line batch.

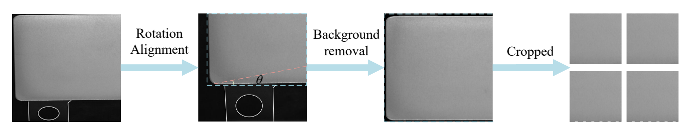


Fig. 4. Illustration of the preprocessing pipeline. Please zoom in for better visualization of details.

is configured as a 12-megapixel monochrome industrial camera equipped with a telecentric lens to guarantee geometric imaging accuracy. The spatial resolution is controlled within 10–15 μm/px, enabling the detection of fine scratches and dents. For illumination, a fixed industrial light source with a total luminous flux of approximately 4000-6000 lm is used in combination with polarizers to suppress strong reflections from the steel-cased surface, thereby ensuring image clarity and stability. The acquired data are first transmitted and stored in the data center, where they are processed through the data preprocessing and industrial anomaly detection modules for analysis and recognition. The detection results are then integrated into an inspection dashboard, which can be remotely accessed in real time via smartphones or computers, allowing users to monitor inspection data and system operation status. In addition, the dashboard supports the export of anomaly data. Based on the exported information, operators can quickly locate the corresponding defective batteries on the production line and perform fault confirmation and maintenance operations. The system acquires surface images from multiple production lines under real manufacturing conditions. The data used in this study were accumulated through long-term industrial collection, followed by manual screening and expert annotation.

B. Workflow of the SCBSAD System

As shown in Fig. 3, the workflow of steel-cased battery surface anomaly detection begins with image capture and preprocessing. Before initiating inspection, the system automatically checks whether an existing inspection configuration is available for the current production line. If available, the corresponding model parameters are loaded to perform anomaly detection. Otherwise, the system initializes a new inspection configuration and collects standard samples to establish a baseline for subsequent inspections. The captured images are then analyzed by the SCBSAD system to locate anomalies, and the results are visualized for operator review. Once the inspection for the current batch is completed, the process either terminates or proceeds to handle the next production line or batch, ensuring continuous and standardized surface anomaly detection throughout industrial operations.

C. Image Processing and Testing Workflow

Raw images of steel-cased batteries are first standardized before being fed into the anomaly detection network. Since the captured images may have a certain angular offset, each battery is first rotated according to the tilt direction of the battery body, so that the main battery region is adjusted to a horizontal state. After rotation alignment, a keypoint-based region of interest (ROI) cropping method is adopted

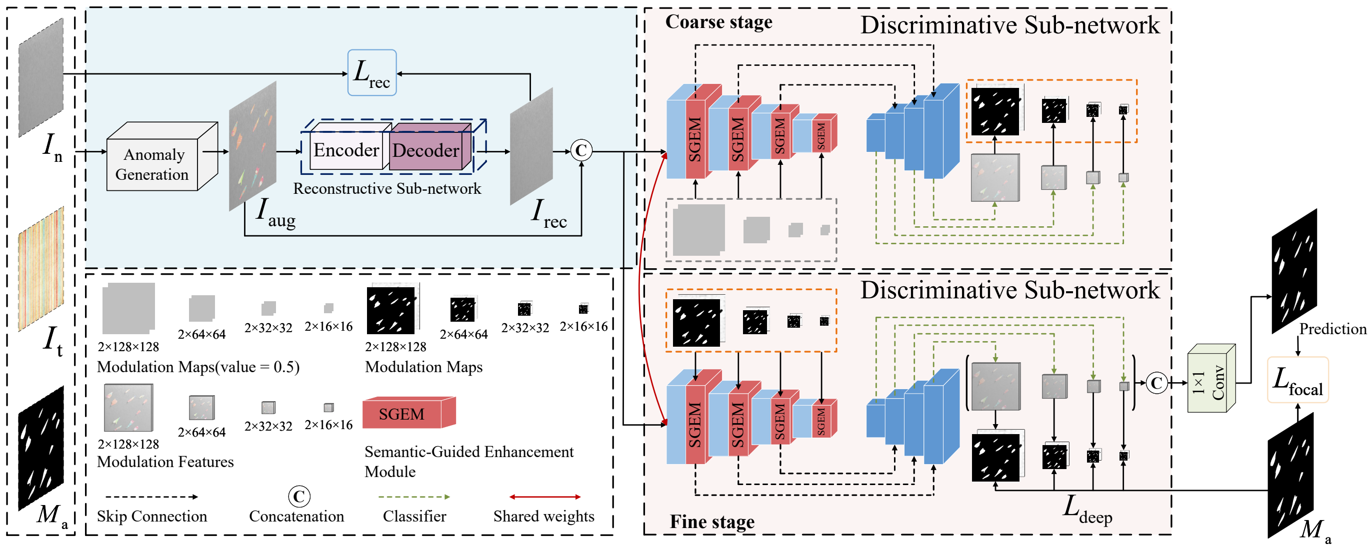


Fig. 5. Overall architecture of the proposed CSG-Net framework. The reconstructive sub-network adopts a U-Net-style encoder-decoder design inspired by the reconstruction branch in DRAEM [12] to generate a nearly defect-free reference image I_{rec}, which is concatenated with the augmented input I_{aug} to form a joint representation. In the coarse stage, an initial constant semantic modulation map (set to 0.5) is used to produce multi-scale coarse predictions, which are further normalized to generate semantic guidance for subsequent refinement. In the fine stage, the semantic guidance map derived from the coarse stage is injected into the encoder through SGEM to progressively refine anomaly localization. Deep supervision with weighted multi-scale losses is applied to the decoder outputs to improve training stability and localization accuracy.

to localize the effective battery surface region and remove irrelevant surrounding background while preserving the main inspection area. This step belongs to the system preprocessing stage; therefore, this study only provides an overview of the processing workflow without further expanding its implementation details. Since the illumination conditions, camera installation positions, and mechanical fixtures in battery production scenarios are usually relatively stable, the keypoint-based ROI cropping model has a certain generalization ability across similar production lines. For cross-line migration, only minor adjustments to the keypoint detection model or ROI cropping rules are usually required when the imaging angle, product specification, or fixture position changes significantly. The resulting cropped regions are then divided into fixed-size patches of 256×256 pixels. In practical inspection, defect regions often occupy only a small fraction of the image area, which increases the difficulty of anomaly localization. Therefore, this preprocessing pipeline reduces background and edge interference and enhances the stability of subsequent feature extraction. The overall procedure is illustrated in Fig. 4.

In practical deployment, the anomaly score map can be converted into a binary defect segmentation map using a predefined operating threshold. This threshold can be adjusted according to specific industrial requirements, such as the acceptable trade-off between missed detections and false alarms. In this study, for quantitative evaluation, the operating threshold for threshold-dependent metrics is determined on the validation set by maximizing the pixel-level Youden index and is then fixed for test-set evaluation.

IV. Methods

This section presents the proposed CSG-Net. It provides an overview of the cascaded CSG-Net framework, followed by a description of SGEM for cross-stage refinement. The anomaly synthesis and reconstructive sub-network are then described, along with the semantic-guided discriminative sub-network for multi-scale anomaly scoring. Finally, the loss functions and training objectives used to optimize the framework are presented.

A. Cascaded Framework Overview of CSG-Net

As illustrated in Fig. 5, CSG-Net aims to localize surface anomalies in steel-cased batteries through a cascaded coarse-to-fine framework. For steel-cased battery surfaces, relying only on reconstruction discrepancy may be insufficient for stable and discriminative anomaly localization. This is because weak-contrast defects, background texture variations, and reflective regions may produce ambiguous or similar anomaly responses, leading to background false activations, unstable responses, or unclear boundaries in single-stage localization results. To address this issue, this study adopts a coarse-to-fine cascade structure, which first generates preliminary anomaly responses in the coarse stage to locate potential defect-related regions and then further utilizes these coarse responses as semantic priors in the fine stage for stage-wise refinement of anomaly localization.

In the coarse stage, the discriminative sub-network D_ϕ receives the concatenated input $\text{Concat}(I_{\text{rec}}, I_{\text{aug}})$, where $I_{\text{rec}} = R_\theta(I)$. The modulation maps $M^{(1)}$ are initialized to 0.5 at all encoder scales to provide uniform guidance. Multi-scale coarse predictions $P^{(1)} = \{p_i^{(1)}\}_{i=1}^{L}$ are extracted and used to generate guidance maps:

$$P^{(1)} = D_\phi\big(\text{Concat}(I_{\text{rec}}, I_{\text{aug}}), M^{(1)}\big) \tag{1}$$

where each feature $p_i^{(1)}$ corresponds to a spatial resolution of $H_i \times W_i$ with C_i channels. The coarse predictions are

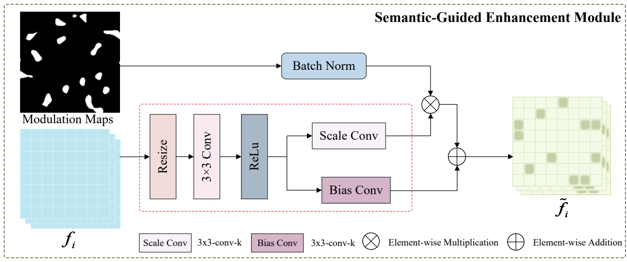


Fig. 6. Architecture of the SGEM. The module introduces semantic modulation features into the multi-scale encoder representations. It normalizes the feature maps and adaptively modulates them using semantic information, thereby enhancing cross-stage semantic interactions.

then normalized via a softmax operation to produce semantic guidance maps $M^{(2)} = \text{Softmax}(P^{(1)})^{(0)}$, which serve as priors for the fine stage, where the superscript (0) denotes the background probability channel.

In the fine stage, the discriminative sub-network D_ϕ is modulated by the semantic guidance maps $M^{(2)}$ to refine multi-scale features $P^{(2)} = \{p_i^{(2)}\}_{i=1}^{L}$ and generate a fused anomaly map $P^{(2)}_{\text{fused}}$. The fused map are further normalized through a softmax operation to produce the final probability map, representing the pixel-level anomaly confidence. Within this process, SGEM is used to introduce the coarse-stage semantic priors into fine-stage feature modulation, while weighted multi-scale deep supervision is applied to constrain predictions at different scales in the fine stage. During inference, the fused map $P^{(2)}_{\text{fused}}$ is used to estimate the final anomaly score and localize defect regions.

B. Semantic-Guided Enhancement Module

Within the CSG-Net framework, each stage progressively refines the anomaly cues produced by the previous stage. Directly using coarse predictions as additional inputs may provide spatial cues, but it may not sufficiently adapt fine-stage feature responses according to the anomaly semantics encoded in the coarse stage. Therefore, SGEM is incorporated at each scale to transform semantic guidance maps into spatially and channel-wise modulation parameters, enabling multi-scale encoder features to be adaptively enhanced for cross-stage refinement, as illustrated in Fig. 6. Let $f_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ denote the input feature map at the i-th encoder layer, and $M^{(s)} \in \mathbb{R}^{1 \times H_i \times W_i}$ the corresponding semantic guidance map. SGEM first aligns the spatial dimensions of $M^{(s)}$ with f_i, producing $\tilde{M}^{(s)}$. A learnable mapping function $\mathcal{F}^{(l)}(\cdot)$ then transforms $\tilde{M}^{(s)}$ into channel-wise adaptive scaling and bias parameters:

$$\gamma^{(i)}, \beta^{(i)} = \mathcal{F}^{(i)}\left(\tilde{M}^{(s)}\right) \tag{2}$$

where $\gamma^{(i)}, \beta^{(i)} \in \mathbb{R}^{C_i \times H_i \times W_i}$ denote the channel-wise and spatially adaptive scaling and bias parameters, respectively, which are applied to modulate the feature map f_i. In practice, $\mathcal{F}^{(i)}(\cdot)$ is implemented with convolutional layers and nonlinear activation, enabling the semantic guidance map to dynamically modulate the encoder feature map f_i. The enhanced feature map $\tilde{f}_i$ is computed as:

$$\tilde{f}_i = f_i \odot \left(1 + \gamma^{(i)}\right) + \beta^{(i)} \tag{3}$$

where $\odot$ denotes element-wise multiplication.

This spatially adaptive modulation enables SGEM to suppress background redundancy and emphasize suspicious regions, enhancing the encoder's sensitivity to fine-grained anomaly patterns and strengthening the representation of subtle anomalous details.

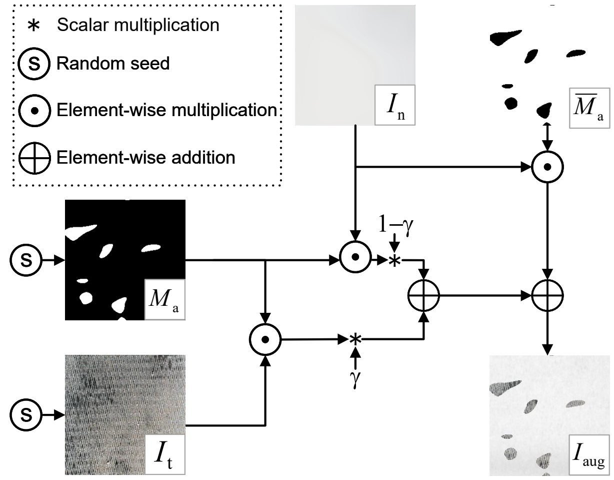


Fig. 7. Structural diagram of the anomaly generation module. The anomaly mask M_{a} and texture image I_{t} both generated from different random seeds, are fused with the normal image I_{n}, where the blending process is governed by the opacity parameter γ. The complementary mask M_{a} is used to preserve the normal regions. Finally, the anomaly augmented image I_{aug} is obtained through element-wise addition and multiplication operations.

C. Anomaly Synthesis and Reconstruction Pipeline

In industrial anomaly detection scenarios, normal samples typically dominate the data distribution, while defective samples are not only extremely scarce but also exhibit a sparse distribution. This class imbalance severely limits the model's ability to learn effective representations of anomalous patterns. Inspired by DRAEM [12], the proposed anomaly generation module synthesizes anomaly-augmented images that exhibit distributional deviation characteristics. The anomaly texture source image I_{t} is obtained through random augmentation and sampling from the describable texture dataset (DTD) [46], which provides a rich variety of natural textures for generating pseudo anomalies. Although DTD is not an industrial dataset, its diverse texture patterns help simulate heterogeneous surface perturbations, thereby enhancing the model's generalization capability to unseen defect appearances. As illustrated in Fig. 7, we randomly combine Perlin noise with a binarized mask of the original sample to generate an anomaly mask M_{a}. The anomaly generation module then takes the anomaly mask M_{a}, the anomaly texture image I_{t}, and the normal image I_{n}, and then computes the background mask $\bar{M}_{\text{a}}$ by inverting M_{a}. The augmented training image I_{aug} defined as:

$$I_{\text{aug}} = \gamma\left(M_{\text{a}} \odot I_{\text{t}}\right) + (1-\gamma)\left(M_{\text{a}} \odot I_{\text{n}}\right) + I_{\text{n}} \odot \bar{M}_{\text{a}} \tag{4}$$

where $\bar{M}_{\mathrm{a}}$ denotes the binary inverse of M_{a}, $\odot$ indicates element-wise multiplication, and γ is an opacity coefficient uniformly drawn from [0.1,1.0].

The anomaly augmented images I_{aug} produced by the previous module are fed into a reconstructive sub-network designed to capture the distribution of normal surface patterns. This network transforms locally perceived features into representations aligned with normal samples, enabling it to suppress abnormal perturbations and restore the underlying normal structure rather than merely reconstructing seen normal images. Specifically, the sub-network consists of an encoder and a decoder: the encoder progressively extracts multi-level semantic features through stacked convolutional layers and downsampling operations, while the decoder gradually restores the spatial structure and fine details via upsampling and skip connections. The reconstructed image I_{rec} is then obtained as:

$$I_{\mathrm{rec}} = D\left(E_i(I_{\mathrm{aug}})\right)_{i=1}^{L} \tag{5}$$

where $I_{\mathrm{aug}} \in \mathbb{R}^{3\times H\times W}$ denotes the anomaly augmented image, $E_i(\cdot)$ represents the output feature map of the i-th layer in the encoder, and $D(\cdot)$ denotes the decoder that performs upsampling and skip fusion across layers to reconstruct $I_{\mathrm{rec}} \in \mathbb{R}^{3\times H\times W}$. The symbol L indicates the total number of encoding layers. Through this process, the network learns to approximate the distribution of normal samples, suppressing anomalous regions in the reconstructed output and facilitating the discriminative sub-network in highlighting structural discrepancies between the input and its reconstruction.

During the training phase, a joint reconstruction loss is formulated by combining the pixel-wise mean squared error (L2) and the structural similarity index measure (SSIM), aiming to simultaneously optimize pixel-level fidelity and local structural consistency. The overall reconstruction loss is defined as:

$$L_{\mathrm{rec}} = L_{\mathrm{SSIM}}(I_{\mathrm{n}}, I_{\mathrm{rec}}) + L_2(I_{\mathrm{n}}, I_{\mathrm{rec}}) \tag{6}$$

where the L_2 term measures the pixel-wise discrepancy between the normal image I_{n} and the reconstructed image I_{rec}. The SSIM loss is further expressed as:

$$L_{\mathrm{SSIM}} = \frac{1}{N_{\mathrm{p}}} \sum_{i=1}^{H} \sum_{j=1}^{W} \left[1 - \mathrm{SSIM}\left(I_{\mathrm{n}}, I_{\mathrm{rec}}\right)(i,j)\right] \tag{7}$$

where H and W denote the height and width of the image, and N_{p} represents the total number of pixels. The term $\mathrm{SSIM}(I_{\mathrm{n}}, I_{\mathrm{rec}})(i,j)$ indicates the structural similarity score computed between local patches centered at pixel (i,j) in the normal image I_{n} and the reconstructed image I_{rec}.

D. Semantic-guided Discriminative Sub-network

1) Encoder–Decoder Architecture: The discriminative sub-network D_ϕ is designed to perform anomaly localization under the semantic guidance propagated from the coarse stage. It follows an FPN-style encoder–decoder topology for multi-scale feature refinement. The encoder extracts multi-scale features, which are modulated by SGEM using the semantic guidance maps derived from the coarse-stage responses. This modulation allows the refinement stage to emphasize anomaly-related regions while suppressing redundant background responses, thereby enhancing the separability between normal and anomalous patterns. Details of SGEM and the generation of semantic guidance maps are provided in Section IV-B.

The decoder follows a top-down pathway that progressively upsamples and fuses encoder features to recover fine spatial structures. Multi-level features are aggregated to produce prediction maps at stage s , defined as:

$$\{p_i^{(s)}\}_{i=1}^{4} = D_\phi\Big(C(I_{\mathrm{rec}}, I_{\mathrm{aug}}), M^{(s)}\Big), \quad s = 1, 2. \tag{8}$$

where $C(I_{\mathrm{rec}}, I_{\mathrm{aug}})$ represents the concatenated reconstruction–augmentation pair, and $M^{(s)}$ denotes the semantic modulation maps generated by SGEM, D_ϕ indicates the mapping performed by the discriminative sub-network to produce multi-scale predictions.

In the coarse stage ($s = 1$), a constant prior $M^{(1)} = 0.5$ stabilizes feature contrast learning. In the refinement stage ($s = 2$), the prior $M^{(2)}$ is derived from the background channel of the coarse predictions $p_i^{(1)}$ to provide semantic cues for boundary refinement and contextual enhancement. This hierarchical modulation strategy enables progressive feature alignment and fine-grained anomaly localization across scales.

2) Multi-scale Prediction Fusion: In the fine stage, the decoder produces multi-scale predictions $\{P_i^{(2)}\}_{i=1}^{4}$. Previous studies have indicated that deep supervision can impose direct constraints on intermediate predictions, allowing features at different scales to participate in model optimization, and can therefore be used to enhance the model's perception of small target regions [47], [48]. Motivated by this, we introduce multi-scale supervision to the fine-stage decoder outputs to encourage the model to preserve defect-related responses across different decoder levels. The multi-scale predictions are then upsampled, concatenated, and passed through a 1×1 convolution to obtain the fused logits $P_{\mathrm{fused}}^{(2)}$, which are converted to the final anomaly segmentation prediction $P^{(2)}$ by softmax normalization:

$$P^{(2)} = \mathrm{Softmax}\big(P_{\mathrm{fused}}^{(2)}\big) \tag{9}$$

The second channel of $P^{(2)}$ represents the pixel-wise anomaly probability. This fusion strategy effectively combines local detail and semantic context to produce the final anomaly segmentation masks. The final fused prediction is supervised by the ground-truth anomaly mask M_{a} using focal loss. To further constrain the fine-stage multi-scale predictions, deep supervision is introduced in the training objective, as described in Section IV-E.

E. Loss Function

To enable end-to-end optimization, the proposed CSG-Net employs a unified joint training strategy, as illustrated in Fig. 5 and summarized in Algorithm 1. Both subnetworks are optimized simultaneously through backpropagation, facilitating collaborative feature refinement and stable convergence.

During training, semantic modulation maps are progressively introduced into the discriminative pathway, guiding multi-scale feature refinement and enhancing the network's sensitivity to subtle anomaly patterns. Concurrently, the reconstruction sub-network is supervised to minimize reconstruction errors, providing complementary priors that support discriminative feature learning. For the discriminative sub-network, the fine-stage multi-scale predictions are supervised by the ground-truth anomaly mask M_{a} using focal and Dice losses. The deep supervision loss is defined as:

$$L_{\text{deep}} = \sum_{i=1}^{4} \lambda_i \left[L_{\text{focal}}(P_i^{(2)}, M_{\text{a}}) + L_{\text{dice}}(P_i^{(2)}, M_{\text{a}}) \right] \quad (10)$$

where λ_i balances the contribution of each scale. The final fused prediction $P^{(2)}$ is further supervised by focal loss, and the total discriminative sub-network loss is formulated as:

$$L_{\text{disc}} = L_{\text{deep}} + L_{\text{focal}}(P^{(2)}, M_{\text{a}}) \quad (11)$$

By combining the reconstruction and discriminative objectives, the overall optimization is formulated as:

$$L_{\text{total}} = L_{\text{rec}} + L_{\text{disc}} \quad (12)$$

thereby ensuring effective alignment between reconstruction and discriminative learning and promoting robust feature adaptation under complex conditions.

Algorithm 1 Training Procedure of CSG-Net

Input: Normal images I_{n}, anomaly augmented images I_{aug}, ground-truth anomaly mask M_{a}, reconstructive sub-network R_θ, discriminative sub-network D_ϕ (comprising an encoder D_ϕ^{enc} and a decoder D_ϕ^{dec}), maximum epochs T

Output: Optimized parameters $\{\theta, \phi\}$, anomaly segmentation masks $P^{(2)}$

for epoch $= 1$ to T **do**
  for each sample pair $(I_{\text{n}}, I_{\text{aug}})$ in batch **do**
    Step 1: Anomaly-free Reference Restoration
    $I_{\text{rec}} \leftarrow R_\theta(I_{\text{aug}})$
    $L_{\text{rec}} \leftarrow \|I_{\text{n}} - I_{\text{rec}}\|_2^2 + (1 - \text{SSIM}(I_{\text{n}}, I_{\text{rec}}))$
    Step 2: Coarse-stage semantic prior extraction
    Initialize modulation maps: $M^{(1)} \leftarrow 0.5$ at all encoder scales
    $\{p_i^{(1)}\}_{i=1}^{L} \leftarrow D_\phi^{\text{dec}}(\text{Concat}(I_{\text{rec}}, I_{\text{aug}}), M^{(1)})$
    for $i = 1$ to L **do**
      $\hat{p}_i^{(1)} \leftarrow \text{Softmax}(p_i^{(1)})$
      $M_i^{(2)} \leftarrow \hat{p}_i^{(1)}[0]$
    end for
    Step 3: Fine-stage semantic-guided feature refinement
    $\{p_i^{(2)}\}_{i=1}^{L}, P_{\text{fused}}^{(2)} \leftarrow D_\phi(\text{Concat}(I_{\text{rec}}, I_{\text{aug}}), M^{(2)})$
    $P^{(2)} \leftarrow \text{Softmax}(P_{\text{fused}}^{(2)})$
    Step 4: Loss calculation and network update
    $L_{\text{disc}} \leftarrow L_{\text{deep}} + \text{FocalLoss}(P^{(2)}, M_{\text{a}})$
    $L_{\text{total}} \leftarrow L_{\text{rec}} + L_{\text{disc}}$
    Update parameters $\{R_\theta, D_\phi\}$ via backpropagation
  end for
end for

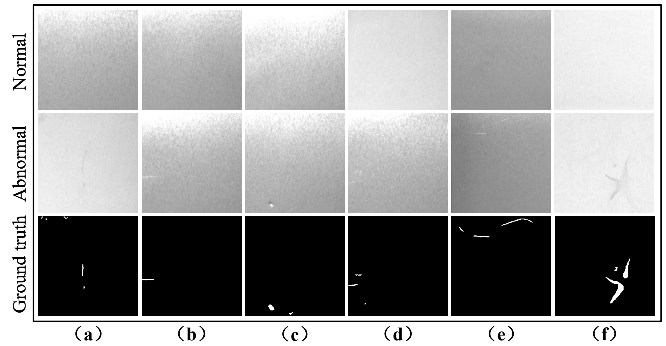


Fig. 8. Examples of normal and anomalous samples with corresponding ground-truth masks from the SCBSA dataset.

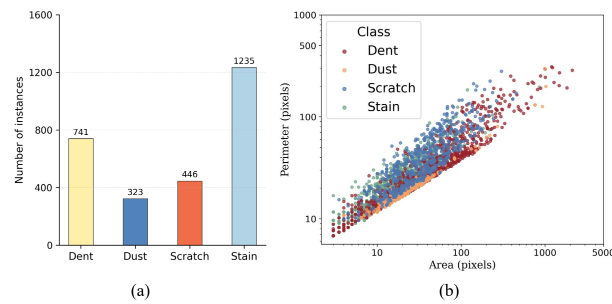


Fig. 9. Attributes of the SCBSA dataset. (a) Distribution of defect samples. (b) Scatter plot of area versus perimeter.

V. Experiments

A. Steel-Cased Battery Surface Anomaly Dataset

To evaluate the effectiveness of the proposed method, experiments were conducted on the SCBSA dataset. The image data were acquired using the AOI inspection system described in Section III for high-resolution surface imaging of steel-cased batteries. The dataset contains 4,207 images and is split into training, validation, and test sets: 1,166 normal samples for training, 315 normal samples and 697 defective samples for validation, and 633 normal samples and 1,396 defective samples for testing. All defective images are annotated with pixel-level ground truth masks by domain experts with industrial inspection experience. In the raw images, before the image processing procedure described in Section III-C, defect pixels account for only about 0.027% of the entire battery area. Notably, the data used for training, validation, and testing were collected from different production lines rather than randomly sampled from a single line, making the evaluation more relevant to practical industrial inspection scenarios.

Fig. 8 presents representative samples from the SCBSA dataset. Besides diverse anomalous morphologies, the dataset also contains appearance variations in normal regions, such as reflective interference, non-uniform brightness, and boundary-like structures, which increase the difficulty of anomaly discrimination. In addition, some anomalous regions are small, low-contrast, or irregular, making accurate localization more difficult. Fig. 9 further indicates that defect categories are

TABLE I
COMPARISON OF ANOMALY DETECTION METHODS ON THE SCBSA DATASET. THRESHOLD-FREE METRICS INCLUDE I-AUROC, P-AUROC, AUPRO, AND PR-AUC. THRESHOLD-DEPENDENT METRICS INCLUDE PRE., REC., F1, MR, AND FAR. ALL VALUES ARE REPORTED IN PERCENT (%). ALL PIXEL-LEVEL METRICS ARE COMPUTED BASED ON FULL-PIXEL STATISTICS. THE OPERATING THRESHOLD IS DETERMINED BY MAXIMIZING THE PIXEL-LEVEL YOUDEN INDEX ON THE VALIDATION SET AND THEN FIXED FOR TEST-SET EVALUATION. "FEATURE. METHODS" DENOTES FEATURE-BASED METHODS, AND "RECON. METHODS" DENOTES RECONSTRUCTION-BASED METHODS.

Method		I-AUROC	P-AUROC	AUPRO	PR-AUC	Pre.	Rec.	F1	MR	FAR
Feature. Methods	BGAD [11]	58.14	68.82	47.79	1.69	0.13	**93.85**	0.27	**6.14**	82.77
	EfficientAD [49]	84.17	88.94	62.76	4.76	**2.27**	56.86	**4.36**	43.14	2.77
	UniNet [50]	68.29	78.03	51.60	1.53	0.41	61.73	0.81	38.27	17.15
Recon. Methods	DRAEM [12]	80.24	87.95	76.41	8.56	0.84	35.80	1.64	64.20	4.52
	SimpleNet [51]	78.92	79.11	54.74	2.02	0.58	59.65	1.15	40.35	11.41
	RD4AD [26]	**91.77**	84.70	82.36	5.25	0.72	83.87	1.42	16.13	52.83
	SK-RD4AD [52]	85.59	84.54	56.20	3.42	0.82	69.56	1.63	30.44	9.48
	BiSiNet [13]	87.91	88.81	76.60	11.99	0.75	64.32	1.48	35.68	9.56
	CSG-Net-2S	87.39	**93.56**	88.21	16.33	1.27	42.52	2.41	57.48	4.23
	CSG-Net-3S	88.64	93.15	**88.58**	**16.70**	1.38	48.59	2.68	51.41	**3.84**

imbalanced and that most defects occupy relatively small regions.

B. Evaluation Metrics

To comprehensively evaluate anomaly detection and defect localization performance, threshold-independent metrics, threshold-dependent metrics, and computational efficiency metrics are adopted. The threshold-independent metrics include image-level AUROC (I-AUROC), pixel-level AUROC (P-AUROC), AUPRO, and PR-AUC. For region-level localization evaluation, the per-region overlap (PRO) is used to measure the coverage of ground-truth defect regions, which is defined as:

$$\mathrm{PRO} = \frac{1}{N} \sum_{n=1}^{N} \frac{\mathrm{TP}_n}{\mathrm{TP}_n + \mathrm{FN}_n} \tag{13}$$

where N denotes the number of ground-truth defect regions, and TP_n and FN_n are the numbers of correctly predicted and missed pixels in the n-th defect region, respectively. The threshold-independent AUPRO is obtained by integrating the PRO curve over a specified false positive rate range:

$$\mathrm{AUPRO} = \frac{1}{\alpha} \int_0^{\alpha} \mathrm{PRO}(\mathrm{FPR})\, d(\mathrm{FPR}) \tag{14}$$

where $\alpha = 0.5$ in this study. In addition, Precision (Pre.), Recall (Rec.), F1, Miss Rate (MR), and False Alarm Rate (FAR) are reported as threshold-dependent metrics. The operating threshold is determined on the validation set by maximizing the pixel-level Youden index and then fixed for test-set evaluation. Pre., Rec., and F1 are used to evaluate pixel-level segmentation performance under the fixed threshold, while MR and FAR indicate missed detections and false alarms, respectively. Finally, model parameters (Params), GFLOPs, and inference speed (FPS) are reported to evaluate the computational efficiency of different model configurations.

C. Implementation Details

All comparative and ablation experiments on the SCBSA dataset were conducted under a consistent hardware and software environment. Training, validation, and testing were performed on an NVIDIA GeForce RTX 3060 GPU with 12 GB of memory using the SCBSA dataset. The models were trained for 200 epochs with a batch size of 2, using the Adam optimizer with an initial learning rate of 0.001 and a cosine annealing schedule for learning-rate adjustment. During training, approximately 25% of the samples were augmented via the anomaly generation module. All experiments were repeated three times, and the average test results are reported using the checkpoint selected according to the pixel-level AUROC on the validation set, while the test set was used only for final performance evaluation. For threshold-dependent metrics in the comparative experiments, including Pre., Rec., F1, MR, and FAR, the operating threshold was determined by maximizing the pixel-level Youden index on the validation set and then fixed for test-set evaluation. Considering the practical requirements of industrial inspection, pixel-level localization is emphasized while maintaining competitive image-level performance, as accurate defect localization is more critical for practical applications.

D. Comparative Experiments

Surface anomaly detection in steel-cased batteries remains challenging due to complex surface textures, subtle defect patterns, and strong reflections, which often obscure the distinction between anomalous and normal regions. To ensure a comprehensive evaluation, two representative categories of unsupervised anomaly detection methods were selected for comparison. The first category includes reconstruction-based approaches that detect anomalies by learning and reconstructing normal data distributions, represented by DRAEM [12], SimpleNet [51], RD4AD [26], SK-RD4AD [52], and BiSiNet [13]. The second category consists of feature-based approaches that identify deviations in learned representations, including BGAD [11], EfficientAD [49], and UniNet [50]. These two groups represent the mainstream strategies in current anomaly detection research and serve as strong baselines for validating the effectiveness of the proposed method. In the comparison, CSG-Net-2S and CSG-Net-3S denote the two-stage and three-stage cascaded variants of the proposed CSG-Net, respectively.

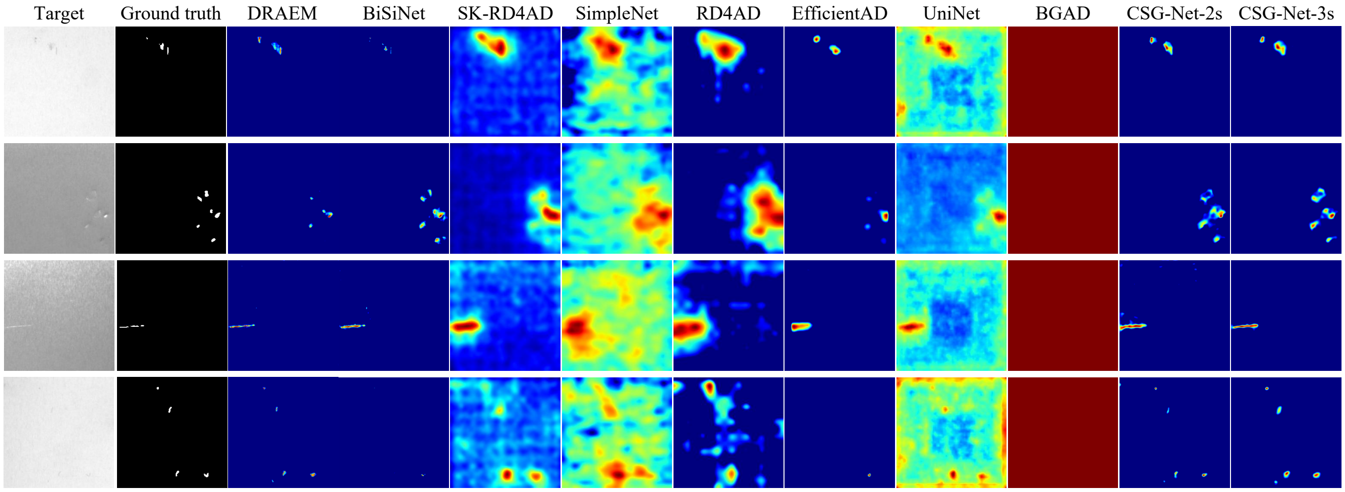


Fig. 10. The figure presents the anomaly localization heatmaps of different methods on the SCBSA dataset. The visualized results include the target images, ground-truth masks, and heatmaps generated by different methods. CSG-Net-2s and CSG-Net-3s denote the 2-stage and 3-stage Cascade versions of the proposed CSG-Net, respectively.

Table I presents the performance comparison of various methods on the proposed SCBSA dataset. In addition to AUROC-based metrics, PR-AUC and threshold-dependent metrics are further reported to provide a more comprehensive evaluation under imbalanced conditions. The results show that reconstruction-based methods generally outperform feature-based methods in localization, indicating the importance of modeling normal appearance for industrial surface anomaly detection. Although some methods show advantages on individual metrics, their performance should be interpreted together with pixel imbalance and false-alarm cost. For example, BGAD [11] obtains a high Rec., but this is accompanied by a very high FAR, indicating that many normal pixels are incorrectly activated as anomalies. EfficientAD [49] achieves the highest Pre., but its Rec. and F1 remain limited, suggesting insufficient defect coverage. RD4AD [26] achieves the best I-AUROC, but its PR-AUC and threshold-dependent metrics are less competitive under pixel-level imbalance. In contrast, CSG-Net-3S achieves the best AUPRO, PR-AUC, F1, and FAR, demonstrating a better balance between fine-grained defect localization and false-alarm suppression. These results indicate that the proposed cascaded semantic guidance improves localization robustness under industrial appearance disturbances and severe pixel-level imbalance. Therefore, in the comparative experiments with other methods, this study adopts the 3-stage Cascade configuration for performance comparison.

Fig. 10 illustrates the anomaly localization heatmaps of different methods on the SCBSA dataset. Overall, existing methods still suffer from varying degrees of incomplete localization or background false responses in challenging scenarios involving low-contrast defects, slender edge defects, and multiple small defects. For example, in the first several samples, some methods can respond to the main defective regions but tend to show response diffusion, blurred boundaries, or activations in non-defective areas. Other methods also produce weak responses to slender defects or local low-contrast defects, resulting in insufficient defect coverage. The fourth-row sample further demonstrates the challenge of localizing small defects. In this sample, the defective regions are small and spatially scattered. Some competing methods can only detect

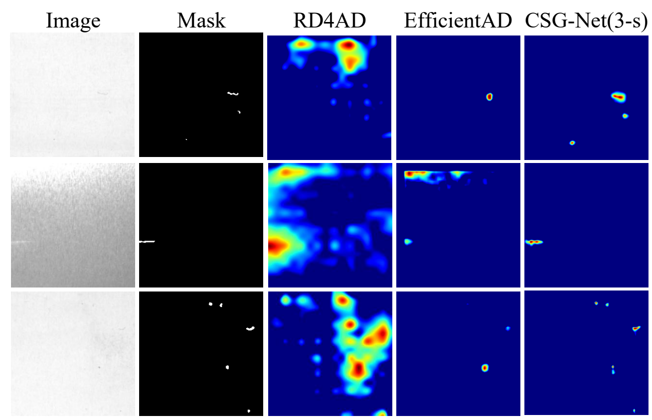


Fig. 11. Qualitative comparison of RD4AD, EfficientAD, and the proposed method on representative anomaly samples from the SCBSA dataset with appearance disturbance.

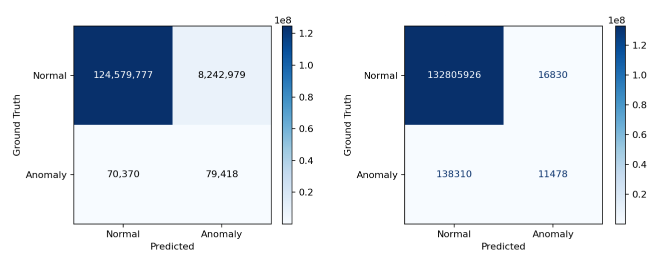


Fig. 12. Pixel-level confusion matrices of the baseline and CSG-Net on the test set using full-pixel statistics. The operating threshold is determined by maximizing the pixel-level Youden index on the validation set and then fixed for test-set evaluation.

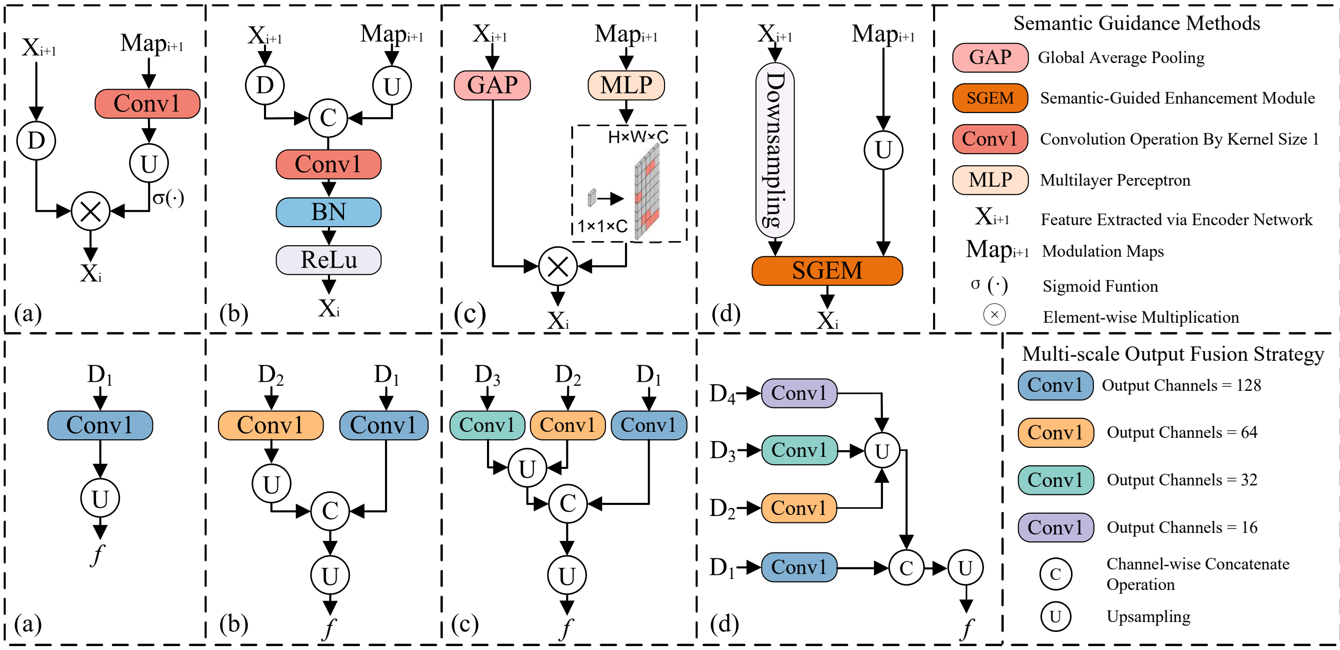


Fig. 13. Ablation variants of encoder modulation: (a) element-wise product; (b) concatenation + Conv–BN–ReLU; (c) SE-like channel modulation; (d) proposed SGEM. Ablation variants of decoder fusion: (a) single-scale output; (b) pairwise concatenation-based fusion; (c) progressive multi-level fusion; (d) full multi-scale fusion.

a few defect points or produce obvious false responses in the background. Although some methods generate anomaly responses, their response locations do not fully match the ground-truth annotations, making it difficult to ensure complete coverage of small defective regions. In contrast, CSG-Net-2s and CSG-Net-3s produce more accurate and concentrated responses around the small defect locations and maintain better coverage completeness for multiple small defective regions. These results indicate that the cascaded semantic-guided mechanism helps enhance the model's perception of small defects and improves the accuracy and completeness of anomaly localization under complex surface conditions.

As shown in Fig. 11, RD4AD, EfficientAD, and CSG-Net(3-s) exhibit different response patterns on representative anomaly samples from the SCBSA dataset. In the first row, the defect region has an appearance similar to the background. RD4AD produces scattered responses with noticeable activations in non-defective regions, while EfficientAD generates only a few local responses. By comparison, CSG-Net(3-s) mainly concentrates its responses on the defect-related region. In the second row, the upper part of the image shows noticeable appearance variation, whereas the true defect is mainly located along the left edge. Both RD4AD and EfficientAD respond to the upper appearance-variation region and the defect region, with RD4AD showing a larger and more scattered response distribution. By contrast, CSG-Net mainly concentrates its response on the left-edge defect region. In the third row, the defect targets are very small. EfficientAD shows insufficient responses to these tiny defects, and RD4AD exhibits both scattered responses and local false activations. CSG-Net responds to part of the tiny defect regions, while still presenting an additional local response point in the lower-right area. Overall, CSG-Net shows a relatively more concentrated anomaly response distribution on these samples.

As shown in Fig. 12, the pixel-level confusion matrices of the baseline and CSG-Net(3-s) are presented to further analyze their pixel-wise classification behavior under severe class imbalance. To ensure consistency with the overall evaluation protocol, the confusion matrices are computed using full-pixel statistics on the test set. The operating threshold is determined by maximizing the pixel-level Youden index on the validation set and then fixed for test-set evaluation. Compared with the baseline, CSG-Net reduces the number of false positives from 8,242,979 to 16,830, indicating that it significantly reduces the misclassification of normal pixels as anomalous pixels. Correspondingly, the false-alarm rate decreases from 6.21% to 0.013%, demonstrating that the cascaded semantic-guided mechanism can effectively suppress false responses in background regions. Although the recall of anomalous pixels decreases in this result, the substantial reduction in false positives indicates that CSG-Net exhibits stronger stability in controlling background interference.

E. Ablation Study

To evaluate the effect of key CSG-Net designs on pixel-level anomaly localization performance, we conducted an ablation study on the SCBSA dataset, as summarized in Table II. The experiment uses a standard U-Net discriminative network [12] as the Baseline, and further compares the performance of the basic CSG-Net variant, the two-stage cascaded structure, the semantic-guided enhancement module (SGEM), and deep supervision under different combinations. Here, the basic

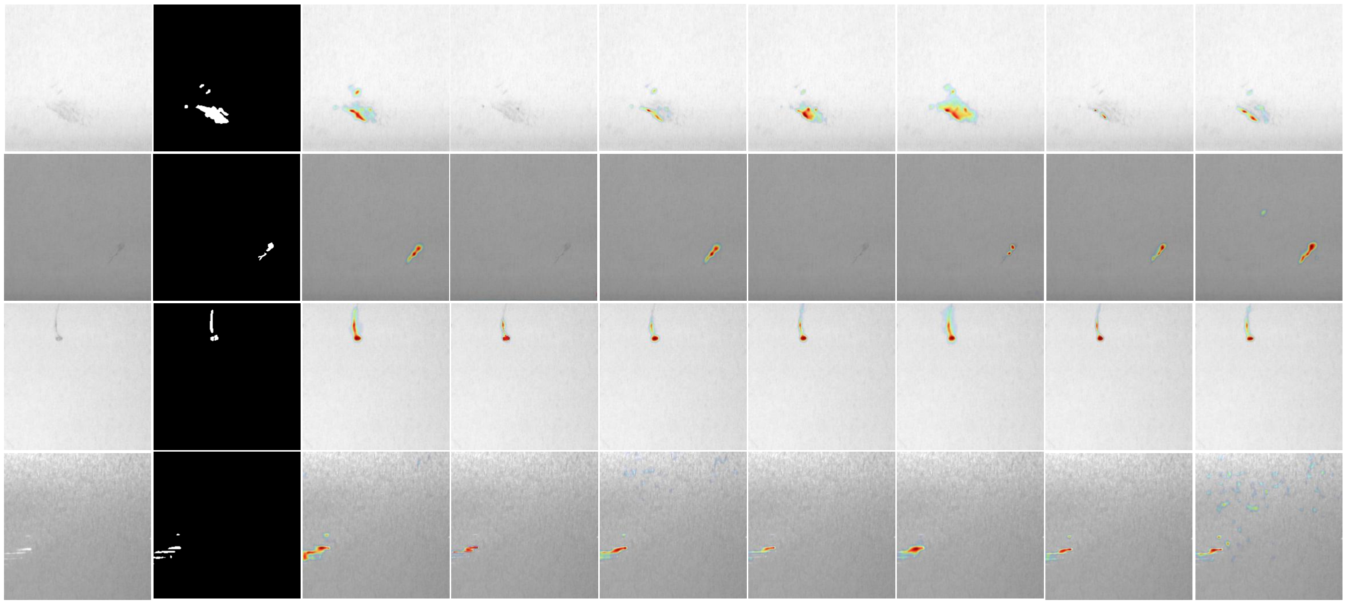


Fig. 14. Qualitative visualization of anomaly localization under different ablation settings. (a) Input image; (b) Ground truth; (c) Full model; (d) Baseline; (e) Deep supervision with uniform weighting (1:1:1:1); (f) Deep supervision applied to both S1 and S2; (g) FiLM-style modulation; (h) SPADE-style modulation; (i) C-BN.

TABLE II
PIXEL-LEVEL AUROC AND AUPRO (%) OF DIFFERENT CSG-NET VARIANTS, WHERE THE CHECK MARK (✓) AND CROSS (×) DENOTE INCLUSION OR EXCLUSION OF THE REFINEMENT STAGE AND SGEM MODULE, RESPECTIVELY.

Variant	Two-Stage	SGEM	Deep Sup	AUROC	AUPRO
Baseline	×	×	×	87.95	76.41
CSG-Net	×	×	×	84.67	69.65
CSG-Net	✓	×	×	91.00	88.20
CSG-Net	✓	✓	×	91.88	79.88
CSG-Net	✓	✓	✓	**93.56**	**88.21**

TABLE III
COMPARISON OF DIFFERENT BACKBONE NETWORKS ON THE SCBSA DATASET IN TERMS OF PERFORMANCE AND EFFICIENCY. PARAMS ARE REPORTED IN M, AND PIXEL-LEVEL AUROC AND AUROC ARE EXPRESSED AS PERCENTAGES (%).

Backbone	Params	GFLOPs	FPS	AUROC	AUPRO
ResNet-18	**98.52**	**374.86**	**34.26**	93.56	**88.21**
ResNet-50	108.09	384.04	33.00	92.97	86.85
ResNet-101	127.08	403.52	30.43	93.53	84.13
WideResNet50	108.09	384.09	32.91	**93.83**	87.11
WideResNet101	127.08	403.60	30.39	93.01	85.77

CSG-Net variant denotes the single-stage model without the two-stage cascaded structure, SGEM, and deep supervision, which is used to analyze the influence of subsequent cascaded refinement and semantic-guided designs. As shown in Table II, the Baseline achieves AUROC and AUPRO values of 87.95% and 76.41%, respectively. The basic CSG-Net variant obtains AUROC and AUPRO values of 84.67% and 69.65%, respectively. After introducing the two-stage cascaded structure into this basic variant, AUROC and AUPRO increase to 91.00% and 88.20%, respectively. With the further incorporation of SGEM, AUROC continues to improve to 91.88%. After introducing deep supervision, the complete CSG-Net achieves the best results on both metrics, with AUROC and AUPRO values of 93.56% and 88.21%, respectively.

1) Backbone for the Discriminative Sub-Network: We further analyzed the impact of different backbone networks on the performance and efficiency of the discriminative model. Under the same decoder configuration, five variants were evaluated using ResNet-18, ResNet-50, ResNet-101 [53], WideResNet-50, and WideResNet-101 [54] as backbones. For each variant, we report the number of parameters, GFLOPs, FPS, and pixel-level localization metrics. As shown in Table III, ResNet-18 has the lowest number of parameters and computational cost, with 98.52M parameters and 374.86 GFLOPs, while achieving the highest FPS of 34.26. In terms of localization performance, WideResNet50 obtains the highest AUROC of 93.83%, whereas ResNet-18 achieves the highest AUPRO of 88.21% and a competitive AUROC of 93.56%. In contrast, ResNet-50, ResNet-101, and WideResNet101 do not provide consistent improvements in both localization metrics, while introducing higher computational overhead and reducing inference speed. Therefore, although deeper or wider backbones generally provide stronger representational capacity, they do not show a clear overall advantage in this task. Considering localization accuracy, region-level overlap performance, and inference efficiency, ResNet-18 is selected as the default backbone for subsequent experiments.

2) Encoder Modulation Mechanisms: The core of encoder modulation is to use semantic guidance to modulate fine-stage

TABLE IV
ABLATION RESULTS OF DIFFERENT SEMANTIC MODULATION STRATEGIES IN THE ENCODER. THE PROPOSED SGEM OUTPERFORMS ELEMENT-WISE PRODUCT, CONCATENATION, AND SE MODULE IN BOTH PIXEL-LEVEL AUROC AND AUPRO (%).

Variant	Upsample	AUROC	AUPRO
element-wise product	Nearest	93.03	87.63
element-wise product	Bilinear	92.00	85.67
concatenation + Conv–BN–ReLU	Nearest	92.80	85.18
concatenation + Conv–BN–ReLU	Bilinear	92.76	86.76
SE-like channel modulation	-	92.19	81.13
SGEM (ours)	Nearest	93.30	87.46
SGEM (ours)	**Bilinear**	**93.56**	**88.21**

TABLE V
COMPARISON WITH REPRESENTATIVE SEMANTIC-CONDITIONED FEATURE MODULATION METHODS. ALL VARIANTS SHARE THE SAME TWO-STAGE BACKBONE AND USE THE SAME STRUCTURAL SEMANTIC PRIOR FOR CONDITIONING. DEEP SUPERVISION IS APPLIED ONLY TO THE FINE STAGE (STAGE-2) FOR ALL VARIANTS. PIXEL-LEVEL AUROC AND AUPRO ARE REPORTED (%).

Modulation	Type	AUROC	AUPRO
FiLM-style modulation [39]	affine (global)	**93.59**	85.27
C-BN [40]	affine (global)	93.53	82.99
SPADE-style modulation [41]	affine (spatial)	92.63	87.67
SGEM (ours)	affine (spatial)	93.56	**88.21**

features for pixel-level anomaly localization. To evaluate the effect of different modulation strategies, we compare four representative variants, including element-wise product, feature concatenation, squeeze-and-excitation (SE)-like channel modulation, and the proposed SGEM, as summarized in Fig. 13. These variants cover direct feature interaction, convolution-based feature fusion, and channel attention mechanisms [53], [55]. As reported in Table IV, different semantic modulation strategies show clear differences in pixel-level localization performance. As a simple feature interaction strategy, element-wise product with nearest upsampling achieves 93.03% AUROC and 87.63% AUPRO, outperforming most alternative variants. In contrast, SE-like channel modulation obtains an AUPRO of only 81.13%, showing relatively weak region-level overlap performance. The proposed SGEM achieves high results under both upsampling strategies, and SGEM with bilinear upsampling performs best, reaching 93.56% AUROC and 88.21% AUPRO.

3) Comparison with Semantic-Conditioned Feature Modulation: We further compare SGEM with representative semantic-conditioned feature modulation designs, including FiLM-style modulation [39], conditional batch normalization (C-BN) [40], and SPADE-style modulation [41]. For a fair comparison, all variants share the same two-stage backbone and use the same structural semantic prior as the conditioning information.

As shown in Table V, different semantic-conditioned modulation methods show different trends in AUROC and AUPRO. FiLM-style modulation achieves the highest AUROC of 93.59%, while SGEM obtains a comparable AUROC of 93.56%. In contrast, SGEM improves AUPRO by 0.54% to 5.22% compared with the other three modulation methods, achieving the highest AUPRO of 88.21%. Fig. 14 presents the qualitative results of different semantic-conditioned modulation methods, where column (c) denotes the complete CSG-Net, and columns (g)–(i) correspond to FiLM-style modulation, SPADE-style modulation, and C-BN, respectively. It can be observed that CSG-Net produces more concentrated anomaly responses that are better aligned with the ground-truth defect regions. Although FiLM-style modulation achieves a high AUROC, it shows incomplete coverage of defect regions in some samples, such as the slender defect in the second row, where its localization is less complete than that of CSG-Net. SPADE-style modulation produces relatively scattered responses in some samples, while C-BN tends to generate weaker or fragmented activations under complex backgrounds.

TABLE VI
ABLATION STUDY ON SUPERVISION STRATEGIES AND LOSS WEIGHTING SCHEMES FOR MULTI-STAGE DEEP SUPERVISION. PIXEL-LEVEL AUROC AND AUPRO SCORES (IN %). FOR LOSS WEIGHTING, WEIGHTS ARE ASSIGNED TO FINE-STAGE FEATURE LEVELS IN THE ORDER OF $p_1^{(2)} \rightarrow p_4^{(2)}$ (FROM HIGH TO LOW SPATIAL RESOLUTION).

Ablation Aspect	Strategy/Weights	AUROC	AUPRO
Supervision Stage	S1 + S2	91.53	85.52
	S2	**93.56**	**88.21**
Weighting Scheme	1 : 1 : 1 : 1	**93.73**	87.67
	1 : 2 : 3 : 4	92.81	88.00
	4 : 3 : 2 : 1	93.56	**88.21**

4) Supervision Strategy and Loss Weighting for Multi-Stage Deep Supervision: To determine the specific setting of multi-stage deep supervision, we compare different supervision stages and loss weighting schemes, as reported in Table VI. Specifically, the supervision-stage experiment compares S1+S2 supervision with S2 supervision, while the loss-weighting experiment compares three multi-scale prediction weighting schemes. As shown in Table VI, applying deep supervision only to S2 achieves AUROC and AUPRO values of 93.56% and 88.21%, respectively, outperforming the S1+S2 supervision setting. This indicates that applying supervision to S1+S2 does not lead to better performance in the current framework. For loss weighting, different schemes show different preferences in AUROC and AUPRO. The equal-weight scheme (1:1:1:1) achieves the highest AUROC of 93.73%, but its AUPRO is lower than that of the 4:3:2:1 setting. In contrast, the 4:3:2:1 weighting scheme achieves the highest AUPRO of 88.21%, while maintaining a competitive AUROC of 93.56%. Considering that this work focuses more on complete defect-region localization, we adopt S2 supervision and the 4:3:2:1 loss weighting scheme as the default setting. The qualitative results in Fig. 14 further show that, with S2 supervision and the 4:3:2:1 weighting scheme, the model produces more complete and concentrated responses to defect regions.

5) Multi-Scale Semantic Output Aggregation Strategies: To compare the aggregation of semantic outputs from different decoder layers, we evaluate different level combinations and fusion strategies, with quantitative results reported in Table VII. For the concatenation strategy, using only the Level 1 output achieves AUROC and AUPRO values of 90.10% and 83.17%, respectively. As more level outputs are included in the aggregation, the results generally improve; when aggregating Level 1–4 outputs, AUROC and AUPRO reach 93.56% and 88.21%, respectively. Under the full-level output aggregation setting, Element-wise Average achieves the highest AUROC and AUPRO, reaching 93.76% and 88.36%, respectively. Concatenation obtains comparable results, with the two metrics being only 0.20% and 0.15% lower, respectively, while Element-wise Addition shows relatively weaker performance. The structural diagrams of different output aggregation strategies are shown in Fig. 13.

6) Multi-Stage Cascade Strategy: To further analyze the setting of the cascaded refinement strategy in CSG-Net, we compare different configurations from 1-stage to 5-stage Cascade, as reported in Table VIII. Specifically, 1-stage Cascade denotes a single-stage setting without coarse-to-fine semantic refinement, 2-stage Cascade introduces one coarse-to-fine semantic-guided refinement stage, and 3-stage to 5-stage Cascade further increase the number of refinement stages based on the 2-stage configuration. As shown in Table VIII, 1-stage Cascade has the lowest number of parameters and computational cost, and achieves the highest FPS of 37.98, but its AUROC and AUPRO are only 84.67% and 69.65%, respectively. After introducing 2-stage Cascade, AUROC and AUPRO increase to 93.56% and 88.21%, respectively, with an FPS of 34.26. When further extending to 3-stage Cascade, AUPRO slightly increases to the highest value of 88.58%, while AUROC decreases to 93.15%, GFLOPs increase to 401.95, and FPS drops to 30.86. When the cascade is further extended to 4-stage and 5-stage configurations, AUROC and AUPRO do not show consistent improvements, while GFLOPs continue to increase and FPS further decreases. Overall, 2-stage Cascade achieves the highest AUROC and near-optimal AUPRO; meanwhile, compared with 3-stage and deeper cascade configurations, 2-stage Cascade has lower computational cost and higher inference speed. Therefore, the ablation experiments adopt 2-stage Cascade as the default cascade configuration.

To further explain the effectiveness of the cascade structure and the SGEM, we supplement the visualization results of SGEM output features under the 3-stage Cascade configuration, as shown in Fig. 15. Specifically, the multi-scale features after SGEM modulation at different cascade stages are directly upsampled to the input image resolution and displayed as heatmaps. From the visualization results, it can be observed that, in the earlier cascade stage, the feature responses can preliminarily cover the defect regions, but still exhibit a certain degree of scattered activations. As the cascade progresses, the SGEM output features gradually focus on the true defect regions, the irrelevant activations in background areas are weakened, and the responses of different-scale features to defect regions become more consistent. This phenomenon indicates that the Cascade architecture can refine anomaly responses through stage-wise refinement, while SGEM uses the semantic priors provided by the previous stage to modulate the features of the current stage, enabling the network to pay more attention to defect-related regions.

TABLE VII
ABLATION OF OUTPUT AGGREGATION STRATEGIES. LEVEL 1 IS THE FINAL DECODER LAYER (HIGHEST RESOLUTION), LEVEL 4 EARLIEST LAYER (LOWEST RESOLUTION, RICHEST SEMANTICS). METRICS IN PERCENT: PIXEL-LEVEL AUROC FOR PER-PIXEL ANOMALY DISCRIMINATION, AND AUPRO FOR REGION-LEVEL LOCALIZATION ACCURACY.

Scheme	Levels 1	2	3	4	AUROC	AUPRO
Concatenation	✓				90.10	83.17
	✓	✓			92.27	79.78
	✓	✓	✓		92.13	85.98
	✓	✓	✓	✓	93.56	88.21
Element-wise Addition	✓	✓	✓	✓	92.72	86.77
Element-wise Average	✓	✓	✓	✓	**93.76**	**88.36**

TABLE VIII
PERFORMANCE AND EFFICIENCY COMPARISON UNDER DIFFERENT STAGE CONFIGURATIONS. PARAMS (M) AND PIXEL-LEVEL AUROC/PROAUC (%) ARE REPORTED. "1-STAGE CASCADE" DENOTES A SINGLE-STAGE SETTING WITHOUT COARSE-TO-FINE SEMANTIC REFINEMENT; "2-STAGE CASCADE" DENOTES COARSE-TO-FINE REFINEMENT WITH SEMANTIC GUIDANCE FROM THE COARSE PREDICTION. K-STAGE CASCADES ($K = 3\sim5$) ITERATIVELY REPEAT THE REFINEMENT STAGE.

Scheme	Params	GFLOPs	FPS	AUROC	AUPRO
1-stage Cascade	**89.64**	**338.97**	**37.98**	84.67	69.65
2-stage Cascade	98.52	374.86	34.26	**93.56**	88.21
3-stage Cascade	98.52	401.95	30.86	93.15	**88.58**
4-stage Cascade	98.52	429.04	27.86	92.88	86.56
5-stage Cascade	98.52	456.13	25.59	93.53	85.13

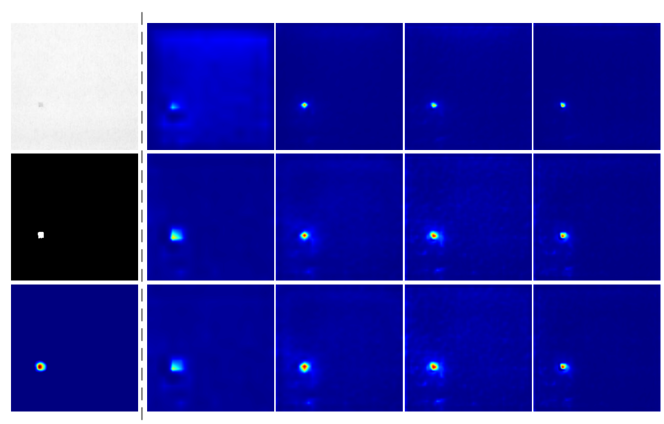


Fig. 15. Visualization of SGEM module output features under the 3-stage Cascade. The left column shows the input image, ground truth, and final prediction from top to bottom. The right part presents the multi-scale SGEM feature responses at different cascade stages, where the rows indicate progressively deeper cascade stages and the columns denote feature levels in the order of $p_1^{(2)} \rightarrow p_4^{(2)}$ from high to low spatial resolution.

VI. EXTENDED VALIDATION ON PUBLIC BENCHMARKS

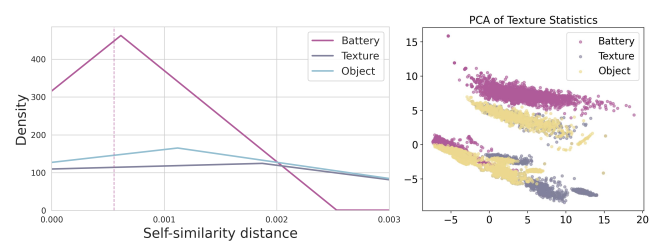
Although the proposed framework is primarily developed for battery steel-shell inspection, all compared methods have been uniformly reproduced and evaluated on the SCBSA dataset, we further conduct an extended validation on two public industrial anomaly detection benchmarks, namely MVTec AD and VisA, to examine its applicability across different industrial contexts. Therefore, the comparison results on MVTec

TABLE IX
IMAGE-LEVEL / PIXEL-LEVEL AUROC (%) ON MVTECAD TEXTURE CATEGORIES FOR ANOMALY DETECTION AND LOCALIZATION. ”-” INDICATES MISSING VALUES. BEST RESULTS IN EACH ROW ARE BOLD.

Category	**BGAD** [11]	**EfficientAD** [49]	**UniNet** [50]	**DRAEM** [12]	**SimpleNet** [51]	**RD4AD** [26]	**SK-RDAD** [52]	**BiSiNet** [13]	**CSG-Net** (Ours)
Grid	72.44/78.83	99.88/91.50	99.50/99.40	99.90/99.70	99.70/99.70	**100.0**/99.00	-/99.30	**100.0**/99.70	**100.0/99.74**
Leather	99.89/99.36	**100.0**/96.64	**100.0**/99.50	**100.0**/98.60	**100.0/100.0**	**100.0**/99.40	-/99.60	**100.0**/99.70	**100.0**/99.80
Tile	94.96/86.48	99.88/88.00	99.50/97.30	99.60/99.20	99.80/**99.80**	99.30/95.60	-/96.10	99.90/96.90	**100.0**/99.69
Carpet	74.82/92.50	99.31/89.65	**100.0**/99.20	97.00/95.50	99.70/**99.70**	98.90/98.90	-/99.20	95.70/97.50	99.20/99.22
Wood	**100.0**/91.36	99.46/85.54	**100.0/99.20**	99.10/96.40	**100.0**/94.50	99.20/95.30	-/95.40	**100.0**/97.10	**100.0**/96.78
Average	88.42/89.71	99.73/90.27	99.80/98.92	99.12/97.88	**99.84**/98.74	99.48/97.64	-/97.92	99.12/98.18	**99.84/99.05**

TABLE X
IMAGE-LEVEL / PIXEL-LEVEL AUROC (%) ON SELECTED CATEGORIES FROM THE VISA DATASET. “-” INDICATES UNAVAILABLE RESULTS. THE BEST RESULT FOR EACH METRIC IN EACH ROW IS SHOWN IN BOLD.

Category	**BGAD** [11]	**EfficientAD** [49]	**UniNet** [50]	**DRAEM** [12]	**SimpleNet** [51]	**RD4AD** [26]	**SK-RDAD** [52]	**BiSiNet** [13]	**CSG-Net** (Ours)
Candle	82.50/89.47	**98.42**/97.60	95.40/99.10	91.88/96.60	84.10/97.60	96.20/99.00	-/98.60	88.00/99.48	90.56/**99.65**
Capsules	68.46/91.04	93.54/96.82	92.20/**99.70**	95.37/98.50	74.10/97.10	90.50/99.40	-/99.10	94.17/98.99	**95.72**/99.55
Cashew	90.62/97.98	97.18/97.35	93.70/96.30	96.58/83.50	88.00/98.90	98.50/95.30	-/98.10	86.50/**98.93**	**98.64**/97.06
Chewinggum	95.49/98.09	**99.88**/97.18	98.90/99.10	98.84/96.80	96.40/97.90	99.30/98.90	-/97.70	96.50/**99.87**	98.44/99.68
Fryum	85.60/96.07	96.55/88.35	96.50/96.20	97.26/87.20	88.40/93.00	97.50/97.00	-/96.70	96.50/**99.81**	**99.02**/99.77
Macaroni1	76.88/96.00	**99.35**/99.54	98.00/99.80	96.15/**99.90**	85.90/98.90	94.60/99.80	-/99.30	95.50/99.26	96.54/99.32
Macaroni2	63.47/95.26	**96.74**/99.30	90.60/99.60	91.97/99.20	68.30/93.20	89.50/99.70	-/99.30	83.50/**99.99**	80.26/**99.99**
Pcb1	84.93/96.37	**98.49**/99.40	97.50/99.70	97.50/88.70	91.60/99.20	97.20/**99.80**	-/99.60	88.00/99.22	96.51/99.73
Pcb2	90.53/96.37	99.50/97.80	97.50/98.30	**99.54**/91.30	92.40/96.60	96.80/99.00	-/98.30	91.00/98.45	98.33/**99.01**
Pcb3	79.98/95.97	**98.91**/97.88	96.60/99.10	96.70/98.00	89.10/97.20	97.00/99.30	-/98.30	90.03/99.83	98.27/**99.86**
Pcb4	95.56/95.82	98.86/96.84	99.80/98.90	99.51/96.80	97.00/93.90	**99.90**/98.70	-/98.60	95.53/**99.66**	98.74/99.46
Pipe Fryum	81.03/99.32	99.46/97.60	**99.70**/99.20	99.09/85.80	90.80/98.50	93.50/99.10	-/99.10	93.50/**99.66**	99.24/99.59
Average	82.13/95.65	98.09/97.14	96.37/98.75	**99.12**/93.52	87.20/96.80	96.42/98.75	-/98.50	91.56/**99.43**	95.86/99.40



(a) (b)

Fig. 16. Statistical comparison across three domains: Battery, Texture, and Object. (a) KDE of self-similarity distances, with the dashed line indicating the mean distance of the Battery domain, suggesting higher intra-image redundancy. (b) PCA visualization of texture statistics, where battery samples form a distinct cluster, indicating clear domain discrepancy.

AD and VisA are adopted from the values reported in the original papers, and unavailable results are denoted by “-” in the tables. MVTec AD contains 15 categories spanning both texture and object types, with diverse defect patterns such as scratches, cracks, and contaminations. Since battery inspection typically involves homogeneous metallic surfaces with localized structural deviations, we focus our detailed evaluation on the texture categories, which are more consistent with the visual characteristics of battery steel shells. The PCA-based feature visualization and self-similarity analysis in Fig. 16 provide supporting evidence for this scenario relevance.

Quantitative results on the MVTec AD texture categories are summarized in Table IX. CSG-Net with the 3-stage Cascade configuration achieves an average image-level AUROC of 99.84%, matching the best result, and an average pixel-level AUROC of 99.05%, which is the highest among all compared methods. These results indicate that CSG-Net also shows strong anomaly detection and pixel-level localization performance on representative public texture anomaly detection scenarios. We further evaluate the proposed method on selected categories from the VisA dataset. As reported in Table X, CSG-Net achieves an average image-level AUROC of 95.86% and a pixel-level AUROC of 99.40%, maintaining competitive performance across different categories. Although CSG-Net does not obtain the top score on every metric, it shows strong pixel-level localization performance on several categories. For example, on Macaroni2, CSG-Net achieves a pixel-level AUROC of 99.99%, while its image-level AUROC is lower than that of EfficientAD. In addition, BiSiNet slightly exceeds CSG-Net in average pixel-level AUROC, indicating that the localization performance of CSG-Net on some public categories could still be further improved. Representative qualitative comparisons on MVTec AD and VisA are shown in Fig. 17. On texture-dominant MVTec AD samples, CSG-Net produces compact and well-aligned activation maps that closely match the ground-truth defect regions while reducing irrelevant background responses. Compared with several reconstruction-based and feature-distance-based baselines, the

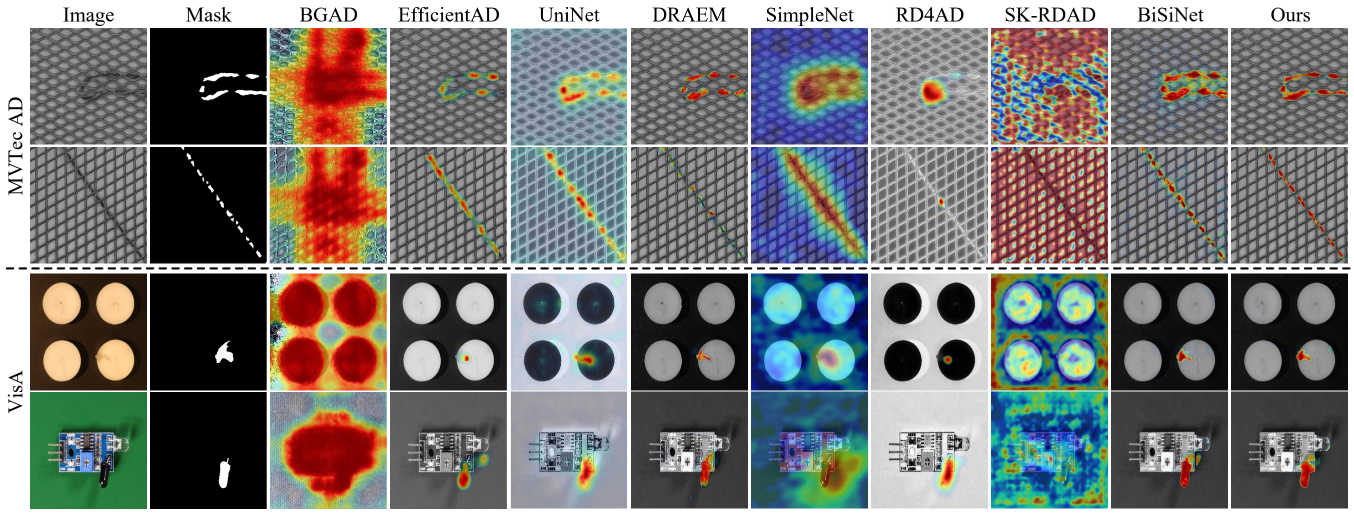


Fig. 17. Qualitative results on MVTec AD and VisA. The upper two rows are derived from MVTec AD, whereas the lower two rows are derived from VisA.

proposed method usually yields clearer boundary localization and fewer scattered false positives. For VisA samples, CSG-Net still maintains concentrated anomaly responses and reduces excessive activation over structurally repetitive regions. In categories such as Pcb3 and Candle, the predicted anomaly regions remain focused on the true defect areas, whereas some competing methods tend to over-expand to surrounding textures or miss subtle defect regions.

VII. Conclusion

This work presents CSG-Net, a cascaded coarse-to-fine framework for surface anomaly localization in steel-cased batteries. The coarse stage combines reconstruction–discrimination sub-network to model normal appearance and provide reliable initial cues, while the fine stage introduces SGEM to adaptively modulate multi-scale semantics for refined localization. Experiments on the battery dataset, together with extended validation on public benchmarks, demonstrate that CSG-Net achieves accurate and reliable defect localization. Despite the strong localization performance, cascaded refinement may not always translate highly localized evidence into equally stable global decision signals in certain challenging cases, reflecting a remaining global–local coupling issue. Future work will investigate cross-stage semantic consistency and alignment constraints to better couple localized evidence with more stable global representations and improve decision reliability. In addition, the cascaded design increases computational and memory cost compared with single-stage alternatives. We will explore parameter-efficient refinement and stage-wise feature sharing to reduce the memory footprint while preserving localization accuracy.

References

[1] G. Ablay, "Online condition monitoring of battery systems with a nonlinear estimator," *IEEE Trans. Energy Convers.*, vol. 29, no. 1, pp. 232–239, 2014.

[2] B. G. Carkhuff, P. A. Demirev, and R. Srinivasan, "Impedance-based battery management system for safety monitoring of lithium-ion batteries," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6497–6504, 2018.

[3] M. Cai, E. Zhang, J. Lin, K. Wang, K. Jiang, and M. Zhou, "Route optimization equalization scheme based on graph theory for liquid metal battery strings," *IEEE Trans. Ind. Appl.*, vol. 59, no. 2, pp. 2502–2508, 2023.

[4] M. A. Reyes and T. Novak, "Injury surveillance and safety considerations for large-format lead-acid batteries used in mining applications," *IEEE Trans. Ind. Appl.*, vol. 52, no. 2, pp. 1925–1930, 2016.

[5] K. Wu, J. Tan, and C. Liu, "Cross-domain few-shot learning approach for lithium-ion battery surface defects classification using an improved siamese network," *IEEE Sens. J.*, vol. 22, no. 12, pp. 11 847–11 856, Jun. 2022.

[6] F. Carere, A. Sardellitti, A. Bernieri, L. Ferrigno, S. Sangiovanni, and M. Laracca, "An eddy current probe for the detection of subsuperficial defects of any orientation," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[7] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "PGA-net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7448–7458, 2020.

[8] Y. Gao, X. Luo, X. Gao, W. Yan, X. Pan, and X. Fu, "Semantic segmentation of remote sensing images based on multiscale features and global information modeling," *Expert Syst. Appl.*, vol. 249, p. 123616, 2024.

[9] S. Qu, X. Tao, X. Gong, Z. Qu, M. Prasad, F. Shen, Z. Zhang, and G. Ding, "Lscad: A large–small model collaboration framework for unsupervised industrial anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–14, 2025.

[10] M. S. Sadabadi, X. Meng, and Z. Liu, "Resilient and robust voltage regulation in shipboard dc microgrids with zip loads under actuator and parameter uncertainties," *IEEE Trans. Transport. Electrific.*, vol. 12, no. 1, pp. 578–588, 2026.

[11] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang, "Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 24 490–24 499.

[12] V. Zavrtanik, M. Kristan, and D. Skočaj, "DRÆM – a discriminatively trained reconstruction embedding for surface anomaly detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021, pp. 8310–8319.

[13] Y. Zhai, W. Pan, Y. Liang, H. Zhu, Z. Long, P. Coscia, A. Genovese, V. Piuri, and F. Scotti, "Bidirectional feature pyramid siamese anomaly detection network with cellular anomaly generation for container marking," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–17, 2025.

[14] Y. Zhu, J. Wang, J. Zhang, and Q. Li, "A two-stage autoencoder for visual anomaly detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2021, pp. 1869–1873.

[15] J. Liu, "A two-stage anomaly detection framework: Towards low omission rate in industrial vision applications," *Adv. Eng. Inf.*, vol. 55, p. 101822, 2023.

[16] G. Tong, Q. Li, and Y. Song, "Two-stage reverse knowledge distillation incorporated and self-supervised masking strategy for industrial anomaly detection," *Knowl.-Based Syst.*, vol. 273, p. 110611, 2023.

[17] Y. Liang, Z. Hu, J. Huang, D. Di, A. Su, and L. Fan, "Tocoad: Two-stage contrastive learning for industrial anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[18] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.

[19] K. Zhang, L. Liao, and Y. Wang, "Discrete wavelet transform-based detection transformer for battery weld defect detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[20] H. Zhu, J. You, Y. Zhai, Y. Xu, T. Wang, Y. Chen, J. Zhou, P. Coscia, A. Genovese, and C. L. P. Chen, "DS^2 A-Former: Battery Surface Defect Segmentation via Dual-Stream Spatial Attention Transformer Network," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–17, 2025.

[21] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but differnet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 1906–1915.

[22] Y. Zhao, X. Wu, and J. Cheng, "Dualflow: Dual-branch flow for unsupervised anomaly detection and localization," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–14, 2025.

[23] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 14 298–14 308.

[24] Q. Wan, L. Gao, X. Li, and L. Wen, "Unsupervised image anomaly detection and segmentation based on pretrained feature mapping," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2330–2339, 2023.

[25] C.-C. Tsai, T.-H. Wu, and S.-H. Lai, "Multi-scale patch-based representation learning for image anomaly detection and segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 3065–3073.

[26] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 9727–9736.

[27] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 1819–1828.

[28] S. Zhuang, H. Zhang, D. Liang, H. Liu, and Z. Gao, "Adaptive sequential bayesian iterative learning for myocardial motion estimation on cardiac image sequences," *IEEE Trans. Med. Imaging*, vol. 45, no. 1, pp. 406–420, 2026.

[29] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[30] R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, and L. Steels, "Generalization and network design strategies," in *Connectionism in Perspective*. Elsevier, 1989, pp. 143–155.

[31] M. Rudolph, B. Wandt, and B. Rosenhahn, "Structuring autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2019, pp. 615–623.

[32] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. V. D. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1705–1714.

[33] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou, "Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8771–8780.

[34] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, 2021, pp. 1–6.

[35] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 4, 2021, pp. 3110–3118.

[36] C. Lv, F. Shen, Z. Zhang, D. Xu, and Y. He, "A novel pixel-wise defect inspection method based on stable background reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[37] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[38] Y. Peng, C. Liu, Y. Yan, N. Ma, D. Wang, C. Liu, and Q. Chen, "Semi-supervised bolt anomaly detection based on local feature reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

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

[40] H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 6594–6604.

[41] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2332–2341.

[42] F. Xue, I. Budvytis, and R. Cipolla, "SFD2: Semantic-guided feature detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5206–5216.

[43] Q. Lin, S. Chang, J. Mao, Q. Liu, Z. Liu, Y. Lu, and Y. Huo, "Anomaly detection in smart IoT systems based on contextual semantics of behavior graphs," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 13 683–13 696, 2025.

[44] H. Gao, H. Luo, F. Shen, and Z. Zhang, "A norm-constrained method for semantic and structured anomaly detection tasks," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[45] K. Wang, S. Lyu, B. Wang, and Y. Zhang, "Anomaly detection via semantically conjugate view learning on industrial temporal data," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 15 479–15 490, 2025.

[46] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3606–3613.

[47] S. Xie and Z. Tu, "Holistically-nested edge detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.

[48] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille, "Deep supervision for pancreatic cyst segmentation in abdominal ct scans," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, 2017, pp. 222–230.

[49] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*. Waikoloa, HI, USA: IEEE, 2024, pp. 127–137.

[50] S. Wei, J. Jiang, and X. Xu, "Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 9994–10 003.

[51] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A simple network for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 20 402–20 411.

[52] E. Park, T. Kim, M. Kim, H. Lee, and G.-J. Lee, "Sk-rd4ad: Skip-connected reverse distillation for robust one-class anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2025, pp. 3945–3953.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[54] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2016, pp. 87.1–87.12.

[55] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.

Ying Xu received the BS. and MS. degrees in automation, and control theory and control engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2004 and 2008, respectively, and the Ph.D. degree in control theory and control engineering from the South China University of Science and Engineering, Guangzhou, China, in 2013.

She is currently an Associate Professor with the School of information and communication engineering, Wuyi University, Jiangmen, Guangdong, China. Her current research interests include image processing, deep learning, and biometric extraction.

Dongyan Lin received the B.S. degree in Software Engineering from Guangdong Baiyun University in 2023. She is currently pursuing the M.S. degree with the School of Electronics and Information Engineering, Wuyi University.

Her research interests include computer vision, pattern recognition, and image processing.

Tianlei Wang received the B.S.E. degree in electrical engineering (academic training) from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, the M.S. degree in signal processing from the South China University of Technology, Guangzhou, China, in 2007, and the Ph.D. degree in safety engineering from Beijing Jiaotong University, Beijing, in 2022.

His research interests include the intelligent systems, control theory, and pattern recognition.

Yingwen Chen received the B.S.degree in Big Data Management and Application from Neusoft Institute Guangdong in 2023. He is currently working toward the M.S. degree with the School of Electronics and Information Engineering, Wuyi University.

His research interests include computer vision, pattern recognition, and image processing.

Yikui Zhai (Senior Member, IEEE) received his Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013.

Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is a Full Professor now. He is also Associate Dean of the School of Electronics and Information Engineering in Wuyi University, since 2021. He has been a Visiting Scholar with Department of Computer Science, the Universita degli Studi di Milano, Italy, during June 2016 to June 2017, August 2023 and January 2024. His research interests include: Image Processing, Deep Learning, Optical Character Recognition, Object Detection, UAV Change Detection, Self Supervise Learning.

Jianhong Zhou (Member, IEEE) is currently pursuing the Ph.D. degree in the School of Electronics and Information Engineering, South China University of Technology. He received the B.S. and MS. degree in Wuyi University of Communication Engineering and Information and Communication Engineering in 2020 and 2024.

His research interests include computer vision, representational contrastive learning, image processing and object detection.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Italy, in 2019.

From 2019 to 2022, he was a post-doctoral researcher at the Universita degli Studi di Padova, Italy. He is currently an Assistant Professor at the Department of Computer Science of the Universita degli Studi di Milano, Italy. He is Co-chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023. His research activities are focused on theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milan, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Universita degli Studi di Milan since 2014.

He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters.

His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Dr. Genovese is an Associate Editor for the Journal of Ambient Intelligence and Humanized Computing (Springer).

C. L. Philip Chen (Life Fellow, IEEE) received the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1988, both in electrical and computer science.

He is the Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is the former Dean of the Faculty of Science and Technology, University of Macau, Macau, China. His research interests include cybernetics, computational intelligence, and systems.

Dr. Chen is a Fellow of the IEEE, AAAS, IAPR, CAA, and HKIE; and a member of Academia Europaea (AE) and the European Academy of Sciences and Arts (EASA). He received the IEEE Norbert Wiener Award in 2018 for his contributions in systems, cybernetics, and machine learning. He was also recognized as a Highly Cited Researcher by Clarivate Analytics from 2018 to 2021.He served as the Editor-in-Chief of IEEE Transactions on Cybernetics from 2020 to 2021, after completing his term as Editor-in-Chief of IEEE Transactions on Systems, Man, and Cybernetics: Systems from 2014 to 2019. He was President of the IEEE Systems, Man, and Cybernetics Society from 2012 to 2013. Currently, he serves as a Deputy Director of CAAI Transactions on AI, and as an Associate Editor of IEEE Transactions on AI, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Fuzzy Systems, and China Sciences: Information Sciences. He has received the Macau FDCT Natural Science Award three times and the First-Rank Guangdong Province Science and Technology Advancement Award in 2019.