Prioritized Local Matching Network for Cross-Category Few-Shot Anomaly Detection

Huilin Deng, Hongchen Luo, Wei Zhai, Yanming Guo, Yang Cao, *Member, IEEE*, and Yu Kang, *Senior Member, IEEE*

***Abstract*—In response to the rapid evolution of products in industrial inspection, this article introduces the cross-category few-shot anomaly detection (C-FSAD) task, aimed at efficiently detecting anomalies in new object categories with minimal normal samples. However, the diversity of defects and significant visual distinctions among various objects hinder the identification of anomalous regions. To tackle this, we adopt a pairwise comparison between query and normal samples, establishing an intimate correlation through fine-grained correspondence. Specifically, we propose the prioritized local matching network (PLMNet), emphasizing local analysis of correlation, which includes three primary components: 1) Local perception network refines the initial matches through bidirectional local analysis; 2) step aggregation strategy employs multiple stages of local convolutional pooling to aggregate local insights; and 3) defect-sensitive Weight Learner adaptively enhances channels informative for defect structures, ensuring more discriminative representations of encoded context. Our PLMNet deepens the interpretation of correlations, from geometric cues to semantics, efficiently extracting discrepancies in feature space. Extensive experiments on two standard industrial anomaly detection benchmarks demonstrate our state-of-the-art performance in both detection and localization, with margins of 9.8% and 5.4%, respectively.**

***Impact Statement*—Anomaly detection (AD) plays an indispensable role in industrial inspection. The recent rapid evolution of products necessitates the swift adaptation of AD models. While existing AD research demands reestimation for novel object categories, this article introduces a local-priority comparison framework, enabling direct AD for novel categories through minimal normal samples, without reestimation or fine-tuning. With a significant increase of 9.8% in detection and 5.4% in localization performance against previous state-of-the-art methods, this technology is ready to support fast-paced industrial inspection by offering both minimal sample dependency and high recognition efficiency.**

Manuscript received 18 August 2023; revised 2 February 2024; accepted 1 April 2024. Date of publication 5 April 2024; date of current version 10 September 2024. This work was supported by the National Natural Science Foundation of China under Grant 62033012. This article was recommended for publication by Associate Editor Thanos Vasilakos upon evaluation of the reviewers' comments. *(Corresponding author: Yang Cao.)*

Huilin Deng, Hongchen Luo, and Wei Zhai are with the University of Science and Technology of China, Anhui, Hefei 230026, China (e-mail: huilin_deng@mail.ustc.edu.cn; lhc12@mail.ustc.edu.cn; wzhai056@.ustc.edu.cn).

Yanming Guo is with the National University of Defense Technology, Hunan, Changsha 410073, China (e-mail: guoyanming@nudt.edu.cn).

Yang Cao and Yu Kang are with the School of Information Science and Technology, University of Science and Technology of China, Anhui, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center Anhui, Hefei 230088, China (e-mail: forrest@ustc.edu.cn; kangduyu@ustc.edu.cn).

Digital Object Identifier 10.1109/TAI.2024.3385743

***Index Terms*—Anomaly detection (AD), cross-category, few-shot learning, visual correspondence.**

I. Introduction

ANOMALY detection (AD), aiming to identify exceptional samples deviating from the expected patterns, boasts wide-ranging applications across numerous domains, such as industrial inspection [1], medical image analysis [2], and autonomous driving [3]. Recent developments indicate an accelerated iteration in industrial products, where anomaly detection models frequently process unseen objects from multiple classes. There is a necessity for a unified anomaly detection model that not only accommodates various known categories but is also adaptable to previously unseen categories.

The unsupervised anomaly detection [5], [8], [9] trains a model with massive normal samples for a specific object category. Though exhibit commendable performance, their requisite for numerous normal samples to estimate normal distribution proves impractical in real-world production [Fig. 1(a)]. Accordingly, few-shot anomaly detection (FSAD) methods [6], [7], [10] have been proposed, training with fewer normal samples. To reduce the demand for extensive training samples, they either mine the feature commonalities [6] or employ data augmentation techniques [10] to construct more efficient estimators. Nevertheless, they still follow the *one-model-per-category* paradigm, training a dedicated model for each category. As a result, they cannot detect novel categories and fail to handle diverse categories with a unified model [Fig. 1(b)]. However, in real-world application scenarios like flexible industrial production, anomaly detection models frequently process unseen (novel) objects from multiple classes. There is a need for a common anomaly detection model shared among multiple categories and also generalizable to novel categories.

To address the dual challenges, we propose the cross-category few-shot anomaly detection (C-FSAD) task. C-FSAD follows the *one-model-multiple-categories* paradigm, handling diverse categories with a unified model and facilitating the rapid adaptation to unseen categories. As illustrated in Fig. 1(c), once trained on multiple seen categories, the C-FSAD model can directly perform anomaly detection on unseen object classes. The training set consists of an abundance of normal samples and a minimal number of anomalous samples, under supervised conditions. Only several normal samples from novel classes are provided during the test, and defects are detected by directly

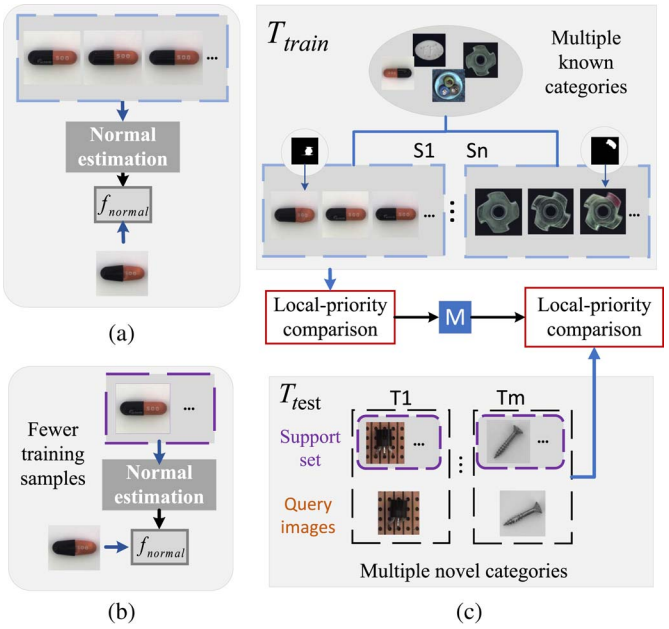

Fig. 1. Task definition. Contrasting with (a) unsupervised [4], [5] and (b) exiting few-shot anomaly detection (FSAD) methods [6], [7] that only confined to trained classes, cross-category few-shot anomaly detection (C-FSAD) broadens anomaly detection to unseen classes with minimal normal samples, instead of retraining (reestimation) or fine-tuning. "m" represents parameter memory of the model. The local-priority comparison are detailed in Fig. 2(b). (c) C-FSAD.

comparing these samples with the test samples. In this context, the diversity of defects and significant visual distinctions between seen and unseen object categories hinder the explicit identification of anomalous regions.

The inherent ambiguity of anomalies results in significant diversity even within the same object category [e.g., different textures, colors, and shapes in Fig. 2(a)]. Yet, the only commonality is their departure from patterns defined as normal within their respective categories. Most existing FSAD methods [6], [7], [11] compare query samples to the normal distribution, leading to unintended loss of pixel information. This can compromise sample efficiency, essential in few-shot settings. To improve sample efficiency, we adopt a direct pairwise comparison between query and normal samples. Leveraging fine-grained correspondence, we establish pixel-level dense correlations between the samples. However, the stability of such pairwise comparison is contested when extended to novel object categories, especially those with pronounced deviations from known classes. As depicted in Fig. 2(a), while scratches on fabric are pronounced, discerning scratches on a hazelnut is nontrivial. This underscores the complexity of pairwise comparison in intricate feature spaces.

Humans can effortlessly discern differences between images. This skill is attributed to *W*eak *C*entral *C*oherence (*WCC*) [12], [13], [14], a cognitive style favoring local detail processing over a holistic view, fostering a heightened sensitivity and deeper understanding of the information. This offers valuable insights for our C-FSAD research. Primarily, a local-centric comparison can suppress global, category-specific information, facilitating cross-category generalization. Moreover, delving deep into local nuances enhances the discernment of essential variances like texture and color shifts, crucial for anomaly detection. Inspired by this, we propose the prioritized local matching network (PLMNet) for C-FSAD, which emphasizes local analysis of correlation to deeply mine discrepancies in feature space.

Specifically, PLMNet, structured under the Siamese network framework, comprises three primary components: LP network, SA strategy, and DSWL. The local-priority comparison mechanism, consisting of LP and SA, operates in two stages as illustrated in Fig. 2(b). Based on initial correlation, the LP network performs bidirectional local analysis to enhance visual cues and enrich the comprehension of intricate features. Subsequently, the SA strategy incrementally aggregates these refined insights through local convolutional pooling, preserving the crucial insights while reducing redundancy. This two-stage process progressively deepens the interpretation of correlation, from geometric cues to semantics. Recognizing the significance of defect-related insights for anomaly detection [15], [16], we incorporate limited defective samples into the training. The DSWL adjusts the final encoded features for more discriminative representations leveraging structural characteristics of defects. To evaluate the proposed PLMNet, we conduct experiments on two standard industrial anomaly detection benchmarks: MVTec AD and MPDD, and achieve state-of-the-art performance.

To summarize, our key contributions are as follows.

1) We introduce a new challenging C-FSAD task along with a large-scale benchmark to support research in fast-paced industrial inspection by performing anomaly detection on unseen categories without fine-tuning.
2) We propose a novel PLMNet that can discern ambiguous anomaly regions through a local-priority comparison, resulting in a good adaptation capability to efficiently perform anomaly detection on multiple unseen categories.
3) Experiments on the MVTec-AD and MPDD datasets show that our PLMNet outperforms the state-of-the-art models and can serve as a strong baseline for future research.

II. Related Work

A. Anomaly Detection

Existing research mainly includes unsupervised anomaly detection and semisupervised anomaly detection. However, the diversity of real-world anomalies complicates the collection of comprehensive anomalous samples. As a result, unsupervised methods have become the mainstream approach. Specifically, one-class classification methods [17], [18] map the extracted features to hyperspherical embeddings. In contrast, flow-based methods [5], [19] transform the normal feature into Gaussian distribution with normalizing flow (NF). On the other hand, reconstruction-based methods [9] typically reconstruct the data, assuming that anomalous samples have higher reconstruction errors.

Nevertheless, learning solely from normal samples often fails to adequately differentiate between normal and complex anomalies, leading to high false positives/negatives. This limitation stems from an insufficient understanding of actual

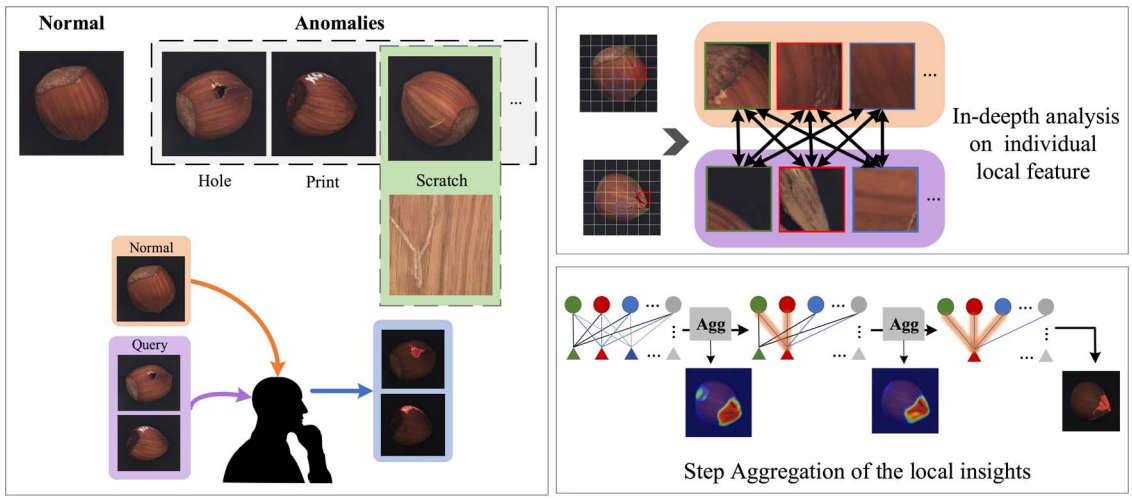

Fig. 2. Motivation of PLMNet. (a) To address the diverse nature of defects, a pairwise comparison is adopted. Motivated by human visual cognition, we further propose the local-priority comparison framework. As illustrated in (b), local-priority comparison operates in two stages: 1) local analysis deepens critical visual cues through in-depth analysis of fine-grained details; 2) step aggregation (SA) progressively aggregates these local insights into a cohesive perception, minimizing redundancy while preserving essential matches.

anomalies [15], [20], [21]. Semisupervised anomaly detection (SSAD) methods [16], [20] advocate incorporating minimal anomaly examples into training, providing essential partial knowledge for anomaly detection despite not encompassing all anomaly classes. DevNet [15] enables end-to-end anomaly score learning via neural deviation learning, with the prior probability of labeled anomalies ensuring statistical deviations. On the other hand, DRA [16] proposes a unique categorization of anomalies and utilizes a multihead neural network to learn disentangled representations. Our work introduces the C-FSAD model to overcome the extensive retraining limitations of current AD methods for new categories. C-FSAD effectively performs anomaly detection with just a single normal sample per category, ensuring rapid adaptation without additional fine-tuning.

B. Few-Shot Segmentation

Few-shot semantic segmentation (FSS) [22], [23] aims at segmenting the foreground of the unseen class with a few annotated samples. Generally, mainstream models can be divided into prototype-based and matching-based methods. Prototype-based methods leverage the prototypes, a semantical representation for the support, to guide the mask prediction. PL [24] introduces prototypical learning into the segmentation task for the first time. Building upon this, PANet [25] proposes a prototype alignment technique to enhance the prototype. However, a single prototype vector may cause semantic ambiguity. PMMs [22] proposes to generate multiple prototype vectors through an EM algorithm. On the contrary, matching-based methods exploited pixel-level features of support images to guide the prediction. CyCTR [26] leverages a transformer for mining the information of support images. HSnet [27] applies efficient 4-D convolution to analyze deeply accumulated features. The setting of C-FSAD and FSS both fall under the paired comparison between support and query samples for novel class prediction, but with a focus on defects and objects, respectively. Locating defective areas requires further exploration of feature correlation, so our basic framework follows the matching-based methods.

C. FSAD

Recent years have witnessed the rapid development of FSAD, which mainly focuses on unsupervised learning (e.g., Patchcore [28], Graphcore [6], and RegAD [11]). Patchcore [28] introduces a memory bank method as the feasible solution for the FSAD issue, while Graphcore [6] enhances the memory bank by incorporating rotation-invariant feature properties. FewSOME [7] adopts a different approach, using a Siamese-like architecture to distill shared features from nominal data. Furthermore, Santos [29] identifies overfitting in current model hyperparameters, indicating the potential for improved performance through optimization. However, these works primarily target AD in trained categories, disregarding the AD in unseen object categories. RegAD [11] marks the first effort at anomaly detection in novel object categories, employing registration, a proxy task to amplify the cosine similarity of features within identical categories. It is notable that RegAD, while effective, can handle one category at a time. In this article, our PLMNet's capability to handle diverse objects addresses the limitation of recent FSAD methods that struggle with multiple object classes, especially those deviating significantly from known classes.

D. Multiclass Novelty Detection

Novelty Detection aims to identify samples from new, unseen classes, where the challenge lies in discerning clear interclass differences ([30], [31], [32]). Multiclass novelty detection, known as multiple inlier single outlier problems (MISO). Multiple classes are considered as inliers and only a single class is

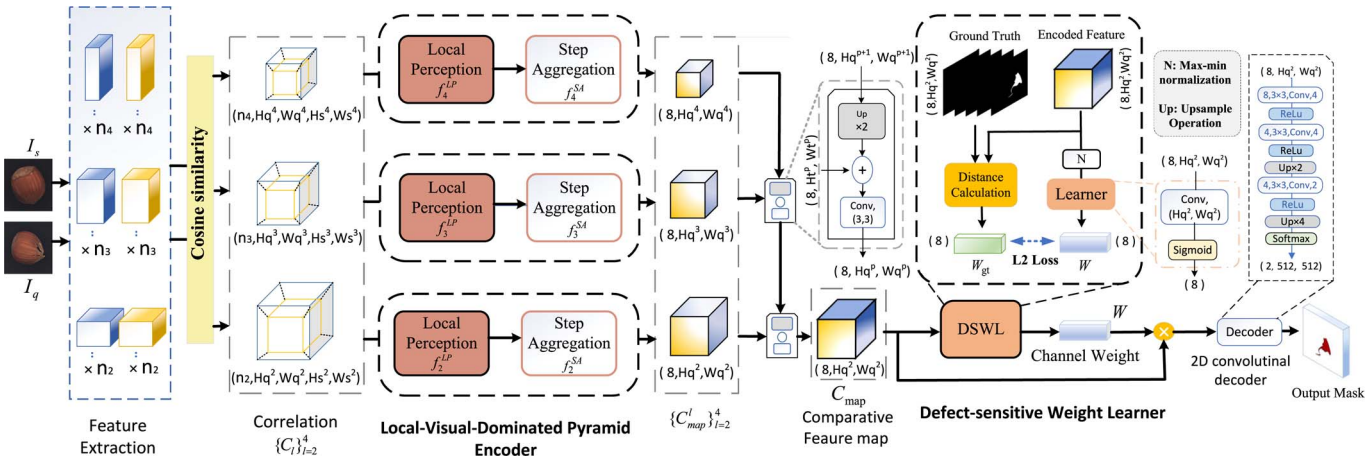

Fig. 3. Overview of the prioritized local matching network. It mainly consists of a LP network (Section III-D), a SA strategy (Section III-E), and a DSWL (Section III-F).

deemed novel. GANomaly [33] focuses on high-dimensional image generation and latent space inference using conditional generative adversarial networks. Deep-embed [34] proposes a novel framework that combines discriminative and generative embeddings to compute novelty scores. On the other hand, NLN [32] leverages the reconstruction error and latent-neighbor distances of nearest neighbors in the latent space to improve the performance of autoencoder models. The key challenge of MISO is to distinguish between multiple inlier classes and a single outlier class. Conversely, the proposed C-FSAD requires a model to discern subtle defect features across multiple novel categories, emphasizing both adaptability and generalization.

III. Method

A. Problem Setup

C-FSAD aims to perform anomaly detection on multiple novel categories. It involves training on C_{train} classes and testing on a distinct test set comprising C_{test} classes. Training sets and testing sets are disjoint to object classes, denoted respectively as D_{train} and D_{test}. Data is fed into the network in pairs, consisting of two sets: the query set $Q = (I_q, M^q)$, where M^q represents the query mask and the support set (I^s). During training, given the query and support set (I_q, I_s) from D_{train}, the model learns the differences between the support and test image to map the mask M^q. During testing, it uses the learned model for evaluation without further optimization.

B. Overall Architecture

As depicted in Fig. 3, our PLMNet comprises three primary components: the LP network (Section III-D), the SA strategy (Section III-E), and the DSWL (Section III-F). Initially, we compute the multilayer correlations between the input pairs (I_q, I_s). Subsequently, the LP network and SA strategy jointly perform the local-priority comparison process on correlation tensors to produce the encoded context. Each pipeline is demonstrated in one-shot setting. The loss function for PLMNet is further elaborated upon in Section III-G, and in Section III-H, we explain how the model can be easily extended to the K-shot setting.

C. Correlation Construction

The dense connections between support and query samples provide initial feature similarity measurement, serving as the preparation for local comparison analysis. Inspired by semantic matching approaches, we leverage intermediate layers of the backbone network, rich in feature representations, to construct multilevel feature correlations. The backbone network consists of L semantic layers, and the lth semantic layer contains n_l layers of intermediate features. Specifically, input a pair of support image and test image (I_q, I_s) to the backbone network, $N = \sum_{l=2}^{4} n_l$ pairs of intermediate feature maps in 2nd, 3rd, and 4th semantic layers are extracted. We further calculate the cosine similarity of the obtained feature $\{F_q^i, F_s^i\}_{i=1}^{N}$ is to construct multilevel feature correlations $\{C^i\}_{i=1}^{N}$, $C^i \in \mathbb{R}^{(H^i, W^i, H^i, W^i)}$. A pair of query and support features at each intermediate layer forms a 4-D correlation tensor $C^i \in \mathbb{R}^{(H^i, W^i, H^i, W^i)}$ using cosine similarity

$$C^i(x_q, x_s) = \text{Relu}\left(\frac{F_t^i(x) \cdot \hat{F}_s^i(x)}{||F_t^i(x)||\,||\hat{F}_s^i(x)||}\right) \tag{1}$$

where x_q and x_s represent the spatial locations of the feature maps F_q^i and F_s^i, respectively. ReLU is used to suppress matching noise. The correlations from the same semantic layer are collected to form the correlations of lth semantic layer C_l. We denote the set of intermediate layers indices in the lth semantic layer as θ_l, i.e, $\theta_l \subseteq \{1, 2, \ldots, N\}, |\theta_l| = n_l$. $\{C^i\}_{i \in \theta_l}$ are concatenated along channel dimension to form the correlations of lth semantic layer $C_l \in R^{(n_l, H^l, W^l, H^l, W^l)}$. Eventually, the multilevel feature correlations are constructed and denoted as $C = \{C_l\}_{l=2}^{4}$.

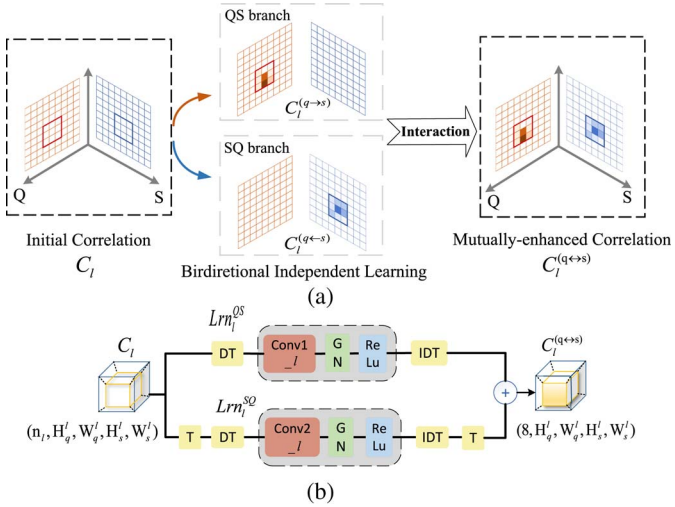

Fig. 4. (a) Two branches, performing local analysis from distinct perspectives, complement and reinforce each other to enrich the comprehension of visual information. (b) Building blocks in LP network: $\{f_l^{\mathrm{LP}}\}_{l=2}^{4}$. Notably, conv1_$l$ and conv2_l share identical structures. "GN" denotes group normalization with four groups.

D. LP Network

To enhance local-level perception, the LP network employs a dual-branch structure, performing bidirectional feature refinement and subsequent interaction of correlation tensors. This bidirectional refinement ensures that features are comprehensively compared from both perspectives. As illustrated in Fig. 4(a), each *Q*uery to *S*upport (QS) and *S*upport to *Q*uery (SQ) branch operates independently, with their outputs mutually interacting for a holistic and in-depth understanding.

1) The QS Branch: The QS branch primarily consists of a local relearning block: $\mathrm{Lrn}^{\mathrm{QS}}$, along with two Dimension Transformation operations: f^{DT} and f^{IDT}, which are inverse operations mutually. The Local Relearning block $\mathrm{Lrn}^{\mathrm{QS}}$ serves to refine the correlation based on the local receptive field, and the operations f^{DT} and f^{IDT} aim to transform the correlation representation space, enabling effective comparison. The overall process for the ith pyramid layer can be described as

$$C_l^{(q\to s)} = f^{\mathrm{IDT}}(\mathrm{Lrn}^{\mathrm{QS}}(f^{\mathrm{DT}}(C_l))). \tag{2}$$

Specifically, f^{DT} involves swapping and merging certain dimensions of the input correlation, facilitating subsequent local analysis. Formally given by

$$f^{\mathrm{DT}} : \mathbb{R}^{(n_l,H_q^l,W_q^l,H_s^l,W_s^l)} \to \mathbb{R}^{(H_q^l\times W_q^l,n_l,H_s^l,W_s^l)}. \tag{3}$$

Next, $\mathrm{Lrn}^{\mathrm{QS}}$, comprising of 2-D convolution, group normalization, and ReLU activation, conducts an in-depth, local-based analysis. In details, $\mathrm{Lrn}^{\mathrm{QS}}$ increases the channel counts while preserving the spatial shape $[\mathbb{R}^{(H_q^l\times W_q^l,8,H_s^l,W_s^l)}]$. On the one hand, maintaining spatial constancy ensures the retention of critical detail information. On the other hand, the augmented channel count bolsters the expressive capacity. Subsequently, the dimension transformation function f^{IDT} is applied to the correlation tensors, acting as the reverse of f^{DT}. This step is critical in reverting the refined correlation $C_l^{(q\to s)}$ to its original spatial configuration.

2) The SQ Branch: This branch contrasts the QS branch with the perspective by viewing the query image from the support image. The dimensional permutation operation, T, is utilized to swap the positions of the query and support in the correlation space. T is added to both the head and the tail of this branch. Initially, T shifts the reference from the query side, $C(x\_q, x\_s)$, to the support side, $C(x\_s, x\_q)$. At the end, T switches the perspective back to the query side, to enable interaction with the QS branch output. This whole process of this branch can be formalized as

$$C_l^{(q\leftarrow s)} = T(f^{\mathrm{IDT}}(\mathrm{Lrn}^{\mathrm{SQ}}(f^{\mathrm{DT}}(T(C_l))))). \tag{4}$$

Specifically, the initial correlations first undergo the dimension permutation, expressed as

$$T : C_l \in \mathbb{R}^{(n_l,H_q^l,W_q^l,H_s^l,W_s^l)} \to C_l^T \in \mathbb{R}^{(n_l,H_s^l,W_s^l,H_q^l,W_q^l)} \tag{5}$$

where T rearranges the 1st, 2nd and 3rd, 4th dimensions of C_l to generate the transposed correlation C_l^T. Subsequently, C_l^T sequentially traverses through f^{DT}, the Local Relearning block $\mathrm{Lrn}^{\mathrm{SQ}}$, and f^{IDT} as outlined earlier. It's notable that $\mathrm{Lrn}^{\mathrm{SQ}}$ and $\mathrm{Lrn}^{\mathrm{QS}}$, despite having identical architecture, learn independently from each other. Subsequently, we apply the T operation once more to revert the correlation order and acquire the output of the SQ branch.

After independent comparative analysis in both directions, we combine the output from two branches to interactively enhance the feature correlation. Combining the unilaterally adjusted correlation: $C_l^{(q\to s)}$ and $C_l^{(q\leftarrow s)}$, we calculate the mutually enhanced correlation as following:

$$C_l^{(q\leftrightarrow s)} = C_l^{(q\to s)} + C_l^{(q\leftarrow s)}. \tag{6}$$

E. SA Strategy

SA strategy performs a progressive aggregation of refined local insights to construct the encoded features, applying multiple stages of local convolutional pooling to down-sample support dimensions. As depicted in Fig. 5(b), SA networks adopt a pyramidal architecture. For the ith pyramid layer, the aggregation process can be formulated as

$$C_{\mathrm{map}}^l = \mathrm{DC}(\mathrm{STA}^l(\mathrm{DM}(C_l^{(q\leftrightarrow s)}))). \tag{7}$$

Initially, we apply a dimension merge operation DM to merge the 2nd and 3rd dimensions of input $C_l^{(q\leftrightarrow s)}$, which facilitates subsequent aggregation. Formally represents as

$$\mathrm{DM} : \mathbb{R}^{(8,H_q^l,W_q^l,H_s^l,W_s^l)} \to \mathbb{R}^{(8,H_q^l\times W_q^l,H_s^l,W_s^l)}. \tag{8}$$

Subsequently, the SA process embarks on a progressive aggregation of local features, learning and encoding intricate associations among them. More precisely, it compresses the 3rd and 4th dimensions of correlation to obtain a more compact feature representation. Each layer of STA^l contains a specific number of blocks, with each block comprising a sequence of 2-D convolution, GroupNorm, and ReLU. Subsequently, we decompose

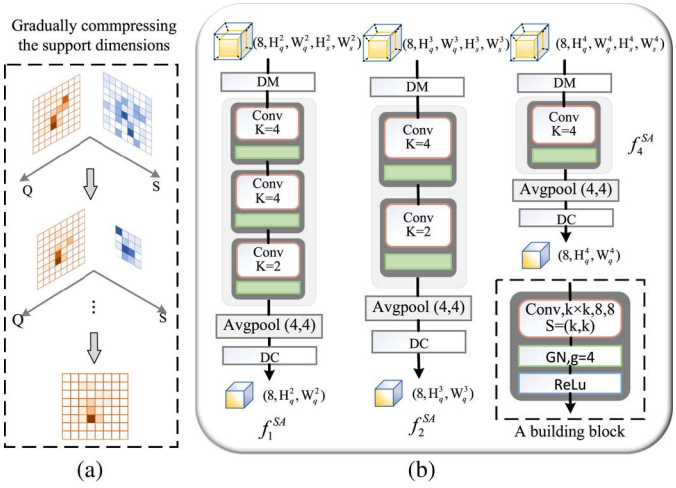

Fig. 5. (a) SA strategy progressive aggregates refined local insights through a dimensional reduction within the support domain. (b) Building blocks in SA strategy: $\{f_l^{SA}\}_{l=2}^{4}$, "g" denotes the number of groups utilized in group normalization.

the 2nd dimension to reestablish the original dimensions of the analyzed correlation

$$\text{DC} : \mathbb{R}^{(8,H_q^l \times W_q^l)} \to \mathbb{R}^{(8,H_q^l,W_q^l)}. \tag{9}$$

Ultimately, we obtain the encoded context for lth layer.

Similarly, a series of comparative features $\{C_{\text{map}}^l\}_{l=2}^{4}$ has been obtained. Subsequently, two outputs from adjacent pyramidal layers are consolidated by elementwise addition after the spatial dimensions of the previous layer are upsampled by a factor of two, similar to the FPN. This process is illustrated in Fig. 3. We sequentially fuse the features from 4th, 3rd and 2nd layers to form the encoded context, a 2-D feature map, i.e., $C_{\text{map}} \in \mathbb{R}^{(8,H_q^2,W_q^2)}$), a compact, multilayer representation of comparison results.

F. Defect-Sensitive Weight Learner

To maximize the defect-related knowledge in D_{train}, the DSWL prioritizes informative channels using structural priors (e.g., texture and shape). The specific structure is depicted in Fig. 3. During training, DSWL refines channel weights W by evaluating the channelwise discrepancies between query mask M_q and the comparison-based map C_{map}. For compatibility in spatial size and channel dimensions, the query mask $M_q \in \mathbb{R}^{(H_q,W_q)}$ is resized to (H^2, W^2) and expanded to eight channels, aligning with the $C_{\text{map}} \in \mathbb{R}^{(8,H^2,W^2)}$. Posttraining, DSWL applies these learned weights to C_{map} for enhanced defect detection. Specifically, the channelwise distance is computed using the structural similarity index (SSIM) [35], denoted as

$$D(C_{\text{map}}, M_q) = \text{SSIM}(C_{\text{map}}, M_q). \tag{10}$$

Notably, DSWL prioritizes structural priors to offer localized cues. Consequently, the luminance component, which mainly evaluates average pixel similarity, is excluded from the SSIM computation due to its limited relevance. The DSWL derives channel weights W directly from the C_{map} through a Learner network composed of a global convolution and a sigmoid layer. Finally, the computed distance, denoted as W_{gt}, is used to guide the learning of W with the MSE loss

$$L_{\text{DSW}}(W_{gt}, W) = \text{MSE}(W_{gt}, W). \tag{11}$$

Finally, the learned channel attention W is applied to the encoded context C_{map}.

G. Loss Function Design

The proposed PLMNet is optimized through two Loss functions: $L_{\text{Seg}}, L_{\text{DSW}}$. L_{Seg}, the segmentation loss, is computed as the cross-entropy between the predicted map $\hat{M}$ and the query mask M_q, formally defined as

$$L_{\text{Seg}}(\hat{M}, M_q) = \text{LCE}(\hat{M}, M_q). \tag{12}$$

The predicted map $\hat{M} \in \mathbb{R}^{(2,H^2,W^2)}$, is produced through a decoder network, the structure of which is illustrated in Fig. 1. This decoder network consists of sequential 2-D convolutional layers, ReLU activation functions, upsampling layers, and a softmax function. Meanwhile, L_{DSW} quantifies the discrepancy between the learned channel weight W and the target weight W_{gt}, which is evaluated using the SSIM metric as detailed in Section III-F. The final total loss, L_{Total}, is a weighted sum of L_{Seg} and L_{DSW}, with weights α and β respectively, as follows:

$$L_{\text{Total}} = \alpha L_{\text{Seg}}(\hat{M}, M_q) + \beta L_{\text{DSW}}(W_{gt}, W). \tag{13}$$

H. Extension to K-Shot Setting

Given K support images $S = \{I_s^k\}_{k=1}^{K}$ and a query image I_q, our model can easily extend to K-shot setting by performing K forward passes to get corresponding mask predictions $\{\hat{M}\}_{k=1}^{K}$. We get anomaly scores for every pixel location by calculating the mean value over all the K predictions. Pixels with anomaly scores exceeding the threshold of 0.5 are designated as anomalous, while those beneath this threshold are considered normal.

IV. Experiments

This section elaborates on the experiments' details, including experiment settings, results, and analysis. Section IV-A presents the dataset and experimental protocol. Section IV-B presents the evaluation metrics and comparison methods we choose. In Section IV-C, we describe the implementation details. Section IV-D analyzes the results of our method. Section V demonstrates the ablation study.

A. Experimental Protocol

To evaluate the models' adaptability to multiple unseen classes, the C-FSAD setting involves training each model on multiple known categories, which is then tested across various unseen categories. Specifically, for the testing set comprising C_{test} classes, we employ a single unified model for each, which

TABLE I
DATASET PARTITION RESULTS OF MVTEC-AD

	5^0	5^1	5^2	5^3	5^4
MVTec-5^i	Toothbrush Metalnut Carpet	Grid Capsule Bottle	Tile Zipper Pill	Leather Transistor Cable	Hazelnut Screw Wood

is trained on a training set with C_{train} classes. The proposed PLMNet is evaluated on two real-world industrial anomaly detection datasets: MVTec-AD [36], encompassing 5354 high-resolution images across 15 categories, and the recently introduced MPDD [37], featuring anomaly detection in the painted metal part fabrication with six classes.

1) MVTec-AD: We implement a 5-fold cross-validation, as outlined in [27], [38], dividing the dataset into fivefolds, with each fold comprising three distinct classes. The partitioning details are in Table I, with each fold denoted as MVTec-5^i. For each fold, the model was trained on the aggregated data of the 12 classes and then tested on the remaining three classes, ensuring nonoverlapping classes. The training uses a 5:1 normal-to-anomalous sample ratio with random defect selection.

2) MPDD: To demonstrate the robustness of our method to domain shift, we experiment with cross-dataset generalization following the recent work of [39] and [26]. Specifically, we evaluate MVTec-trained PLMNet on each class of MPDD.

We primarily evaluate in a one-shot setting to highlight the rapid generalization. Moreover, to assess the impact of synthetic versus real defects on performance, we experiment using synthetic defects in PLMNet following [40].

B. Baselines

For a comprehensive evaluation, we compare PLMNet against 11 typical baseline methods: one FSAD model (RegAD [11]), two few-shot segmentation models (HSnet [27], IPMT [41]), two segmentation models (DeepNetV3+ [42], PSPNet [43]), two saliency detection methods (CPD [44], BASNet [45]), one Camouflage detection (SINet [46]), and three multiclass novelty detection (GANomaly [33], deep-embed [34], NLN [32]). To achieve a fair comparison with RegAD, we perform the reestimation with support samples before generalizing to new classes as mentioned. Multiclass supervised novelty detection models are trained with normal samples on known categories. For novel categories, we implement fine-tuning using a limited number of normal samples and classify all anomalies as outliers.

The primary metric used in anomaly detection is the area under the receiver operating characteristic curve (ROC AUC). We adopt image-level AUC for classification accuracy, pixel-level AUC for segmentation precision, and per-region overlap (PRO) for region-level performance. Image-level AUC is derived from the true positive rate of correctly identified anomalies versus the false positive rate of misclassified normal images. Pixel-level AUC is obtained by comparing the pixel-by-pixel anomaly map against the ground truth. Instead of treating every pixel independently, region-level metrics PRO averages the performance over each connected component of the ground truth. The PRO can be computed as

$$\text{PRO} = \frac{1}{N}\sum_i \sum_k \frac{|P_i \cap C_{i,k}|}{|C_{i,k}|} \tag{14}$$

where $C_{i,k}$ denotes the set of pixels marked as anomalous for a connected component k in the ground truth, and P_i denotes the set of pixels predicted as anomalous. For PRO, the detection of smaller anomalies is considered equally important as the detection of larger ones. For MVTec-AD, the mean area under the receiver operating characteristic curve (AUROC) is computed over categories for each fold and then across folds. For MPDD, it is computed per object category and averaged over the six categories.

C. Implementation Details

To ensure fair comparisons, the ResNet50 [47] backbone, pretrained on ImageNet [48], is consistently utilized across experiments. In (13), the loss proportions are specified as $\alpha = 0.7$ and $\beta = 0.3$. The network optimization utilizes the Adam [49] algorithm with a learning rate of $1e-3$, and image spatial dimensions are set to 512. All experiments are conducted on a server equipped with four 3090 GPUs, each handling four input pairs at a time.

D. Results and Analysis

1) K-Fold Cross-Validation on MVTec-AD: To evaluate PLMNet's generalization and robustness over novel classes, we conduct a comprehensive evaluation, integrating both numerical metrics and qualitative visualizations. Table II summarizes the results on unseen categories, while Table III details the results for categories included in the training. Meanwhile, Fig. 6 graphically illustrates the metrics for each unseen category. 1) *Generalization*. Table III reveals that PLMNet outperforms competing methods in trained categories, particularly in terms of average AUROC at the pixel and region levels, with margins of 2.3%p and 1.8%p compared to HSNet [27]. This might be attributed to PLMNet's feature-capturing ability and keen sensitivity to local details. Moreover, Table II demonstrates PLMNet's superior performance on unseen categories, boasting enhancements of 9.8%p image and 5.6%p pixel AUROC, compared to RegAD [11]. While the registration proxy task utilized by [11] contributes to cross-category adaptability, its emphasis on global registration loss potentially neglects critical localized features. These suggest that PLMNet effectively handles both familiar and novel categories, displaying remarkable adaptability and generalization capabilities. Fig. 6 further reveals our advantage across nearly all unseen classes, demonstrating PLMNet's robust adaptability in diverse anomaly detection scenarios. Furthermore, Fig. 7 visually displays the anomaly localization prowess, highlighting PLMNet's improved segmentation precision and reduced false positives in nonanomalous regions. 2) *Robustness.* In further experiments where real defects are substituted with synthetic ones, our method continues the SOTA results, underscoring its robustness in Cross-Category anomaly detection, without an exclusive dependence on genuine defect

TABLE II
ONE-SHOT ANOMALY DETECTION PERFORMANCE ON UNSEEN CATEGORIES ACROSS DIFFERENT FOLDS ON MVTEC-5^i

Methods	Fold 1			Fold 2			Fold 3			...	Fold 5			Average		
	Img	Pixel	PRO	Img	Pixel	PRO	Img	Pixel	PRO		Img	Pixel	PRO	Img	Pixel	PRO
DeepNetV3+ [42]	73.9	70.0	56.1	59.1	83.3	45.1	64.8	65.1	48.7		51.1	79.1	49.7	62.2	74.6	56.7
PSPNet [43]	87.4	76.3	65.1	55.2	85.3	50.3	79.6	80.4	48.1		60.2	80.4	39.8	70.6	85.6	70.4
CPD [44]	64.9	84.0	78.2	56.8	88.0	60.5	62.5	90.0	78.0		55.0	80.6	68.8	59.8	88.2	68.7
BASNet [45]	56.2	64.4	39.9	54.7	82.6	54.8	52.6	76.6	51.2		52.7	88.0	93.9	53.4	78.0	53.9
SINet [46]	70.6	88.5	76.0	61.1	82.4	43.2	65.4	77.5	70.2		58.8	78.5	57.7	64.0	81.7	63.6
HSNet [27]	93.8	98.2	92.5	85.4	92.6	86.0	86.2	91.8	87.9	...	71.9	91.6	75.3	84.3	93.6	83.9
IPMT [41]	94.9	97.5	91.7	83.4	88.5	84.1	86.8	90.9	86.1		69.8	88.6	73.6	83.7	91.4	82.4
RegAD [11]	84.3	91.5	85.4	78.8	86.3	81.6	80.8	94.3	88.3		80.2	94.6	81.2	80.8	91.7	84.1
GANomaly [33]	68.4	75.1	59.8	67.2	68.8	58.5	80.2	81.3	76.3		63.7	76.6	51.2	65.7	80.3	59.7
Deep-embed [34]	79.1	83.1	69.6	82.7	86.1	74.2	81.5	84.1	70.3		71.9	73.5	60.2	78.9	83.6	73.2
NLN [32]	86.7	83.1	85.3	78.9	85.6	83.9	93.5	86.0	81.5		76.4	78.9	86.4	81.1	85.4	80.2
PLMNet	**95.4**	**98.3**	**94.6**	**89.3**	**96.3**	**90.5**	**90.7**	**97.2**	**91.1**		**82.7**	**96.3**	**84.1**	**89.6**	**97.0**	**88.2**
PLMNet†	92.4	94.6	92.4	84.6	94.9	87.1	83.7	96.5	89.8		72.2	93.5	79.7	83.3	94.8	86.3

Note: The best results are in **bold** while the second best are underlined. PLMNet† denotes PLMNet trained with an equal number of synthetic defects as mentioned in Section IV-A. "Img" and "Pixel" represent the mean image-level and pixel-level AUROC.

TABLE III
ONE-SHOT ANOMALY DETECTION PERFORMANCE ON TRAINED CATEGORIES ACROSS DIFFERENT FOLDS ON MVTec-5^i

Methods	Fold 1			Fold 2			Fold 3			...	Fold 5			Average		
	Img	Pixel	PRO	Img	Pixel	PRO	Img	Pixel	PRO		Img	Pixel	PRO	Img	Pixel	PRO
DeepNetV3+ [42]	81.2	83.2	73.2	68.6	73.2	58.2	82.3	86.2	78.4		80.2	86.4	83.2	75.2	82.6	71.2
PSPNet [43]	92.4	94.3	86.5	67.4	73.5	69.9	88.5	94.4	90.2		68.9	91.4	83.4	80.1	92.6	76.5
CPD [44]	78.9	91.0	84.1	76.2	89.2	67.2	69.5	87.6	82.1		66.3	83.6	74.8	77.8	92.2	79.2
BASNet [45]	77.5	82.3	73.6	79.2	82.6	68.2	63.6	71.6	67.2		52.7	85.1	87.9	76.4	82.1	67.3
SINet [46]	85.6	90.5	76.0	69.1	82.4	77.2	80.4	82.3	76.1		61.8	71.5	62.7	79.2	82.7	78.9
HSNet [27]	98.2	99.2	93.1	92.4	95.7	89.2	89.1	93.7	89.2	...	91.9	98.6	89.1	93.3	96.6	90.5
IPMT [41]	99.2	**99.4**	94.2	91.2	96.3	89.7	94.7	93.4	87.3		78.8	82.6	88.1	92.1	94.7	89.4
RegAD [11]	97.8	99.1	92.8	93.8	95.2	87.6	94.4	98.1	88.3		**89.2**	94.6	87.9	94.2	93.1	86.1
GANomaly [33]	77.4	87.1	67.3	78.1	67.8	65.1	83.2	82.3	78.3		72.7	74.1	61.2	78.7	84.3	68.4
Deep-embed [34]	84.1	86.2	74.6	83.7	87.1	89.2	89.5	92.4	80.1		78.2	78.1	67.8	87.9	89.4	87.1
NLN [32]	95.2	93.1	88.1	84.6	84.4	89.4	93.3	86.8	82.1		89.1	81.9	87.4	91.1	92.4	85.2
PLMNet	**99.4**	98.9	**93.6**	**95.1**	**98.9**	**92.4**	**97.5**	**97.2**	**95.1**		**86.7**	**98.1**	**88.5**	**94.9**	**98.9**	**92.3**
PLMNet†	94.2	97.2	94.0	92.6	98.2	89.2	90.4	91.0	92.1		80.1	98.3	82.8	87.9	95.3	88.7

Note: Bold entries indicate the best results.

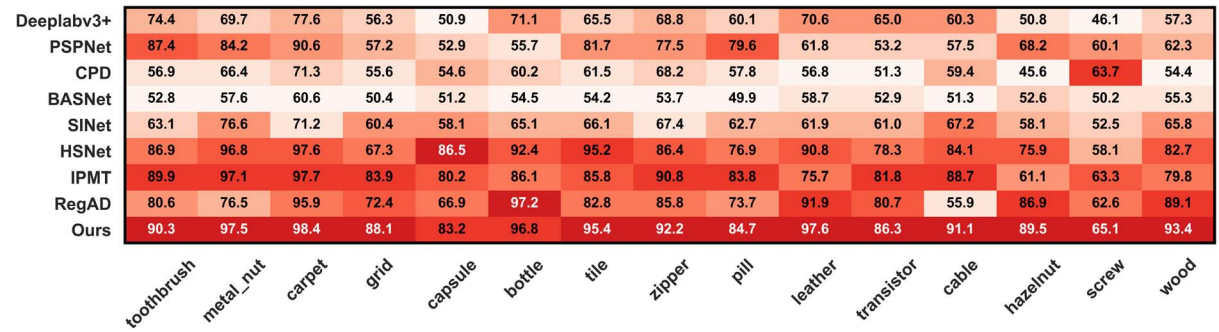

Fig. 6. Image-level AUROC on MVTec-AD across each category (one-shot). Darker colors represent better performance.

data. This robustness may stem from our direct comparison in the correlation space, enhancing sensitivity in distinguishing anomalous regions, contrary to other FSAD methods [6], [7] that calculate distances in the distribution space, potentially losing critical information inherent in direct features.

2) Domain Shift (MVTec → MPDD): As outlined in Section IV-A, we conduct cross-dataset experiments from MV Tec-AD to MPDD dataset. As detailed in Table IV, trained without data augmentation, PLMNet maintains state-of-the-art performance in the presence of large domain gaps, achieving a 7%–20% AUC-PR improvement compared to other methods. Impressively, PLMNet also boasts the shortest inference times (0.17 s), exceeding that of [11] and [45] by factors of 14 and 10.7, respectively. These attributed to their resource-intensive processes, for RegAD, the reestimation and data augmentation overheads, and BASNet, its intricate architecture centered on

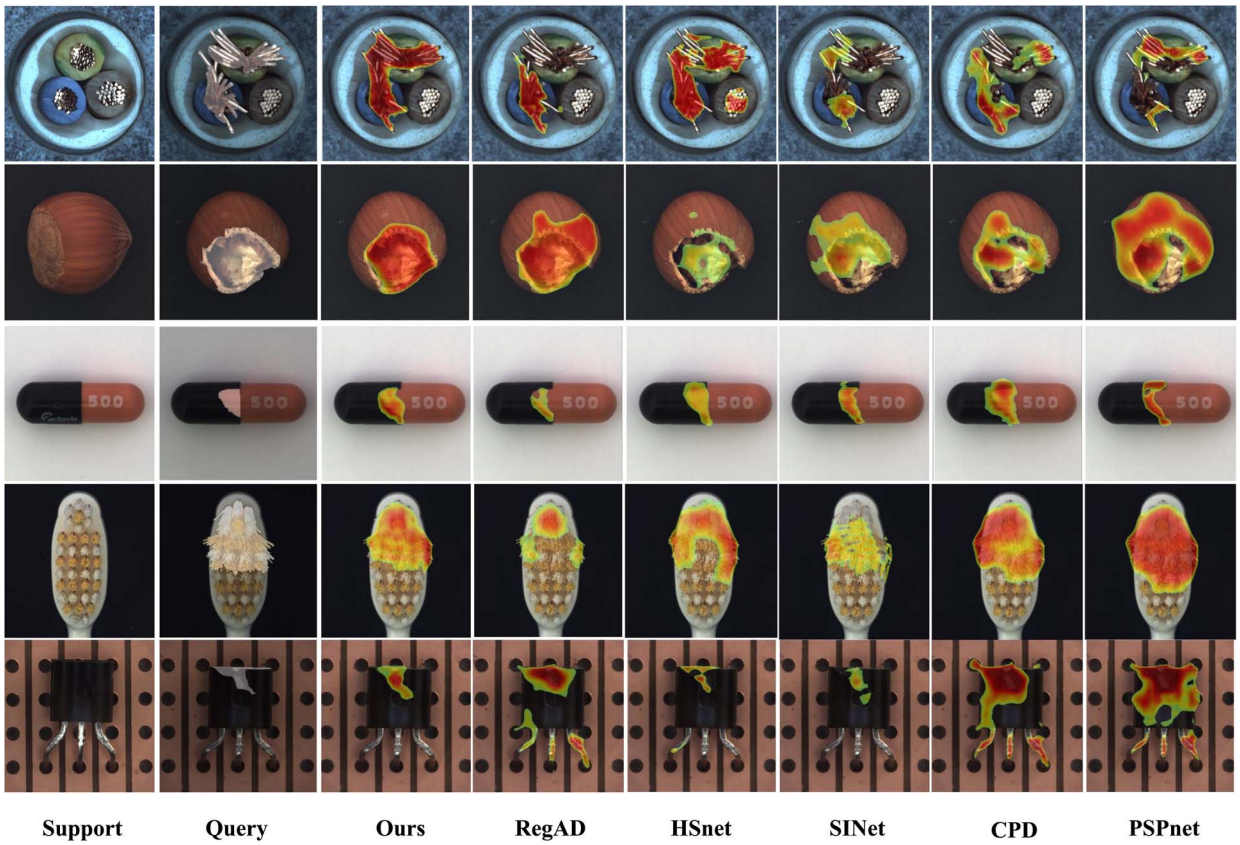

Fig. 7. Visualization of competing methods and PLMNet on MVTec-AD (one-shot).

TABLE IV
DOMAIN SHIFT RESULTS

Methods	One-Shot		Five-Shot		Aug	Times(s)
	Img	Pixel	Img	Pixel		
DeepNetV3+ [42]	57.6	84.7	67.8	83.5	3	0.36
PSPNet [43]	60.5	89.2	75.3	88.7	0	0.41
CPD [44]	59.2	86.1	73.9	87.5	3	0.37
BASNet [45]	53.3	83.7	71.2	85.6	2	1.82
SINet [46]	56.4	87.1	68.3	83.1	2	0.36
HSNet [27]	69.2	90.6	76.1	89.2	0	0.19
IPMT [41]	66.8	88.6	74.8	88.7	2	0.31
RegAD [11]	65.2	85.6	70.3	85.3	9	2.48(0.39)
GANomaly [33]	49.7	76.3	52.3	79.4	2	0.76
Deep-embed [34]	58.9	80.1	67.8	83.5	0	0.46
NLN [32]	64.4	84.1	71.4	86.4	0	0.28
PLMNet	**76.1**	**91.9**	**81.8**	**92.6**	0	0.17
PLMNet †	78.1	96.3	77.9	91.7	0	0.17

Note: The "Aug" signifies the number of applied data augmentation. In the notation, 2.48(0.39), 0.39 s is the data augmentation time, and 2.48 s is the total inference time. Bold entries indicate the best results.

global image contrast. Notably, our model particularly excels in pixel-level analyses, achieving a substantial 5.6%–13.2% AUC-PR improvement over multiclass novelty detection methods. This underscores PLMNet's exceptional capability to detect fine-grained, localized defects. In contrast, multiclass novelty detection methods focus primarily on broader, macroscopic feature variations among object categories, a strategy that proves less effective for defect detection.

TABLE V
ABLATION STUDY TO EXPLORE THE CONTRIBUTION OF LP NETWORK, SA STRATEGY, AND DEFECT-SENSITIVE WEIGHT LEARNER

LP	SA	DSWL	One-Shot		Five-Shot	
			Img	Pixel	Img	Pixel
			82.2	89.7	83.6	91.6
✓			85.6	93.5	87.2	93.1
	✓		83.3	91.6	85.6	92.5
		✓	84.9	92.3	86.2	93.9
✓	✓		87.2	94.8	89.6	95.2
	✓	✓	86.3	95.2	88.8	95.7
✓		✓	88.6	96.3	90.5	97.2
✓	✓	✓	**89.9**	**97.1**	**91.2**	**98.1**

Note: Bold entries indicate the best results.

V. ABLATION STUDY

In this section, we conduct a comprehensive ablation study to validate the contributions of major modules in our model. Furthermore, we evaluate the impact of defective samples in terms of both quantity and source. All ablation experiments are performed with K-fold cross-validation on MVTec-AD.

A. Ablation Analysis of Main Components

To validate the contributions of the major components in our model, we selectively turn these modules on (✓) or off (×) and present the experimental results in Table V. It reveals that activating the LP network alone enhances the AUROC by 3.4%p at the image level and 3.8%p at the pixel level.

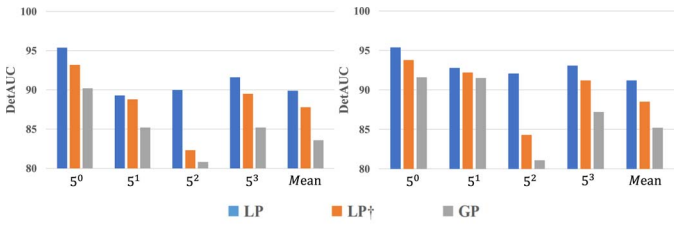

Fig. 8. Ablation study on LP network on MVTec-AD [36] dataset in one-shot (left) and five-shot (right) detection results. "GP" applies global transformers as feature learning units instead of local convolutions. "LP†" removes the query-to-support branch.

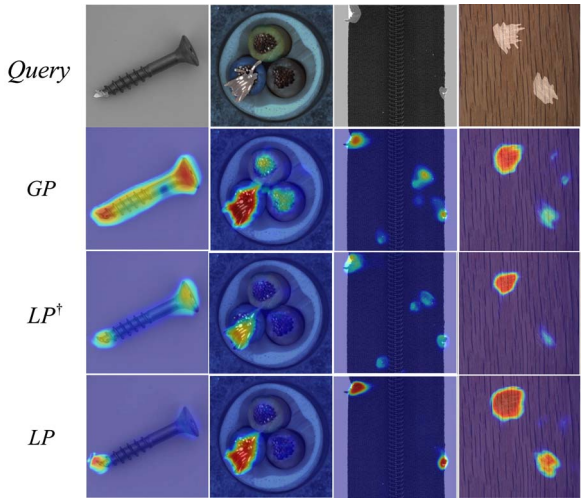

Fig. 9. Visualization of the LP Network ablation study.

Omitting the SA strategy does not significantly impact the results, while the inclusion of DSWL notably boosts performance by 2.7%p for image-level and 2.6%p for pixel-level AUROC. Individually, LP improves detail-focused learning, and DSWL effectively adjusts channel significance. The limited effect of feature aggregation suggests the sufficiency of the features from localized processing.

B. LP Network

Fig. 8 shows a significant performance decline when the LP network is substituted with global transformer blocks, highlighting the efficacy of local processing in anomaly detection. Over-reliance on global analysis, as demonstrated in Fig. 9, leads to higher misjudgment rates and incorrect segmentations due to unintended integration of category-specific information. Additionally, Fig. 8 indicates improved matching accuracy with enhanced dual-branch interaction, emphasizing the importance of branch synergy in anomaly localization. The dual-branch structure, as shown in Fig. 9, effectively facilitates precise anomaly detection and a more nuanced understanding of defects. This supports the adoption of dual-branch structures for improved accuracy and detailed analysis in anomaly detection tasks.

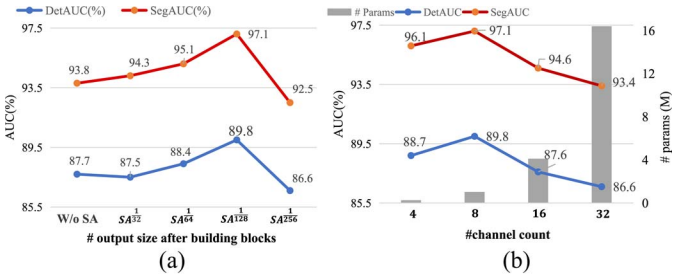

Fig. 10. Parameter sensitivity analysis. (a) Effect of depths in building blocks: f_l^{SA}. We use output sizes after the building blocks as proxies to reflect depth variations. (b) Impact of channel count in DSWL.

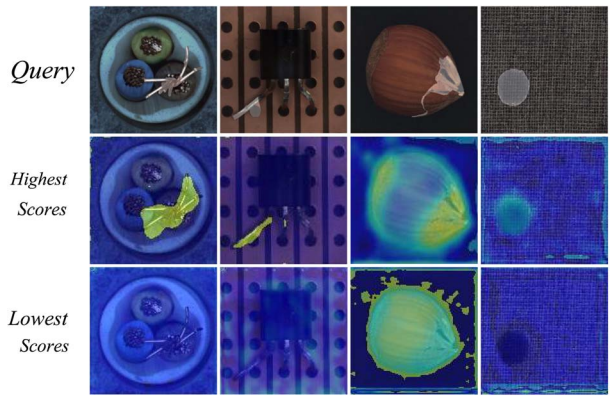

Fig. 11. Visualization of feature channels in the encoded context. The second and third rows display feature channels with the highest and lowest scores, learned by DSWL.

C. SA Strategy

We experiments with varying number of 2-D conv layers in the building blocks: f_p^{SA}. Fig. 10(a) plots one-shot results on the MVTec-AD, the x-axis represents the postconvolution output size (indirectly reflecting the number of 2-D conv layers). It reveals that appending additional 2-D convolution layers in the building blocks provides a clear performance increase. However, when the output size attains 1/128 of the original image dimensions, the performance stabilizes. We, therefore, deploy stacks of 2-D layers [4], [4,2], and [4,2,2] for the 4th, 3rd, and 2nd layers, respectively.

D. Defect-Sensitive Weight Learner

To thoroughly evaluate the DSWL, we conduct ablation studies on channel count and visualization experiments on encoded features. In Fig. 10(b), as the channel count increases, performance improves but plateaus at eight channels, indicating redundant additions beyond this. Consequently, we ascertain eight as the optimal channel count. In Fig. 11, channels with the highest scores closely reflect defect structures, capturing significant defect information. In contrast, lower-score channels focus more on backgrounds or nondefect details. This observation indicates a channel preference for defect features, with DSWL adeptly prioritizing these pertinent channels to boost generalization.

TABLE VI
ABLATION EXPERIMENT ON THE NUMBER OF ANOMALOUS SAMPLES

		(5)	[1/30]	[1/20]	[1/10]	[1/5]	S[1/5]	S[1/10]
One-shot	Img	86.0	88.0	88.3	88.9	**89.9**	84.6	83.2
	Pixel	95.9	96.5	96.3	96.6	**97.1**	94.4	93.3
Five-shot	Img	86.8	89.7	89.9	90.3	**91.2**	85.5	84.7
	Pixel	95.1	96.2	97.1	97.6	**98.1**	96.1	95.6

Note: (5) denotes five actual defective samples for each class. [1/5] presents a ratio of anomalous samples to normal ones at 1:5 for each category in the training set. So as [1/10], [1/20], [1/30]. $S[1/5]$ and $S[1/10]$ indicate synthetic defect samples. Bold entries indicate the best results.

E. Discussion on Anomalous Samples

To assess our model's robustness to anomalous samples, we adjust authentic defect counts and substitute them with synthesized ones as per [40]. As depicted in Table VI, when trained on synthesized defects equivalent in number, the performance significantly trails those trained on actual defects (e.g., from [1/5] to S[1/5], the detection AUROC drops by 6%, and the segmentation AUROC drops by 2.7%). It reveals the limitations of current synthetic anomaly generation methods, which lack the inherent complexity and diversity of real-world anomalies. Remarkably, with just five samples per class, PLMNet outperforms [11] by 4–6%p AUROC, demonstrating its efficiency in leveraging limited defect samples, which is a significant advantage given the scarcity of high-quality defect samples in real-world settings.

VI. CONCLUSION AND DISCUSSION

This article tackles the novel and challenging task of C-FSAD, which holds significant implications for visual inspection in flexible industrial production. Specifically, we propose a novel PLMNet, discerning ambiguous defect regions and performing anomaly detection on unseen categories effectively. PLMNet features an innovative LP module for detailed analysis in the correlation space, coupled with a SA module to synthesize local insights into an encoded context. This context is further refined through a defective-sensitive learner module, enhancing the discernment of subtle feature discrepancies. Our experiments, utilizing a limited number of real defective samples, reveal valuable information of even limited real defects, a fact previously underexplored in FSAD research. Comprehensive evaluation results confirm the superiority of the proposed PLMNet, with an approximate 10% AUC-PR improvement over existing FSAD methods on two industrial anomaly detection datasets.

A. Potential Applications

This section discusses several potential applications that could benefit from the research of C-FSAD.

1) Visual Inspection in Flexible Industrial Production: In the context of flexible manufacturing systems that are required to produce different types and quantities of products with minimal reconfiguration, the challenge of multiclass anomaly detection holds significant importance [3], [50]. Anomaly detection models must learn to differentiate between normal and defective regions across multiple categories [7]. C-FSAD methods offer a crucial solution by rapidly learning the definition of normalcy with minimal samples.

2) Obstacle Detection for Autonomous Driving: In the context of autonomous driving, obstacle detection is critical for ensuring the vehicle's and passengers' safety. While the perception of autonomous vehicles performs well under closed-set conditions, they struggle to handle the unexpected [3], [51]. C-FSAD can be leveraged to detect obstacles in complex and diverse driving scenarios, enabling the system to adapt to new types of obstacles quickly, and enhancing the robustness of obstacle detection in autonomous driving applications.

B. Future Research Directions

In this part, we discuss potential research directions for advancing C-FSAD.

1) Vision-Language Anomaly Detection: In the C-FSAD task, we address anomaly detection in novel classes using a limited number of support images. However, this constraint on visual data can hinder performance in new categories [52], [53]. Several researches ([54], [55]) suggest that incorporating text input can effectively support zero-shot anomaly detection. Our future research will consider integrating textual data to improve model performance in FSAD with limited visual information.

2) Anomaly Detection From Imbalanced Data: Due to the scarcity of anomalous samples, anomaly detection datasets typically show imbalanced data distribution, with certain categories possessing only a few defective samples. Data imbalance presents a considerable obstacle in anomaly detection, as models must rapidly adapt and mitigate the issue of memory retention, especially when faced with the sparse training data typical of long-tailed categories ([15], [16], [56], [57]). Despite these challenges, addressing this issue holds substantial practical value and merits further effort.

REFERENCES

[1] V. Zavrtanik, M. Kristan, and D. Skocaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8330–8339.
[2] L. Tomasetti, K. Engan, L. J. Høllesli, K. D. Kurz, and M. Khanmohammadi, "Exploiting 4D ct perfusion for segmenting infarcted areas in patients with suspected acute ischemic stroke," 2023, *arXiv:2303.08757*.
[3] D. Bogdoll, M. Nitsche, and J. M. Zöllner, "Anomaly detection in autonomous driving: A survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4488–4499.
[4] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Workshops Challenges Pattern Recognit. (ICPR),* Virtual Event, Milan, Italy: Springer, 2021, pp. 475–489.
[5] 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.*, 2022, pp. 98–107.
[6] G. Xie, J. Wang, J. Liu, F. Zheng, and Y. Jin, "Pushing the limits of fewshot anomaly detection in industry vision: Graphcore," 2023, *arXiv:2301.12082*.
[7] N. Belton, M. T. Hagos, A. Lawlor, and K. M. Curran, "Fewsome: Few shot anomaly detection," 2023, *arXiv:2301.06957*.
[8] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.
[9] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3379–3388.

[10] S. Sheynin, S. Benaim, and L. Wolf, "A hierarchical transformation-discriminating generative model for few shot anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8495–8504.
[11] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang, "Registration based few-shot anomaly detection," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Springer, 2022, pp. 303–319.
[12] P. Series and A. R. Seitz, "Learning what to expect (in visual perception)," *Frontiers Human Neurosci.*, vol. 7, Oct. 2013, Art. no. 668.
[13] F. Happé and U. Frith, "The weak coherence account: Detail-focused cognitive style in autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 36, no. 1, pp. 5–25, 2006.
[14] S. Baron-Cohen, E. Ashwin, C. Ashwin, T. Tavassoli, and B. Chakrabarti, "Talent in autism: Hyper-systemizing, hyper-attention to detail and sensory hypersensitivity," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 364, no. 1522, pp. 1377–1383, 2009.
[15] G. Pang, C. Ding, C. Shen, and A. v. d. Hengel, "Explainable deep few-shot anomaly detection with deviation networks," 2021, *arXiv:2108.00462*.
[16] C. Ding, G. Pang, and C. Shen, "Catching both gray and black swans: Open-set supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7388–7398.
[17] L. Ruff et al., "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.,* PMLR, 2018, pp. 4393–4402.
[18] X. Tao, S. Yan, X. Gong, and C. Adak, "Learning multiresolution features for unsupervised anomaly localization on industrial textured surfaces," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 127–139, Jan. 2024.
[19] J. Yu et al., "Fastflow: Unsupervised anomaly detection and localization via 2D normalizing flows," 2021, *arXiv:2111.07677*.
[20] C. Cao, Y. Lu, P. Wang, and Y. Zhang, "A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20392–20401.
[21] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2041–2050.
[22] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Springer, Aug. 2020, pp. 763–778.
[23] W. Huang et al., "Boundary-aware network with topological consistency constraint for optic chiasm segmentation," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1504–1513, Dec. 2023.
[24] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning," in *BMVC*, vol. 3, no. 4, 2018.
[25] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9196–9205.
[26] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21984–21996.
[27] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6941–6952.
[28] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14318–14328.
[29] J. Santos, T. Tran, and O. Rippel, "Optimizing patchcore for few/many-shot anomaly detection," 2023, *arXiv:2307.10792*.
[30] I. I. Osman and M. S. Shehata, "Few-shot learning network for out-of-distribution image classification," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1579–1591, Dec. 2023.
[31] D. Karimi and A. Gholipour, "Improving calibration and out-of-distribution detection in deep models for medical image segmentation," *IEEE Trans. Artif. Intell.*, vol. 4, no. 2, pp. 383–397, Apr. 2023.
[32] M. Mesarcik, E. Ranguelova, A.-J. Boonstra, and R. V. van Nieuwpoort, "Improving novelty detection using the reconstructions of nearest neighbours," *Array*, vol. 14, 2022, Art. no. 100182.
[33] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proc.* 14th Asian Conf. Comput. Vis. (ACCV), Perth, Australia, Springer, Feb. 2019, pp. 622–637.
[34] P. Burlina et al., "Where's wally now? Deep generative and discriminative embeddings for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 11507–11516.
[35] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 2366–2369.
[36] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9592–9600.
[37] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. Skotak, "Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions," in *Proc. 13th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 66–71.
[38] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Springer, 2020, pp. 142–158.
[39] D. Kang and M. Cho, "Integrative few-shot learning for classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 9979–9990.
[40] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9664–9674.
[41] Y. Liu, N. Liu, X. Yao, and J. Han, "Intermediate prototype mining transformer for few-shot semantic segmentation," 2022, *arXiv:2210.06780*.
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc.* 15th Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Springer, Sep. 2018, pp. 801–818.
[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2881–2890.
[44] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3907–3916.
[45] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.
[46] X. Hu et al., "Sinet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1010–1019, Mar. 2019.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 248–255.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[50] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *J. Artif. Intell. Res.*, vol. 46, pp. 235–262, 2013.
[51] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, Apr. 2020.
[52] S. Ni and H.-Y. Kao, "Masked Siamese prompt tuning for few-shot natural language understanding," *IEEE Trans. Artif. Intell.*, vol. 5, no. 2, pp. 624–633, Feb. 2024.
[53] W. Zhai, Y. Cao, J. Zhang, H. Xie, D. Tao, and Z.-J. Zha, "On exploring multiplicity of primitives and attributes for texture recognition in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 403–420, Jan. 2024.
[54] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19606–19616.
[55] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," 2023, *arXiv:2310.18961*.
[56] W. Liu et al., "Margin learning embedded prediction for video anomaly detection with a few anomalies," in *Proc. IJCAI*, 2019, pp. 3023–3030.
[57] M. Ochal, M. Patacchiola, J. Vazquez, A. Storkey, and S. Wang, "Few-shot learning with class imbalance," *IEEE Trans. Artif. Intell.*, vol. 4, no. 5, pp. 1348–1358, Oct. 2023.