IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 September 2023, accepted 6 October 2023, date of publication 12 October 2023, date of current version 20 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324314

RESEARCH ARTICLE

Reliable Anomaly Detection and Localization System: Implications on Manufacturing Industry

QING TANG[1], (Member, IEEE), AND HAIL JUNG[2], (Member, IEEE)
[1]Data Science Group, INTERX, Ulsan 44542, Republic of Korea
[2]Department of Business Administration, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea

Corresponding author: Hail Jung (hail95@seoultech.ac.kr)

This work was supported by the Korea Technology and Information Promotion Agency for Small and Medium Enterprises (TIPA) funded by the Korean Government [Ministry of SMEs and Startups (MSS)] under Grant RS-2022-00140573.

ABSTRACT Industry 4.0 has placed significant emphasis on interconnectivity, digitalization, and automation. Among the myriad innovative technologies that have surfaced, artificial intelligence (AI) stands out as especially instrumental in the development of fully autonomous factories. Product quality inspection is a critical component of industrial manufacturing. An accurate and reliable AI-based Anomaly Detection and Localization (ADL) system for industrial product quality inspection is essential in real-world manufacturing factories. Collecting massive anomalous products is difficult because the number of anomalous products is limited and rare in a realistic manufacturing scenario. Therefore, the paper treats the ADL problem as a cold-start challenge, training the defects inspection network only using nominal (non-defective) images. Significantly, the paper aims to bridge the gap between academic research and real-world manufacturing industry applications. The paper lists issues that current state-of-the-art academic research faces when applied in real-world manufacturing settings, then a Reliable Anomaly Detection and Localization (RADL) system is developed to solve the issues. RADL is improved in three aspects. Firstly, the common image pre-processing method is modified by considering the characteristics of real-world industrial images. Secondly, a Fake Defect Feature Augmentation (FDFA) strategy to mitigate the scarcity of real-world data. Thirdly, a Hardness-aware Cross-Entropy loss (HCELoss) is adopted to enhance the stability and reliability of the system. On the public MVTec AD benchmarks, the proposed RADL outperforms previous methods with 99.53% in I-AUROC, 97.85% in P-AUROC, and 91.60% in PRO. Furthermore, RADL is evaluated under industrial manufacturing settings in two real-world datasets collected from industrial production lines. The experimental results demonstrate the superiority of the proposed strategies in a public dataset and real-world manufacturing industrial environments.

INDEX TERMS Industry 4.0, anomaly detection, manufacturing industry, system reliability, autonomous factory.

I. INTRODUCTION

Detecting defects/anomalies play an important role in various manufacturing industrial production domains for holding product quality standards [1], such as Semiconductor Manufacturing Processes [2], Electronics Manufacturers [3], and Automation Manufacturers [4], etc. Vision-based Anomaly Detection and Localization (ADL) is usually performed at the final stage of the manufacturing process for inspecting product quality and identifying product defects. The quantity and severity of a defect significantly impact and determine the price of a product. In traditional and manual inspection scenarios, workers examine the quality of products one by one after the product is produced on a production line. Obviously, relying solely on human inspection becomes increasingly challenging with the increasing manufacturing volume and growing demands. Moreover, the subjective judgments and biases of workers result in inconsistent quality inspection criteria, which lead to a higher defect escape rate.

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

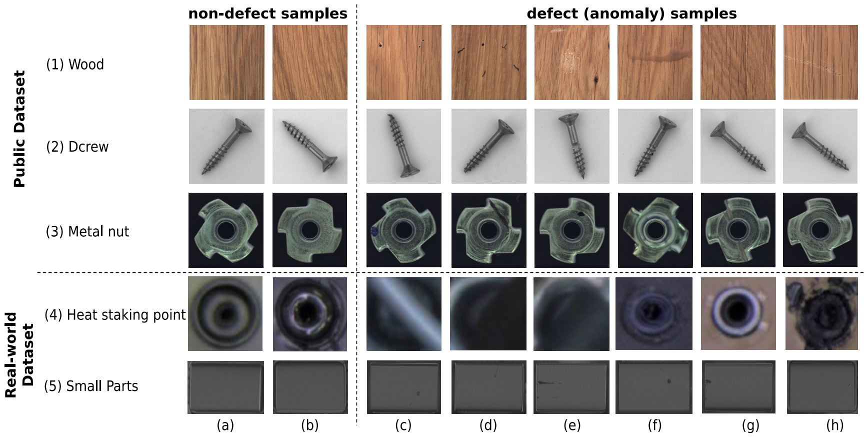


FIGURE 1. Sample images of public dataset MVTec AD [5] and two real-world manufacturing industrial datasets.

In the last decade, Artificial Intelligence (AI) and deep vision detection technology have been rapidly developing in real-world applications, such as autonomous vehicles [6], [7], surveillance systems, medical imaging [8], etc. In the meantime, AI-based or Deep Neural Networks (DNN) also have been adopted in factories for detecting and identifying defects (anomalies) of a product because they provide higher accuracy and faster inspection speed than traditional inspection methods. AI-based quality inspection system brings benefits to the manufacturing industry [1], [9], [10]. Firstly, the AI-based inspection system reduces the workload of workers by automating some manufacturing processes. Secondly, an AI-based model leads to more accurate defect inspection accuracy with less inspection time [11], [12], [13], [14], therefore increasing the overall efficiency of the manufacturing process for helping manufacturers meet higher demand. Thirdly, the AI-based inspection system enforces consistent inspection criteria across the whole production line to ensure that all products undergo consistent quality. Overall, the AI-based inspection system helps manufacturers increase productivity and efficiency in manufacturing processes.

An intuitive deep-learning method for defect detection and quality inspection is the classification of defective or non-defective using supervised binary classification networks [15], [16], [17] and the detection of the defective region. However, unlike the general object learning cases, industrial defect inspection is particularly difficult. Challenges of manufacturing industrial product defect inspection are summarized as follows:

1) The imbalance between non-defective (nominal) and defective (anomalous) data samples [10], [18], [19]. In real-world manufacturing industrial cases, anomaly data is difficult to collect because most production lines are faultless. Therefore, the number of nominal samples is usually the majority, while the number of anomaly samples is rare relatively.
2) The defect types are diverse. The defect can vary from subtle texture changes such as thin scratches to larger structural defects.

Directly adopting general deep learning methods for inspecting industrial product defects cannot achieve satisfactory performance.

To address the above imbalance and diversity problems of defects in industrial product inspection, current studies primarily focus on Unsupervised Anomaly Detection [11], [12], [13] that only uses nominal (non-defective) data as the training set. These methods can be considered as a subset of One-Class Classification (OCC) problems [20].

Nowadays, anomaly detection and localization methods are categorized into reconstruction-based methods, embedding-based methods, and synthesizing-based methods. The key idea of reconstruction-based methods is that only nominal images can be reconstructed and anomalous can not be reconstructed. Therefore, the reconstructed error can serve as an anomaly score to distinguish nominal and anomalous images. Commonly used anomaly detection Autoencoder [5], [21] and Generative Adversarial Networks (GAN) [22], are often employed for image reconstruction. The reconstruction-based methods are interpretable but unstable in performance in hard anomalous images [23], [24]. The fundamental concept of embedding-based methods involves utilizing pre-trained Deep Neural Networks (DNNs) [25] to generate representative embedding features and subsequently distinguish anomalies based on the distance between nominal and anomalous embeddings. Previously, embedding-based methods excelled in anomaly detection but struggled to accurately locate the anomalous regions. The proposals of PaDim [11] and PatchCore [12] have enabled embedding-based methods to achieve state-of-the-art performance, combining both anomaly detection and localization. The synthesizing-based methods attempt to generate fake anomalous images using nominal images. CutPaste [26] generates fake anomalies by cutting a nominal image patch and randomly pasting it onto another nominal image. Synthesizing-based methods face a limitation in accurately depicting real anomalies, because the appearance of real defects is diverse and unpredictable. Therefore, generating synthetic anomalies from nominal images can never fully represent the complexity of real anomalies. Moreover, recent research by OpenGAN [27] suggests that generating synthesized features, rather than synthesized images, leads to better model performance. This approach is advantageous because it: 1) eliminates potential noise when extracting features from synthesized images and 2) reduces the model's capacity by synthesizing in feature space. Similarly, to tackle the challenges posed by synthesized images, SimpleNet [13] proposed synthesizing anomalies in the feature space rather than on images.

Previous studies [12], [13], [14], [26] have achieved significant results on MVTec Anomaly Detection (MVTec AD) benchmark [5], which is a widely used anomaly detection public dataset covering various industrial object categories. But they [11], [12], [13] have given less consideration to the accuracy gap between public datasets [5] and real-world manufacturing industrial applications. To design an ADL model for real-world manufacturing industries, our study follows the approach of previous methods in two key aspects. (1) We address the ADL as the cold start challenge by training the network exclusively with non-defective images. Specifically, during the training step, only non-defective data is used. During the inference step, both non-defective and defective data are incorporated. (2) Our study leverages the strengths of both synthesizing-based and embedding-based methods. Furthermore, our study improves upon the previous methods in three aspects.

Firstly, the paper modifies the commonly used preprocessing strategy of state-of-the-art methods [11], [12], [13]. Sample images from public dataset [5] and real-world datasets are shown in Fig. 1. The real-world datasets are collected directly from industrial production lines. For the public dataset, it is common to observe that 1) Objects are positioned in the center of the image, 2) images have a fixed size, resolution, and square aspect ratio, and 3) backgrounds are clear and have a consistent appearance. On the contrary, real-world datasets might not possess the above characteristics. We consider that the commonly used preprocessing strategy, including resizing and center cropping, poses a potential risk by neglecting defects located in the edge regions as shown in Fig. 2 (b). Neglecting those defects results in the model's inability to fully manifest its inherent performance in real-world industrial applications. Therefore, we suggest using padding instead of center cropping to preserve the network's applicability in real-world industrial environments.

Secondly, the paper proposes a Fake Defect Feature Augmentation (FDFA) method used in the fake feature flow, to mitigate the weakness of a Data Augmentation (DA) strategy that cannot be used in a true feature flow [12], [13]. Examples of non-defect and defect images are shown in Fig. 1. The defect from one product can vary from subtle texture change as Fig. 1. (1)(c) to larger structural defects as Fig. 1. (1)(e). A data augmentation strategy [28] provides an effective solution in the data space for addressing the challenge of limited industrial data by increasing data quantity. However, the application of data augmentation strategies faces difficulties in the context of ADL. There are two primary reasons. Firstly, only non-defective data is trained in OCC manner but augmenting non-defective data may not yield significant benefits due to its sufficiency. Secondly, ADL needs prior knowledge for class-retaining augmentations [12], [13]. GraphCore [14] demonstrates the efficacy of rotation as an augmentation technique. However, GraphCore also points out that there will be more complex and realistic industrial anomaly image datasets that cannot be adequately addressed by rotation. Therefore, this paper aims to augment fake feature flow (defect) rather than real feature flow (non-defect).

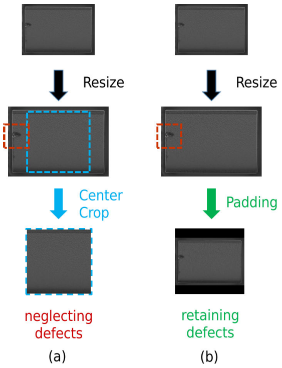


FIGURE 2. Example of (a) pre-processing strategy in previous studies and (b) pre-processing strategy in this paper on real-world datasets. Blue dash boxes: the center crop region. Red dash boxes: defect regions.

Dataset augmentation can be performed in image space and feature space [29]. Recent studies [13], [27], [29], [30] claimed that performing in feature space is better than in image space in model accuracy, robustness, and generalization. Devries and Taylor [29] demonstrated that augmenting feature space improves the performance of supervised learning algorithms. Verma et al., [30] found that augmenting feature space improves generalization and robustness in deep neural networks. OpenGAN [27] further proposed to generate fake features instead of fake images to achieve better performance in open-set recognition tasks. Similarly, our baseline method SimpleNet [13] applied this method in the ADL task by adding noise into a non-defect feature to generate a fake defect feature. Following the above ideas, this paper proposed FDFA which adds noise in feature space and then augments the noise by randomly picking its standard derivation $\tilde{\sigma}$.

Thirdly, this study proposed to employ a Hardness-aware Cross-entropy Loss (HCELoss), which demonstrates superior performance in terms of both accuracy and stability. Previous works used cross-entropy loss [14] or $L1$ loss [11], [12], [13] in industrial anomaly detection. The cross-entropy loss remains stable but fails to achieve higher accuracy compared to the $L1$ loss. Conversely, the $L1$ loss achieves better accuracy but exhibits instability. Detailed experimental results and discussion are presented in Fig. 4 and Sec. V.A.

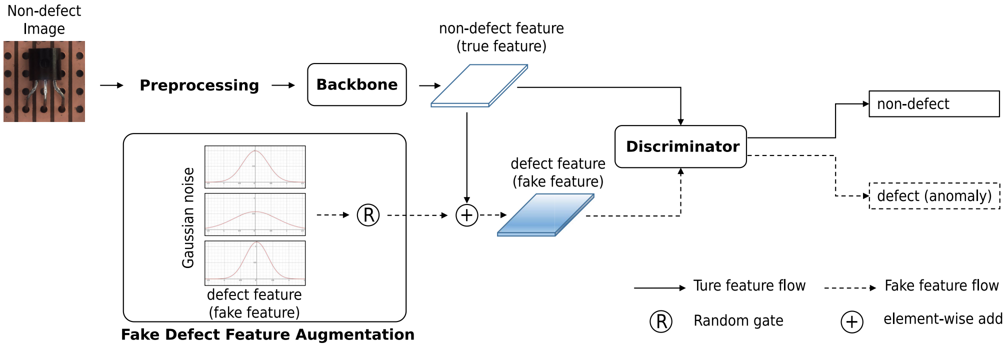


FIGURE 3. The illustration of the proposed Reliable Anomaly Detection and Localization (RADL) network. A backbone network extracts feature maps. An Augmented Anomalous Feature Generator (FDFA) generates fake defect features. A discriminator with HCELoss is trained to distinguish non-defect features and fake defect features. The FDFA will not be used in the test step.

More specifically, the $L1$ loss causes unpredicted large fluctuations in three evaluation metrics and loss value. It is time-wasting to evaluate and compare model performance in every training epoch to select the optimal model. Moreover, there is no labeled ground-truth annotated products in real-world factory environments for calculating the model's accuracy. Therefore, the existing anomaly detection methods are not suitable and applicable in real-world industrial manufacturing applications. A stable loss function is necessary for building a reliable ADL application. To bridge the gap between academic research and real-world industrial manufacturing, HCELoss is utilized in this paper and demonstrates its importance in real-world applications. The results demonstrate the high accuracy achieved by our proposed HCELoss in the public dataset MVTec AD, and its necessity in real-world industrial applications.

The architecture of the proposed unsupervised anomaly detection and localization methods RADL is illustrated in Fig. 3. Non-defect (nominal) images are pre-processed and then fed into an ImageNet [25] pre-trained backbone network to extract their true features. Augmented fake defect (anomaly) features are generated by adding augmented Gaussian noise $\mathcal{N}$ with a random standard deviation $\tilde{\sigma}$ to true non-defect (nominal) features. A discriminator is used to find defect images by training with both true non-defect and fake defect features. The FDFA is only used in training and is removed in inference.

The contributions of this work are fivefold,

1) To better align with real-world manufacturing environments, this paper has refined the commonly used preprocessing strategy.
2) To enhance the generalization capabilities of the proposed RADL, this paper introduces the Fake Defect Feature Augmentation (FDFA) strategy.
3) To enhance the stability of RADL in real-world industrial environments, this paper introduces the Hardness-aware Cross-Entropy Loss (HCELoss).
4) Demonstrating superior results, the RADL achieved 99.53% in I-AUROC(%), 97.85% in P-AUROC(%), and 91.60% in PRO(%) on the public dataset MVTec AD benchmarks.
5) Pushing academic research closer to real-world manufacturing industry applications, the paper evaluates RADL's efficacy in real-world manufacturing scenarios using two real-world datasets collected from production lines.

The remainder of this paper is organized as follows. The whole architecture of the proposed RADL is introduced in Section II. Section III introduces the details of the experiment setting, including datasets, evaluation metrics, and implementation details. Section IV reports the experimental results on the public MVTec AD benchmarks and two real-world industrial manufacturing datasets. Section V concludes the work.

II. METHOD

The architecture of Reliable Anomaly Detection and Localization (RADL) is shown in Fig. 3. Nominal images are pre-processed and processed by the backbone network to extract features. FDFA generates augmented fake defect features by adding augmented Gaussian noise to non-defect (nominal) features. The Discriminator is used to distinguish defects.

A. IMAGE PREPROCESSING

As shown in Figure 2(a), previous studies [12], [13] achieved enhanced model performance on the public dataset [5] by utilizing center cropping, effectively eliminating the surrounding background. The success of center cropping on public datasets MVTec AD [5] can be attributed to the fixed and square shape of the images, as well as the positioning of target objects at the center of these images.

However, there are images in real-world scenarios that are varied and non-square in shape, and the target object is not necessarily located at the center of the images. As a consequence, the prior preprocessing strategy causes the model to neglect defects located in the edge regions in real-world inspection applications. The example is illustrated in Fig. 2(a).

To accommodate real-world manufacturing environments, we suggest using padding instead of performing center cropping, as shown in Fig. 2(b). By padding the resized images, the complete image can be retained while still achieving a consistent input size required for subsequent processing steps. This modification ensures that the preprocessing strategy remains compatible with real-world manufacturing environments and preserves the integrity of the data for accurate defect inspection. The visualization results presented in Fig. 6 illustrate that the location of anomaly regions can be detected accurately with our proposed preprocessing strategy on real-world manufacturing industrial datasets.

B. BACKBONE

In the training phase, given an input image x, the pre-trained backbone network $B(\cdot)$ extracts features from different k layers. The classification network WideResnet50 [15] is used here as the backbone model. The feature map from level $k \in K$ is denoted as $f_k \in \mathbb{R}^{H_k, W_k, C_k}$, where H_k, W_k, and C_k are the height, width, and channel size of the feature map f_k.

Following CutPaste [26] and PatchCore [12], an image is cut into $p \times p$ patches then using its patch features to represent the whole image. For an f_k at location (h, w), its feature map with patch size $p \in P$ is defined as,

$$f_k^{p+} = \{(a, b) | a \in [h - \lfloor p/2 \rfloor] \ldots, [h + \lfloor p/2 \rfloor], \\ b \in [w - \lfloor p/2 \rfloor] \ldots, [w + \lfloor p/2 \rfloor]\}, \tag{1}$$

The second and third intermediate layers ($k = \{2, 3\}$) of the backbone are aggregated to generate non-defect features using an average pooling operator as,

$$f^{p+} = AvgPooling(f_{k=2}^p, Resize(f_{k=3}^p)), \tag{2}$$

where $Resize(\cdot)$ is an interpolation operator to enlarge size of $f_{k=3}^p$ to the same size with $f_{k=2}^p$.

C. FAKE FEATURE FLOW

The solution of RADL is to exploit real non-defective samples and synthetic defective samples to train the Discriminator $\mathcal{D}(\cdot)$. The fake feature flow generates and augments synthetic defective features as shown in Fig. 3.

1) FAKE DEFECT FEATURE GENERATION

As mentioned in [12] and [13], data augmentation is not easy to apply in the true flow because the prior knowledge about class-retaining augmentations is not provided in MVTec AD [5]. Therefore, our study focuses on the fake feature flow.

Moreover, previous research proves that generating synthetic defect features is more efficient than generating synthetic defect images [13], [27]. Following the above ideas, our study adds noise into the feature space of non-defect samples to synthesize defect features.

Liu et al. [13] element-wise added a Random Gaussian Noise $\mathcal{N}(\mu, \sigma^2)$ with fixed mean μ and standard deviation σ on non-defect features f^{p+} as follows,

$$f^{p-} = f^{p+} + \mathcal{N}(\mu, \sigma^2), \tag{3}$$

where f^{p-} is the synthesized fake defect features corresponding to real non-defect features f^{p+}. All these three features f^{p-}, f^{p+} and $\mathcal{N}$ are 1536-dimensional vectors. The choice of 1536 as the dimension for the final layer of feature fusion is manually determined. $\mu = 0$ and $\sigma = 0.015$ are pre-defined and proven that can achieve the best performance in their research [13].

2) FAKE DEFECT FEATURE AUGMENTATION (FDFA)

As mentioned in Sec. I, the defect from one product can vary from subtle texture change to larger structural defects. The naive $\mathcal{N}$ cannot represent diverse defects. In other words, the naive $\mathcal{N}$ is too easy for Discriminator $\mathcal{D}(\cdot)$ to synthesize generalized features of defects. Our study tries to simulate diverse defects by introducing more varied noise.

Based on the above ideas, we propose a Fake Defect Feature Augmentation (FDFA) method, which is performed in the fake feature flow to augment fake features. To validate our insight, we have adapted SimpleNet [13] as the baseline method for our model.

Intuitively, we extended this idea to feature-based data augmentation. The augmented and synthesized fake defect feature is generated by adding an augmented Gaussian Noise $\mathcal{N}(\mu, \tilde{\sigma}^2)$ on non-defect features f^{p+}, where $\tilde{\sigma}$ represents a random standard deviation value in range $[\sigma_{low}, \sigma_{high}]$ as follows,

$$\tilde{f}^{p-} = f^{p+} + \mathcal{N}(\mu, \tilde{\sigma}^2), \tag{4}$$

where $\tilde{f}^{p-}$ is the augmented synthesized fake defect features corresponding to real non-defect features f^{p+}. With the help of $\tilde{\sigma}$, this study diverse and varied synthesized fake defect features are generated.

D. DISCRIMINATOR WITH HARDNESS-AWARE CROSS-ENTROPY LOSS (HCELoss)

RADL trains a discriminator to recognize the non-defect and defect images using the real non-defect features and fake defect features. The discriminator enforces the real non-defect features $f^{p+} \rightarrow 1$ and fake defect features $f^{p-} \rightarrow 0$.

As mentioned in Sec. I, the cross-entropy loss is stable but cannot achieve better accuracy than $L1$ loss. On the contrary, the $L1$ loss is able to achieve good accuracy but is unstable during the whole training procedure. This study employs a Hardness-aware Cross-entropy Loss (HCELoss) to achieve good performance in terms of both accuracy and stability.

The research [31] demonstrated that temperature value in contrastive loss plays a role in controlling the strength of penalties on the hard samples. The hardness-aware property is also observed in softmax-based loss functions. Therefore, our idea is to add the hardness-aware property to the commonly used softmax-based binary cross-entropy loss by adding the temperature value τ.

Hence, the loss of non-defect features f^{p+} is computed as,

$$loss^{+} = -[y^{+} * log(f^{p+}/\tau) + (1 - y^{+}) * log(1 - f^{p+}/\tau)], \tag{5}$$

The loss of defect features f^{p-} is computed as,

$$loss^{-} = -[y^{-} * log(f^{p-}/\tau) + (1 - y^{-}) * log(1 - f^{p-}/\tau)], \tag{6}$$

where $y^{+} = 1$ denotes the labels for real non-defect sample, and $y^{-} = 0$ denotes the labels of the synthesized fake defect features.

The total loss for non-defect images and defect features is computed as,

$$HCEloss = loss^{+} + loss^{-} \tag{7}$$

III. EXPERIMENTAL SETUP

A. DATASET

To perform comprehensive studies and evaluate the proposed RADL in real-world manufacturing industrial applications, we employ three datasets in total, including one public dataset MVTec-AD and two real-world manufacturing industrial datasets.

1) ONE PUBLIC DATASET MVTec-AD

The MVTec-AD dataset [5] contains image classes that are typically found in an industrial setting. It contains 5 classes belonging to textures (Carpet, Grid, Leather, Tile, and Wood) and 10 classes belonging to objects (Bottle, Cable, Capsule, Hazelnut, Metal, Nut, Pill, Screw, Toothbrush, Transistor, and Zipper). Example images are illustrated in Fig. 1(a). MVTec-AD contains 3629 training images and 467 normal images and 1258 abnormal images in the test set. Following previous works, the training set only consists of an amount of nominal (no defect) images. Nominal images are treated as positive data. The trained model is tested on both normal and abnormal images of the same class, where abnormal images are treated as negative data.

2) TWO REAL-WORLD MANUFACTURING INDUSTRIAL DATASETS

a: THE HEAT STAKING POINTS (HSP) DATASET

The HSP [4] contains door trim images. The training set contains 200 nominal heat staking point images. The test set contains 50 nominal and 50 anomaly images. Examples of non-defect (nominal) and defect (anomaly) heat staking point images are shown in Fig.1(4). In HSP, target objects are not always located at the center of images.

b: THE REAL-WORLD SMALL-PART (SP) DATASET

The training set contains 200 small-part images. The test set contains 50 nominal and 50 anomaly images. The original resolution of each image is different, ranging from 658 × 457 to 673 × 468. Examples of non-defect (nominal) and defect (anomaly) small-part images are shown in Fig. 1(5). The SP dataset is used to evaluate the RADL performance in tiny defect detection and localization.

B. EVALUATION METRICS

1) I-AUROC(%)

Image-level anomaly detection performance, which is computed via the standard Area Under the Receiver Operator Curve (AUROC) to measure the classification performance of good samples and anomalies.

2) P-AUROC(%)

pixel-wise anomaly segmentation performance, which is computed via AUROC to measure the segmentation and localization performance of good samples and anomalies.

3) PRO(%)

Pre-Region Overlap (PRO) is a region-level metric, which is more capable than I-AUROC and P-AUROC in assessing the ability of fine-grained anomaly detection. The PRO score takes into account the overlap and recovery of connected anomaly components to better account for varying anomaly sizes.

C. IMPLEMENTATION DETAILS

The ImageNet pre-trained Wide-Resnet50 [15] is used as the backbone, and the second and third intermediate layers of the backbone are used for feature fusion. We follow a similar experiment setting with baseline method [13]. All images are resized to 256 × 256. The dimension of the last layer of feature fusion is set to 1536. The discriminator consists of a linear layer, a batch normalization layer, a Leaky-Relu activation layer, and a linear layer. The Adam optimizer is used with a learning rate of 0.0002 and weight decay of 0.00001. Training epochs are set to 40 for each dataset and batch size is 8. The number of iterations varies for each dataset depending on the number of images in the dataset and the training batch size. For MVTec [5], HSP, and SP datasets, the number of training iterations is 1800, 1000, and 1000, respectively.

The experiment is performed on an Nvidia GeForce GTX 3080Ti GPU and a 12th Gen Intel® Core^{TM} i7-12700K × 20 CPU. The training time for 15 classes on AD MVTec is about 15 hours. The training time for HSP and SP is about 30 minutes The inference time for one image is about 77 FPS.

IV. EXPERIMENTAL RESULT

A. ABLATION STUDIES

We report accuracy and reliability analysis for the HCELoss and the FDFA to verify these two methods are reasonable and more suitable solutions in the real-world manufacturing industry. Additionally, the reliability analysis of methods is measured by training loss in Fig. 4.

TABLE 1. The comparison experimental results of three settings. Setting A: Baseline method [13], which contains backbone network, the fake defect feature generation, and the discriminator with $L1$ loss. Setting B (this paper): Baseline + HCELoss. Setting C (this paper): Baseline + HCELoss + FDFA. The best-performing method is in bold.

	Setting A (baseline)			Setting B (this paper)			Setting C (this paper)		
Defect Classes	I-AUROC	P-AUROC	PRO	I-AUROC	P-AUROC	PRO	I-AUROC	P-AUROC	PRO
Carpet	99.48	97.92	86.47	99.52	**98.06**	**86.52**	**99.64**	98.00	86.07
Grid	99.50	97.63	92.01	99.25	**98.81**	**95.17**	**99.67**	97.85	94.69
Leather	**100.0**	99.2	96.86	**100.0**	**99.26**	97.05	**100.00**	99.25	**97.10**
Tile	99.86	**95.85**	87.29	**99.93**	95.01	**89.58**	99.82	95.31	89.17
Wood	**100.0**	93.73	**83.56**	**100.0**	93.50	82.86	**100.0**	**93.76**	83.44
Avg. Texture	99.77	96.87	89.24	99.74	**96.93**	**90.24**	**99.83**	96.83	90.09
Bottle	**100.0**	97.97	90.12	99.92	97.92	90.32	99.92	**97.98**	**90.79**
Cable	**99.94**	97.40	90.34	99.76	97.76	**91.58**	99.79	**97.77**	91.40
Capsule	**97.69**	98.90	**93.45**	97.13	**98.94**	**92.81**	**96.97**	**98.92**	93.31
Hazelnut	99.75	97.56	82.34	**99.82**	97.72	**83.64**	**99.82**	**97.75**	82.9
Metal Nut	**100.0**	98.56	88.69	**100.0**	98.88	89.70	**100.0**	**98.92**	**90.73**
Pill	**98.83**	98.24	94.20	98.72	**98.64**	**94.94**	98.52	98.59	94.58
Screw	98.48	99.23	96.29	98.73	**99.27**	**96.51**	**99.12**	99.26	96.25
Toothbrush	99.72	98.52	93.04	**100.0**	97.9	91.56	99.72	**98.53**	**93.35**
Transistor	**100.0**	96.72	**94.00**	**100.0**	96.69	93.68	**100.0**	**97.03**	93.91
Zipper	99.90	98.87	95.94	**99.92**	**98.89**	96.07	99.87	98.86	**96.35**
Avg. Object	**99.43**	98.20	91.84	99.40	98.26	92.07	99.37	**98.36**	**92.36**
Average	99.50	97.75	90.97	99.51	97.82	91.47	**99.53**	**97.85**	**91.60**

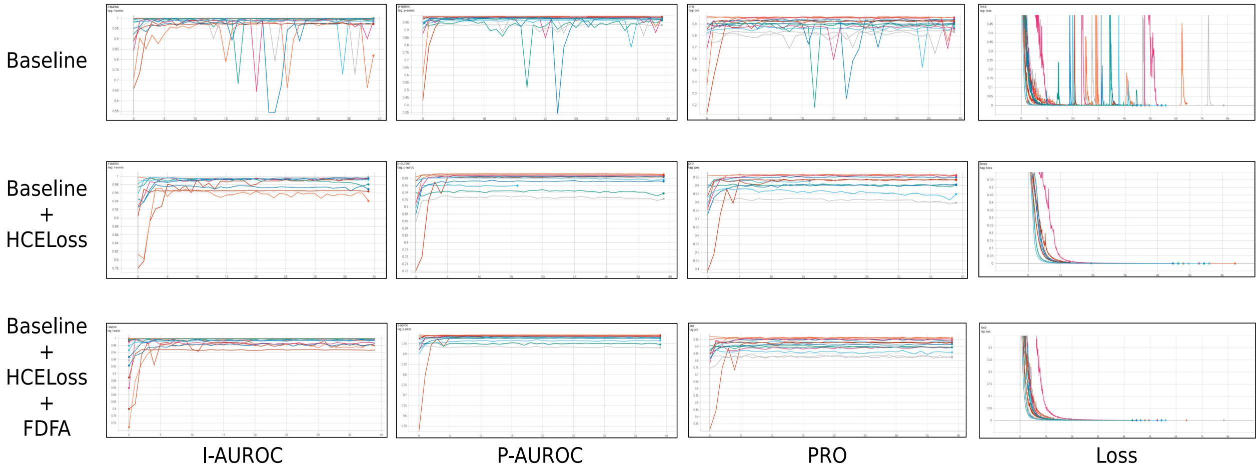


FIGURE 4. The change of three evaluation metrics I-AUROC, P-AUROC, PRO, and loss in each epoch of three settings. Each column shows one evaluation metric, and each row shows Setting A, Setting B, and Setting C. Lines with different color denotes each class of the MVTec AD dataset. Setting A: Baseline method [13], which contains backbone network, the fake defect feature generation, and the discriminator with $L1$ loss. Setting B (this paper): Baseline + HCELoss. Setting C (this paper): Baseline + HCELoss + FDFA. Among them, RADL holds the most reliable result because it can ensure network convergence to the flattest and most stable state.

TABLE 2. Performance on MVTec AD dataset with varied τ in Eq. (3).

τ	I-AUROC	P-AUROC	PRO
1	**99.5**	**97.8**	91.0
0.5	**99.5**	**97.8**	**91.5**
0.25	**99.5**	**97.8**	91.4
0.1	99.4	**97.8**	91.4

1) THE EFFECTIVENESS OF FDFA AND HCELoss

The effectiveness of FDFA and HCELoss in terms of anomaly detection and localization accuracy can be demonstrated in Table 1. **Setting A** indicates the baseline method [13], which contains the backbone network, the fake defect feature generation, and the discriminator with L1 loss. **Setting B** indicates the baseline method using our proposed HCELoss in discriminator. **Setting C** indicates the baseline method using our proposed HCELoss and FDFA. The highest scores are bold.

Specifically, with the help of FDFA and HCELoss, the I-AUROC, P-AUROC, and PRO of the baseline method in MVTec AD [5] are improved from 99.50% to 99.53%, from 97.75% to 97.5%, and from 90.97% to 91.60%.

The effectiveness of FDFA and HCELoss in terms of system reliability can be demonstrated in Fig. 4. Lines with different color denotes each class of the MVTec AD dataset. The network was trained for 40 epochs for each defect class. It is observed that the baseline method exhibits significant and unpredictable fluctuations in the three evaluation metrics and loss. More specifically, the performances suddenly dramatically decreased to 54.0% in I-AUROC, to 34.0% in P-AUROC, and to 25.0% in PRO in epoch 22.

TABLE 3. Comparison of RADL with state-of-the-arts works on MVTec AD. Image-wise AUROC (I-AUROC)(%), pixel-wise AUROC(%), and PRO(%) are shown for every defect class. The best-performing method is in bold.

	PaDim [11]		PatchCore [12]		SimpleNet [13]			RADL(this paper)		
Classes	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	PRO	I-AUROC	P-AUROC	PRO
Carpet	**99.8**	99.4	98.7	99.0	99.48	97.92	**86.47**	99.64	98.00	86.07
Grid	93.7	97.3	98.2	**98.7**	99.50	97.63	92.01	**99.67**	97.85	**94.69**
Leather	**100.0**	99.2	**100.0**	**99.3**	**100.0**	99.20	96.86	**100.00**	99.25	**97.10**
Tile	98.1	94.1	98.7	95.6	**99.86**	**95.85**	87.29	99.82	95.31	**89.17**
Wood	99.2	94.9	99.2	**95.0**	**100.0**	93.73	**83.56**	**100.0**	93.76	83.44
Avg. Texture	95.5	96.9	99.0	**97.5**	99.77	96.87	89.24	**99.83**	96.83	**90.09**
Bottle	99.1	98.3	**100.0**	**98.6**	**100.0**	97.97	90.12	99.92	97.98	**90.79**
Cable	97.1	96.7	99.5	**98.4**	**99.94**	97.40	90.34	99.79	97.77	**91.40**
Capsule	87.5	98.5	**98.1**	98.8	97.69	98.90	**93.45**	96.97	**98.92**	93.31
Hazelnut	99.4	98.2	**100.0**	**98.7**	99.75	97.56	82.34	99.82	97.75	**82.9**
Metal Nut	96.2	97.2	**100.0**	98.4	**100.0**	98.56	88.69	**100.0**	**98.92**	**90.73**
Pill	90.1	95.7	96.6	97.4	**98.83**	98.24	94.20	98.52	**98.59**	**94.58**
Screw	97.5	98.5	98.1	**99.4**	98.48	99.23	**96.29**	**99.12**	99.26	96.25
Toothbrush	**100.0**	**98.8**	100.0	98.7	99.72	98.52	93.04	99.72	98.53	**93.35**
Transistor	94.4	**97.5**	**100.0**	96.3	**100.0**	96.72	**94.00**	**100.0**	97.03	93.91
Zipper	98.6	98.5	**99.4**	98.8	**99.90**	**98.87**	95.94	99.87	98.86	**96.35**
Avg. Object	96.0	97.8	99.2	**98.4**	**99.43**	98.20	91.84	99.37	98.36	**92.36**
Average	95.8	97.5	99.1	**98.1**	99.50	97.75	90.97	**99.53**	97.85	**91.60**

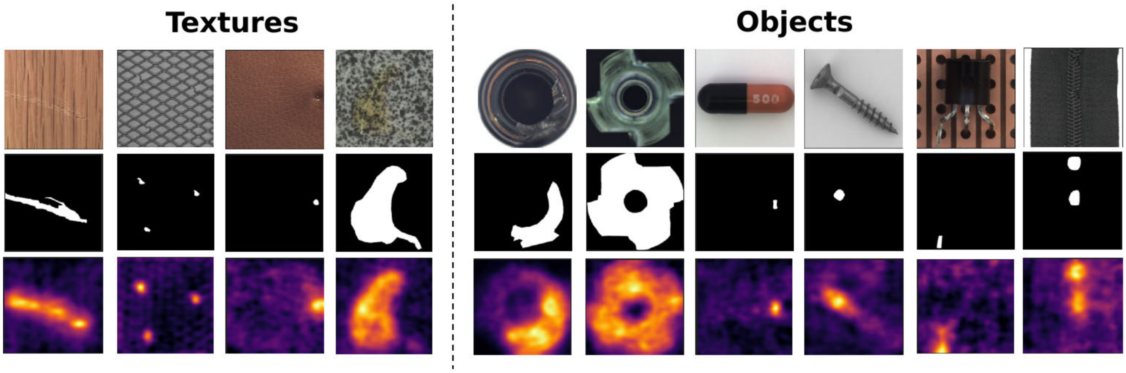


FIGURE 5. Visualization of anomaly detection and localization results in Public MVTec AD Dataset.

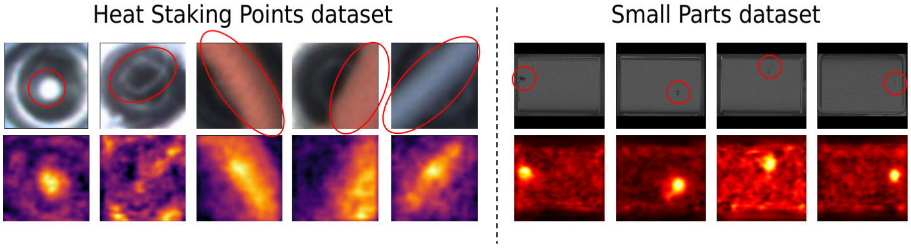


FIGURE 6. Visualizationn result in real-world manufacturing industrial datasets. Defect regions are circled by red ellipses (curves).

Such fluctuations in accuracy and loss highlight the unreliability of the baseline method in real-world manufacturing industrial applications, as workers may stop training the network at a bad performance point. Additionally, labeled ground-truth annotations may not be available to compute model accuracy n real-world manufacturing industrial environments. Hence, workers cannot rely on the evaluation metrics or loss values to assess the network's performance and choose the model at its optimal point. In such scenarios, the importance of having a stable model becomes paramount for real-world manufacturing industries.

In contrast to the "Baseline" method, our proposed approach "Baseline + HCELoss + FDFA" achieves convergence in all defect classes for the three evaluation metrics and loss within 10 epochs. Moreover, our method maintains a stable state after convergence, ensuring consistent performance.

TABLE 4. Comparison in model size and inference time.

Methods	Inference FPS
PaDim [11]	1
PatchCore [12]	10
SimpleNet [13]	**77**
RADL (this paper)	**77**

The experimental results provide evidence of the effectiveness of the proposed HCELoss and FDFA in constructing a reliable anomaly detection and localization network suitable for real-world applications.

2) ANALYSIS OF HARDNESS-AWARE HYPER-PARAMETER τ

In Table 2, we investigate the effect of the temperature fact τ in Eq. (5) and Eq. (6). When $\tau = 1$, HCELoss is a normal cross-entropy loss. The best results are produced when $\tau = 0.5$. The performance remains consistent when $\tau \leq 0.5$, indicating that τ is a quite robust hyper-parameter.

B. COMPARISON ON PUBLIC DATASET MVTec AD

1) ACCURACY

This study conducts a comparative analysis between the proposed RADL and state-of-the-art methods on the public dataset MVTec AD [5]. Experiment results of SimpleNet [13] is re-implemented in our local environment using their open-source code, with a fixed seed for experiment consistency. Table 3 summarizes the comparisons. For texture defect (Avg. Texture), the proposed RADL achieves the best average performance with 99.83% in I-AUROC(%), 96.83% in P-AUROC(%), and 90.09% in PRO(%). For object defect (Avg. Object), the proposed RADL achieves the best average performance with 99.37% in I-AUROC(%), 98.36% in P-AUROC(%), and 92.36% in PRO(%). For all 15 classes of defects in MVTec AD, the proposed RADL achieves the best average performance with 99.53% in I-AUROC(%), 97.85% in P-AUROC(%), and 91.60% in PRO(%).

2) INFERENCE TIME

Training inference time is also an important factor for manufactory industrial model deployment. The inference time of our model and state-of-the-art methods on inference time is reported in Table 4. All the methods are evaluated on our local PC with the same hardware configuration as mentioned in Sec III-C. Our method has the same inference time as SimpleNet [13] because our proposed modules are only used in training and are removed at inference. It is noteworthy that our proposed HCELoss and FDFA improve network accuracy without adding any inference time. The RADL can work in real-time manufacturing applications.

3) VISUALIZATION

Visualization results of the proposed RADL on MVTec AD are shown in Fig. 5. Each column represents a different defect class. The second row shows binary ground truth maps (labels), where black indicates non-defect regions and white indicates anomaly regions. The third row shows predicted defect (anomaly) maps. The anomaly region of the low anomaly score is colored black, and the high anomaly score is colored orange. The results demonstrate that RADL can locate defect regions accurately on various textures and objects.

C. EXPERIMENTS ON REAL-WORLD DATASETS

To evaluate the model performance in real-world applications, we perform quantitative experiments on two real-world manufacturing industrial datasets, which are collected from the production line of factories.

Visualization results of the proposed RADL on real-world Manufacturing Industry Datasets, i.e., HSP and SP, are shown in Fig. 6. Sample images and predicted defect (anomaly) maps are shown in the first and second rows, respectively. The defect region is circled by red ellipses (curves). The RADL accurately detects defect locations even when the input image sizes vary. RADL can effectively identify tiny defective regions in SP.

V. CONCLUSION

This study identifies the existing gap between academic research and real-world manufacturing industry applications. In response to this gap, an accurate and reliable anomaly detection and localization system called RADL is introduced. RADL is specifically designed to address the challenges of real-world industrial product quality inspection. Compared with previous research, the performance achieved on the MVTec AD dataset and real-world manufacturing industrial datasets validate the effectiveness of RADL. The proposed inspection system holds the potential to enhance the efficiency of real-world manufacturing industry processes.

REFERENCES

[1] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," 2023, *arXiv:2301.11514*.

[2] W. Huang and P. Wei, "A PCB dataset for defects detection and classification," 2019, *arXiv:1901.08204*.

[3] S. Liao, C. Huang, H. Zhang, J. Gong, M. Li, and Z. Wang, "Object detection of welding defects in SMT electronics production based on deep learning," in *Proc. 23rd Int. Conf. Electron. Packag. Technol. (ICEPT)*, Aug. 2022, pp. 1–5.

[4] H. Jung and J. Rhee, "Application of YOLO and ResNet in heat staking process inspection," *Sustainability*, vol. 14, no. 23, p. 15892, Nov. 2022.

[5] 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. (CVPR)*, Jun. 2019, pp. 9584–9592.

[6] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020.

[7] Q. Tang, G. Cao, and K.-H. Jo, "Integrated feature pyramid network with feature aggregation for traffic sign detection," *IEEE Access*, vol. 9, pp. 117784–117794, 2021.

[8] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[9] X. Jiang, G. Xie, J. Wang, Y. Liu, C. Wang, F. Zheng, and Y. Jin, "A survey of visual sensory anomaly detection," 2022, *arXiv:2202.07006*.

[10] Y. Chen, Y. Ding, F. Zhao, E. Zhang, Z. Wu, and L. Shao, "Surface defect detection methods for industrial products: A review," *Appl. Sci.*, vol. 11, no. 16, p. 7657, Aug. 2021.

[11] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *ICPR Workshops*, 2020, pp. 475–489.

[12] 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. (CVPR)*, Jun. 2022, pp. 14298–14308.

[13] 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)*, Jun. 2023, pp. 20402–20411. [Online]. Available: https://api.semanticscholar.org/CorpusID:257766673

[14] 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*.

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

[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[17] X. Feng, X. Gao, and L. Luo, "A ResNet50-based method for classifying surface defects in hot-rolled strip steel," *Mathematics*, vol. 9, no. 19, p. 2359, Sep. 2021.

[18] Q. Wan, L. Gao, and X. Li, "Logit inducing with abnormality capturing for semi-supervised image anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[19] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[20] P. Perera, P. Oza, and V. M. Patel, "One-class classification: A survey," 2021, *arXiv:2101.03064*.

[21] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714. [Online]. Available: https://api.semanticscholar.org/CorpusID:102353587

[22] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," 2018, *arXiv:1805.06725*.

[23] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2893–2901. [Online]. Available: https://api.semanticscholar.org/CorpusID:84186723

[24] V. Zavrtanik, M. Kristan, and D. Skoaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognit.*, vol. 112, Jan. 2020, Art. no. 107706. [Online]. Available: https://api.semanticscholar.org/CorpusID:225114154

[25] 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.*, Jun. 2009, pp. 248–255.

[26] 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. (CVPR)*, Jun. 2021, pp. 9659–9669.

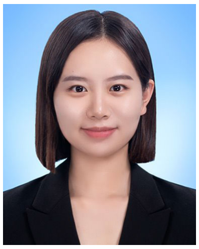
[27] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 793–802.

[28] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[29] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," 2017, *arXiv:1702.05538*.

[30] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–6. [Online]. Available: https://api.semanticscholar.org/CorpusID:59604501

[31] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2495–2504.



QING TANG (Member, IEEE) received the bachelor's degree in vehicle engineering from the School of Automotive Engineering, Shanghai University of Engineering Science, Shanghai, in 2015, and the Ph.D. degree in electrical and computer engineering from the Graduate School of Electrical Engineering, University of Ulsan, Ulsan, South Korea, in 2022.

She is currently a Senior Researcher with the Data Analysis Team, INTERX, South Korea. Her current research interests include computer vision, machine learning, intelligent manufacturing, surveillance systems, and transportation systems.

HAIL JUNG (Member, IEEE) is currently an Assistant Professor with the Seoul National University of Science and Technology. He is also a founding member and the CTO of INTERX, a manufacturing AI solution provider. He has served ad-hoc referee for multiple academic journals, such as the *Journal of Business Research*.