Int. J. of Precis. Eng. and Manuf.-Green Tech.
https://doi.org/10.1007/s40684-026-00871-z

Online ISSN 2198-0810
Print ISSN 2288-6206

REGULAR PAPER

# Dual-Encoder Generative Anomaly Detection for Casting Surface Defect Detection

**Seongmin Kwak**[1] · **Sooyoung Lee**[1,2]

Received: 5 December 2025 / Revised: 19 January 2026 / Accepted: 24 February 2026
© The Author(s) 2026

**Abstract**

The detection of surface defects in casting materials presents significant challenges due to coarse textures, irregular patterns, and class imbalance, which hinder the development of highly reliable automatic classification methods. To address these challenges, this paper proposes a dual-encoder-based generative anomaly detection framework using only normal data. Specifically, the proposed method is designed to integrate pixel responses, critic intermediate features, and latent representations into a unified anomaly score and to enforce consistency across input-reconstruction-latent spaces. A dual-encoder architecture combined with a Wasserstein GAN with gradient penalty backbone is proposed to effectively learn and align these multi-perspective representations. Both quantitative and qualitative results demonstrate that the proposed model achieves superior performance in defect classification and reliability, achieving an Area under the receiver operating characteristic curve of 0.998 and an accuracy of 99.1%. Balanced values in precision, recall, and F1-score indicates effective suppression of false positives and prevention of missed detections, showing clear improvements compared to baseline models. This study not only enables precise detection of surface defects in casting processes but also holds potential to contribute to anomaly detection across diverse manufacturing processes.

**Keywords** anomaly detection · deep learning · surface defect detection · unsupervised learning

## 1 Introduction

Mechanical defects play a critical role in determining product performance and safety throughout manufacturing and design processes [1–3]. These defects can be categorized into various forms, such as cracks, voids, surface defects, and shape deviations occurring during manufacturing processes. They have attracted considerable attention to various manufacturing processes, including additive manufacturing, casting, and welding [4–9]. For example, casting defects arise when the molten metal solidifies unevenly as a result of variable cooling rates. These initial flaws serve as stress concentrators under cyclic loading, which accelerates the accumulation of fatigue damage [10]. Therefore, precise detection and control of potential defects are essential to maintain structural integrity and ensure functional reliability of the product.

While accurate detection of mechanical defects is critical to maintaining product quality, challenges remain due to their complexity and diversity of defect types. Conventional approaches are typically categorized into non-destructive testing and classical machine vision methods. First, non-destructive testing methods are designed to detect both internal and surface defects without damaging the product. However, its dependence on the proficiency of the inspector results in high costs and slow inspection speed, which require substantial professional experience to ensure consistency of the results [11–13]. Second, classical machine vision algorithms, such as thresholding and edge detection, can rapidly detect defects using relatively simple rules. Nevertheless, their reliance on predefined criteria reduces the detection accuracy of small defects and complex geometries, and requires strict control of illumination and surface conditions for reliable application [14–16].

To overcome the challenges mentioned above, deep learning (DL) as a data-driven approach has recently been applied to the field of mechanical defect detection. Generally speaking, DL-based approaches are designed to hierarchically

✉ Sooyoung Lee
lees@cau.ac.kr

1 School of Mechanical Engineering, Chung-Ang University, Seoul, Republic of Korea

2 School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea

Published online: 20 March 2026

extract and learn abstract-level features through deeply constructed neural networks. With large-scale datasets and advances in DL architectures, they have demonstrated high accuracy even for subtle surface defects and complex geometries. Previous studies can be broadly classified into two primary approaches: supervised learning and unsupervised learning. For example, as a supervised approach, Lin et al. [17] introduced a deep convolutional neural network (CNN) augmented with a visual attention mechanism for detecting casting defects. Besides, Gupta et al. [18] developed an ensemble of a pretrained model and a lightweight CNN model, achieving 98.18% accuracy for industrial casting images. In the unsupervised domain, Chamberland et al. [19] proposed a clustering-based defect detection method that utilizes the reconstruction error of a CNN autoencoder trained on normal casting images, reaching 88.52% accuracy. Although these studies demonstrate the potential of DL for casting defect detection, several limitations hinder their practical application in industrial environments. Supervised learning approaches are constrained by the significant cost and time required to collect large-scale labels, severe data imbalance, and their limited capability with respect to unpredictable defect cases. On the other hand, unsupervised approaches for anomaly detection generally rely on learning representations of normal samples without explicit defect labels. However, their effectiveness is often limited by the difficulty of capturing complex surface variations and by their dependence on low-level reconstruction errors, which can fail to highlight subtle defects such as fine cracks or pores.

To address these limitations, this study proposes a DL-based approach based on the latent space representation of normal data to enable accurate detection of mechanical defects and reliable interpretation.

Specifically, the proposed framework is designed to accurately detect defects by leveraging learned representations of normal casting surface data, while simultaneously offering interpretability into the correlations among pixel-level features, perceptual characteristics, and the latent space. To ensure stable representation learning and establish a reliable threshold for defect detection, we propose a dual-encoder architecture based on a pre-trained Wasserstein generative adversarial network (GAN) with gradient penalty [20, 21], which creates a consistent mapping between the input, its reconstruction, and the latent space. The defect classification performance of the proposed method is quantitatively evaluated using the area under the receiver operating characteristic (AUROC) [22] curve and standard classification metrics. Additionally, defect detection results are qualitatively assessed by visualizing the differences between generated and original images as error maps. The results indicate that the proposed method effectively identifies diverse defects and remains robust under challenging variations in surface conditions.

The remainder of this paper is organized as follows. Section 2 conveys the data description as well as our proposed method including model architecture and training strategy. Section 3 presents quantitative and qualitative analyses of the comparative models for surface defect detection. Finally, Sect. 4 concludes the paper.

## 2 Materials and Methods

### 2.1 Data Description

In this study, we consider a casting product dataset obtained from real-world manufacturing processes. We mainly utilize a top-view 256×256 image dataset of a casting impeller product, which highlights the coarse surface texture and irregular microscopic defects characteristic of the casting process [23]. For training, we utilize a total of 2,875 normal samples for unsupervised anomaly detection. The evaluation set consists of 3,138 normal and 4,211 defective samples. Among them, 600 samples from each class (approximately 20%) are randomly selected in a class-balanced manner for decision threshold determination, while the remaining samples are used for final performance assessment. In unsupervised anomaly detection, defective samples are not required during training, and the training set size is determined by the data needed for stable model convergence. Given the availability of a sufficient number of defective samples, we aim to enable a more rigorous evaluation across diverse anomaly types. To mitigate evaluation bias, we make efforts to keep the.

number of normal samples in the test set comparable to that of defective samples, resulting in a test set larger than the training set [24–26]. In terms of data preprocessing, the entire dataset is normalized to the range $[-1, 1]$ during preprocessing.

To analyze the distributional characteristics of the dataset, normal and defective casting surface cases are projected onto a two-dimensional space. Figure 1(a) presents representative examples of normal and defective images from the dataset. As shown in the figure, the defect cases exhibit surface irregularities and subtle texture deviations compared to the normal samples, highlighting the challenges in distinguishing fine-grained defects on casting surfaces. On the other hand, we employ the uniform manifold approximation and projection (UMAP) method [27] to facilitate visual interpretation of high-dimensional distributions. The UMAP method is designed to embed the high-dimensional representations into a low-dimensional space while preserving pairwise relationships among samples. As illustrated in Fig. 1(b), normal and defective samples are distributed on a similar manifold, forming a pattern of concentric circles with overlapping regions without distinct boundaries. This demonstrates the lack of separation in the initial representation, highlighting the

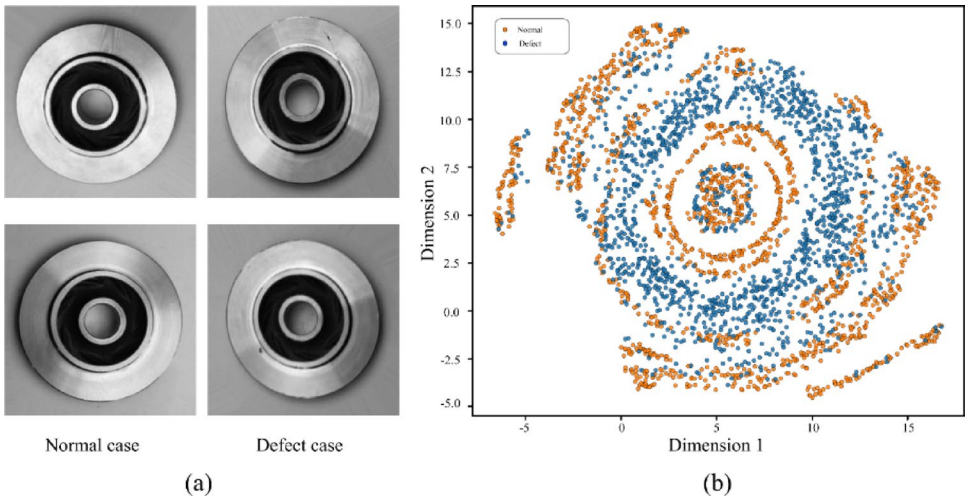



**Fig. 1** Representative examples of the utilized dataset and its distributional characteristics: (**a**) normal and defective casting surface cases, (**b**) UMAP visualizations obtained from the original data distribution of image-level normal and defective samples

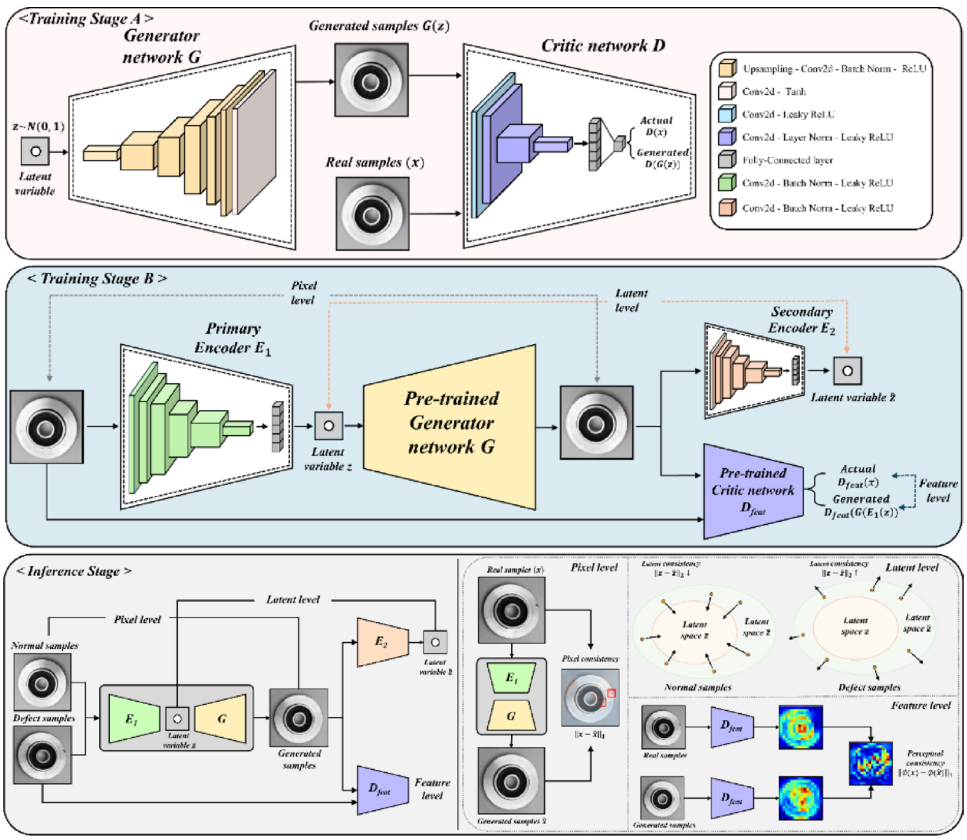



**Fig. 2** Architectural overview of the proposed framework for casting defect detection, illustrating the training stages and the inference stage

inherent complexity and difficulty of the considered casting surface anomaly detection task. Therefore, the primary objective of this study is to achieve precise separation between the highly entangled distributions of normal and defective samples using a DL approach, enabling reliable defect detection.

## 2.2 Proposed Method

This study introduces a DL framework that utilizes the latent representations of normal samples to reliably identify and interpret defects. Figure 2 illustrates the overall description of the proposed framework. Specifically, the proposed model integrates a dual encoder with a generator-critic backbone to maintain consistency among the input, reconstruction, and latent spaces, enabling the generation of anomaly scores. The training process comprises two stages: Training Stage A employs a data generation backbone to achieve the distribution mapping of normal samples, while Training Stage B utilizes a dual-encoder module to align the reconstruction and latent spaces for anomaly score estimation. After training, the inference stage evaluates

deviations from the learned multi-level correspondence to identify anomalous samples. Details on the training strategies are introduced in Sect. 2.3.

In Training Stage A, the proposed framework models the distribution of normal data while generating normal cases based on a generative adversarial network (GAN) approach. In this study, we adopt the Wasserstein GAN with gradient penalty (WGAN-GP) [20, 21] framework to effectively learn the distribution of the normal samples with improved training stability. Overall architecture consists of two sub-networks: a generator network $G$ and a critic network $D$, which are trained adversarially so that the model distribution approximates the true distribution of normal data. First, the generator network $G$ is designed to generate normal surface images by utilizing a latent vector $z$ that is sampled from a standard normal distribution. $G$ progressively upsamples the feature maps via a series of nearest-upsampling convolutional blocks with batch normalization and rectified linear unit (ReLU) activation. This process can be expressed as follows:

$$U_{(l+1)} = H\left(U_l; \theta\right), \tag{1}$$

where $H$ represents the composite operation consiting of a convolutional layer, a batch normalization, and a nonlinear activation, while $\theta$denotes the set of learnable parameters in the block. At the final layer, a tanh activation function is used to generate normalized normal surface images, with output pixel values ranging between $-1$ and 1.

In addition, the model is jointly trained with a discriminative process to enhance the alignment of generated images with the ground-truth normal distribution. The discriminative process utilizes the critic network $D$, which encodes spatial feature maps through two-dimensional convolutional blocks from either ground-truth or generated data representations. Each block applies a layer normalization and a LeakyReLU activation function. The output from the final convolutional block is flattened into a vector, serving as the intermediate feature embedding $\varphi(x) = D_{\text{feat}}(x)$ for Training Stage B. Subsequently, a fully connected layer is applied to $\varphi(x)$ to yield the critic output $D(x)$. Overall, the critic loss function $\mathcal{L}_D$ as well as the generator loss function $\mathcal{L}_G$ can be expressed as

$$\mathcal{L}_D = \mathbb{E}_z\left[D\left(G\left(z\right)\right)\right] - \mathbb{E}_x\left[D\left(x\right)\right] + \lambda \mathbb{E}_{\hat{x}}\left[\left(\left\|\nabla_{\hat{x}} D\left(\hat{x}\right)\right\|_2 - 1\right)^2\right], \tag{2}$$

$$\mathcal{L}_G = -\mathbb{E}_z\left[D\left(G\left(z\right)\right)\right]. \tag{3}$$

Here, $\hat{x}$denotes $\alpha x + (1-\alpha)G\left(_x\right)$, where $\alpha$ is sampled from $\mathcal{U}(0,1)$. The coefficient $\lambda$ in $\mathcal{L}_D$ serves as the weighting factor for the gradient penalty term, which enforces the 1-Lipschitz constraint on $D$, thereby improving the numerical stability of the training process [21]. Through the adversarial training between the two networks, $D$ learns to maximize the discrepancy between ground-truth and generated distributions, while $G$ learns to produce synthetic samples that closely approximate the normal surface distribution. The overall training procedure for the proposed framework is summarized in Algorithm 1.

**Input**: Normal training set $S_{train}$
  **for** each $x \in S_{train}$ **do**
    $z \sim N(0, I)$
    $\tilde{x} \leftarrow G(z)$
    $\hat{x} \leftarrow \alpha x + (1 - \alpha)\tilde{x},\ \alpha \sim U(0, 1)$
    $L_D = E_z[D(\tilde{x})] - E_x[D(x)] + \lambda E_{\hat{x}}\left(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1\right)^2$
    $L_G = -E_z[D(\tilde{x})]$
    $D \leftarrow \arg\min_D L_D$
    $G \leftarrow \arg\min_G L_G$
  **end for**
  G, D **fixed**
  $\phi(\bullet) \leftarrow D_{feat}(\bullet)$
  **for** each $x \in S_{train}$ **do**
    $z \leftarrow E_1(x)$
    $\tilde{x} \leftarrow G(z)$
    $\tilde{z} \leftarrow E_2(\tilde{x})$
    $L_{DE} = \omega_{img}\|x - \tilde{x}\|_1 + \omega_{fm}\|\varnothing(x) - \varnothing(\tilde{x})\|_2 + \omega_{lat}\|z - \tilde{z}\|_1$
    $E_1, E_2 \leftarrow \arg\min L_{DE}$
  **end for**
**Output:** $G, D, E_1, E_2$

**Fig. 3** Training procedure of the proposed method

Furthermore, Training Stage B employs a dual-encoder structure to effectively quantify the degree of abnormality with latent-space consistency and feature-level alignment. We employ the pre-trained $G$ and $D$ backbone, which remains fixed throughout the training process. As shown in Fig. 2, the proposed architecture comprises a primary encoder $E_1$ that maps input image $x$ into a latent feature, a pre-trained $G$ that reconstructs the image from the latent variable, a secondary encoder $E_2$ that projects the reconstructed image into the latent space, as well as a pre-trained $D_{\text{feat}}$ from the fixed critic $D$. Initially, $E_1$ maps the input image $x$ into a latent variable $_{x} = E_1(x)$, and a pre-trained $G$ reconstructs it as $\widetilde{x} = G(_{x})$. Subsequently, $E_2$ encodes the reconstructed image $\widetilde{x}$ into another latent representation $\widetilde{}_{x} = E_2\left(\widetilde{x}\right)$, improving consistency among the input, reconstruction, and latent spaces. Therefore, $E_1$ determines the reconstruction pathway and contributes to pixel-level consistency, while $E_2$ complements this process by projecting the reconstructed image back into.

**Input**: $S_{test}$,G,D,$E_1$,$E_2$
  $\phi(\bullet) \leftarrow D_{feat}(\bullet)$
  **for** each x $\in$ $S_{test}$ **do**
    $z \leftarrow E_1(x)$
    $\tilde{x} \leftarrow G(z)$
    $\tilde{z} \leftarrow E_2(\tilde{x})$
    $s_{img}(x) = \|x - \tilde{x}\|_1$
    $s_{fm}(x) = \|\varnothing(x) - \varnothing(\tilde{x})\|_2$
    $s_{lat}(x) = \|z - \tilde{z}\|_1$
    $S(x) = \alpha s_{img}(x) + \beta s_{fm}(x) + \gamma s_{lat}(x)$
  **end for**
**Output:** $\{S(x)\}_{x \in S_{test}}$

**Fig. 4** Inference procedure of the proposed method

The latent space and enforcing consistency at the latent level. This combination of pixel-level reconstruction consistency and latent-level consistency enforcement allows the model to effectively capture deviations associated with anomalous samples. On the other hand, the $D_{feat}$ provides high-dimensional embeddings $\varphi(x)$ and $\varphi\left(\widetilde{x}\right)$ for feature-level matching. In this wise, the dual-encoder training is formulated with a hybrid loss that enforces pixel-wise reconstruction, perceptual consistency, and latent-space alignment. The objective function of the dual-encoder framework can be expressed as:

$$\mathcal{L}_{DE} = \omega_{img}\|x - \tilde{x}\|_1 + \omega_{fm}\|\varnothing(x) - \varnothing(\tilde{x})\|_2 + \omega_{lat}\|z - \tilde{z}\|_1 \quad (4)$$

where $\omega_{img}, \omega_{fm}$, and $\omega_{lat}$ denote the weighting factors for pixel reconstruction, perceptual alignment, and latent consistency losses, respectively. After training processes, we obtain the anomaly score $S(x)$ for a input $x$ by evaluating the three components in the following:

$$S(x) = \alpha\|x - \tilde{x}\|_1 + \beta\|\varnothing(x) - \varnothing(\tilde{x})\|_2 + \gamma\|z - \tilde{z}_1\| \quad (5)$$

As a result, the proposed dual-encoder structure simultaneously enforces (i) pixel-level reconstruction consistency, (ii) perceptual alignment in the critic feature space, and (iii) consistency in the latent space, thereby ensuring multi-level correspondence between normal and defective representations. The overall inference process for obtaining the final anomaly score is summarized in Algorithm 2.

## 2.3 Model Training Strategies

Training strategies of the proposed model are summarized as follows. In the proposed framework, there are two separate training phases, denoted as $P_1$ and $P_2$. Training phase $P_1$ corresponds to the adversarial training process based on the WGAN-GP framework defined in the earlier section. The model is trained using the Adam optimizer [28] with a learning rate of $1 \times 10^{-4}$and a batch size of 32. The momentum coefficients $\beta_1$ and $\beta_2$ are assigned to 0.5 and 0.999, respectively. Hyperparameters are determined via random search processes. A gradient penalty coefficient of $\lambda = 10$ is applied to ensure the 1-Lipschitz constraint of the critic and to stabilize the numerical training process. To maintain the training balance, the number of critic updates per generator iteration is set to $n_{\text{critic}} = 1$, updating both $D$ and $G$ once per mini-batch over 300 epochs. In addition, the Fréchet Inception Distance (FID) [29] is employed to monitor the convergence of distributions and the fidelity of generated samples, evaluated every five training iterations. Through this process, the generator learns the intrinsic distribution of normal surfaces and produces synthetic samples that are consistent with ground-truth normal samples.

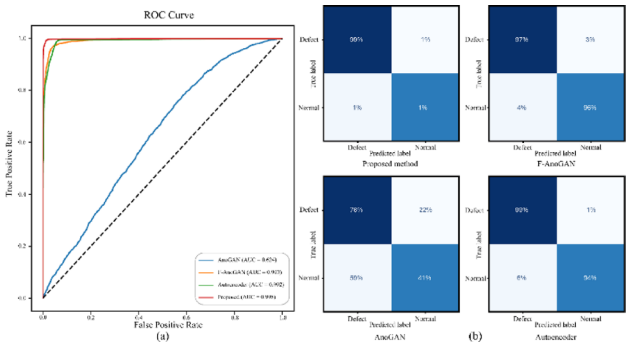



**Fig. 5** Quantitative comparison of defect detection performance: (**a**) ROC curves of the proposed and baseline models, and (**b**) confusion matrices at the optimal threshold

Training phase $E_2$ represents the dual-encoder training process based on the hybrid loss function defined in Eq. (4). In this phase, pre-trained $G$ and $D$ in $P_1$ remain fixed, and only the encoder parameters are updated. Encoders are trained using the same Adam optimizer with the same settings as in $P_1$, including a learning rate of $1 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, a batch size of 32, and a total of 200 epochs. The hybrid loss consists of three components: reconstruction loss, critic feature-matching loss, and latent consistency loss, with respective weights of $\omega_{img} = 5.0$, $\omega_{fm} = 1.0$, and $\omega_{lat} = 1.0$, respectively. While the weighting parameters were selected empirically, a more detailed description of the ablation study under different weighting scenarios is provided in Sect. 3.2. The anomaly score $S(x)$, as defined in Eq. (5), is calculated by assigning equal weights of $\alpha = \beta = \gamma = 1.0$ to the pixel, critic, and latent terms. This approach reinforces consistency among input, reconstruction, and latent representations, which improves the reliability and discriminative capability of the unified anomaly score. During inference, the three loss terms are aggregated to compute the final anomaly score.

## 3 Results and Discussion

This section presents both quantitative and qualitative analyses of the proposed method for casting surface defect detection. We first provide a detailed description of the evaluation metrics and comparative models used for performance evaluation. Subsequently, we demonstrate the detection performance of the proposed approach through quantitative and qualitative results, and finally discuss its potential strengths and limitations.

**Table 1** Quantitative comparison in terms of evaluation metrics

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1 score(%) |
|---|---|---|---|---|
| AnoGAN [31] | 62.5 | 64.2 | 78.2 | 70.5 |
| F-AnoGAN [32] | 96.5 | 97.1 | 96.7 | 96.9 |
| Auto-encoder [33] | 97.0 | 95.9 | 98.9 | 97.4 |
| **Proposed** | **99.0** | **99.1** | **99.1** | **99.1** |

### 3.1 Evaluation Metrics and Comparative Methods

We quantitatively evaluate the performance of the proposed framework using the following evaluation metrics. First, we consider the area under the receiver operating characteristic (AUROC), where the ROC curve represents the relationship between the true positive rate (TPR) and the false positive rate (FPR). This metric provides a threshold-independent measure of classification capability. A higher AUROC value indicates better separability between the two classes. Subsequently, we employ standard classification measures of accuracy, precision, recall, and F1 score to quantitatively evaluate the detection performance for normal and defective cases under a fixed decision threshold. The decision threshold is determined based on Youden's Index [30], which reflects the relationship between recall and FPR.

Regarding the comparison approaches, three representative baseline models are considered to compare the quantitative performance of the proposed method. The baseline models are as follows: (1) AnoGAN, a generative adversarial network that identifies anomalies by optimizing individual samples after training exclusively on normal data; (2) F-AnoGAN, which conducts anomaly detection through embedding-based feature matching with a pre-trained encoder; and (3) an Autoencoder, which applies reconstruction error to detect anomalies [31–33]. The comparison with the proposed model evaluates how integrated learning of pixel-level reconstruction, perceptual consistency, and latent-space alignment affects discriminative performance.

To ensure consistency, all models are trained with identical preprocessing and data-splitting settings, and the evaluation is conducted on the same test dataset.

### 3.2 Quantitative Evaluation of Detection Performance

First of all, we analyze the quantitative defect detection performance of the proposed and the baseline models. To quantitatively evaluate the performance the AUROC values of comparative models are shown in Fig. 5(a). As shown in the

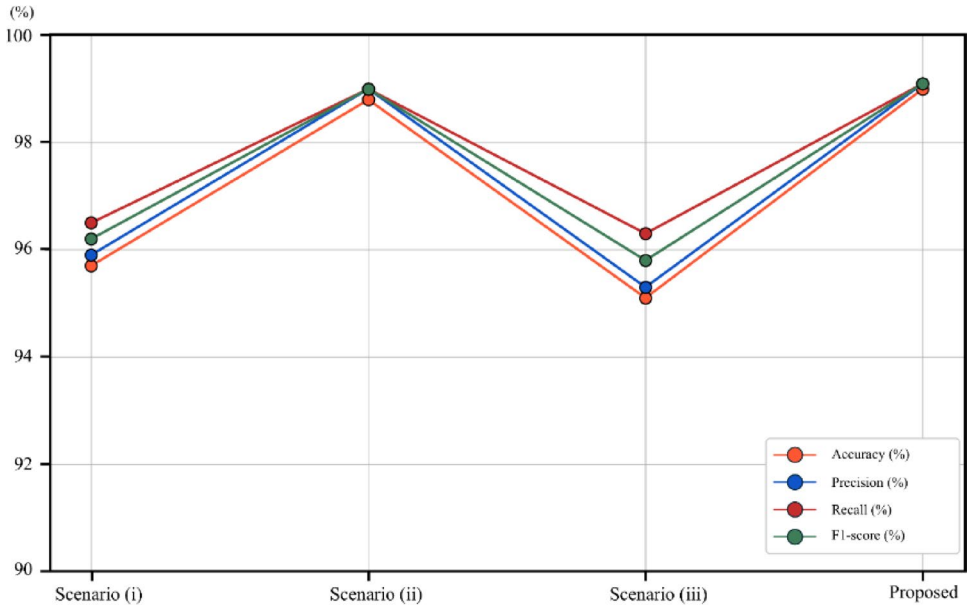



**Fig. 6** Quantitative comparison of the average test results under different weighting scenarios

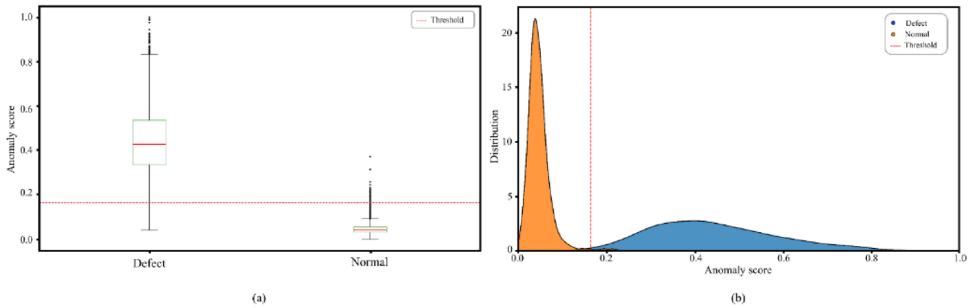



**Fig. 7** Distribution of anomaly scores for normal and defective samples: (**a**) boxplot comparison and (**b**) KDE showing score separation with the selected threshold

figure, we find that the proposed method achieves the highest AUROC score of 0.998, outperforming all baseline models. While the baseline approaches exhibit lower AUROC values, we observe some unstable detection behavior with respect to threshold variation. These results indicate that the proposed model consistently maintains a low false positive rate (FPR) and a high true positive rate (TPR) across the entire threshold range. It is observed that the proposed model attains a higher TPR at the same FPR,

implying that its classification performance is less sensitive to threshold fluctuations. These results confirm that the proposed framework achieves robust separability between normal and defective samples and delivers consistent anomaly detection performance under diverse surface conditions.

Once the optimal decision threshold is determined, the evaluation metrics are obtained to assess the detection capability under this condition. Table 1 summarizes overall quantitative comparison among the methods. Specifically, the proposed model achieves near-perfect accuracy, precision, recall, and F1 score, indicating both high consistency and stability. In contrast, the baseline models show lower accuracy and imbalanced precision and recall, failing to maintain stable classification performance at their respective optimal thresholds. The confusion matrices in Fig. 5(b) further illustrate these differences. The proposed method maintains both false positive (FP) and false negative (FN) samples at relatively low and balanced levels, reflecting stable separation between normal and defective samples. In particular, the balanced FP and FN distribution suggests that the proposed model is not biased toward a specific defect type and secures both precision and recall. On the other hand, the baseline models exhibit distinct error patterns, showing limited generalization across defect types with higher FN and FP values. These findings confirm that the proposed model effectively balances sensitivity and specificity, ensuring stable and accurate defect detection across diverse casting conditions.

To analyze the effect of the weighting factors in Eq. (4), we conduct an ablation study under different weighting scenarios: (i) considering only the pixel reconstruction term $(\omega_{img}, \omega_{fm}, \omega_{lat}) = (1,0,0)$, (ii) assigning equal weights to all terms $(1,1,1)$, (iii) greater emphasis on pixel reconstruction only $(5,0,0)$, and the proposed setting $(5,1,1)$. As shown in Fig. 6, considering only the

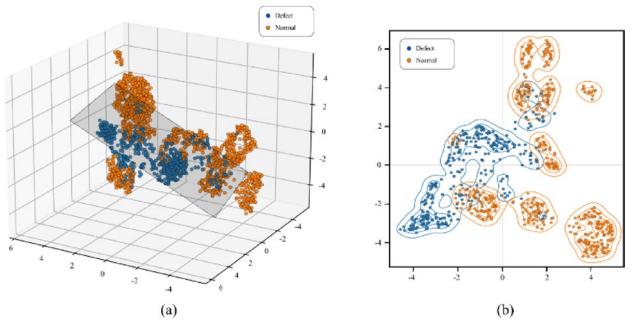



**Fig. 8** UMAP Visualizations projected from the 128-dimensional latent space of the secondary encoder $E_2$ in the proposed model after training: (**a**) three-dimensional representation, (**b**) two-dimensional representation

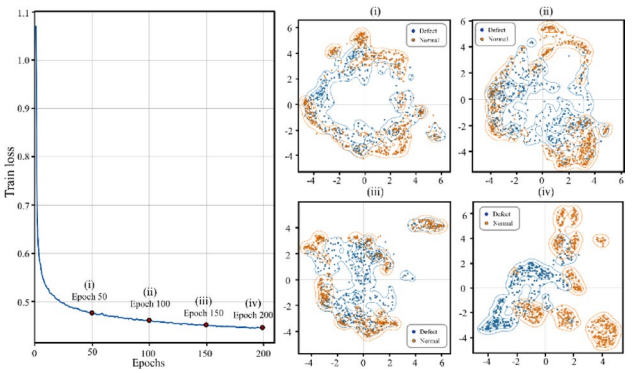



**Fig. 9** Latent space UMAP visualizations of the proposed model across different training phases

pixel reconstruction term (1, 0, 0) results in limited overall anomaly detection performance, while stronger emphasis on pixel reconstruction (5, 0, 0) tends to increase variability across evaluation metrics. In contrast, assigning equal weights to all objective.

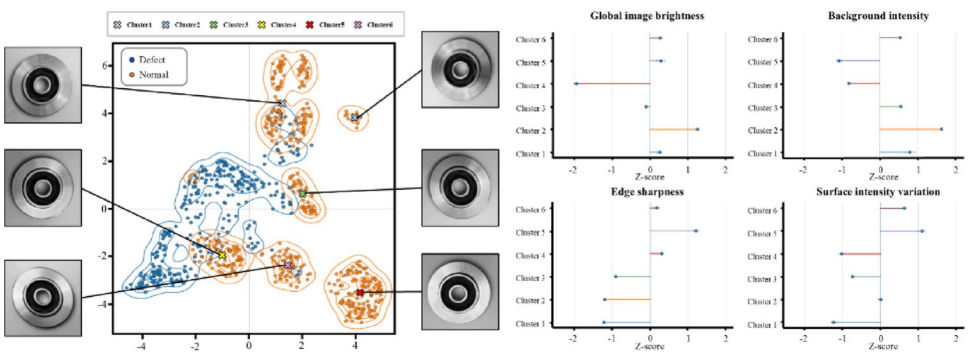
terms (1, 1, 1) significantly improves the overall performance, suggesting the benefit of considering not only pixel-level information but also multi-correspondence information. Finally, we observe that the proposed weighting setting (5, 1, 1) achieves the best performance among different weighting scenarios. This suggests that an appropriate balance between pixel-level reconstruction and higher-level correspondence constraints can contribute to effective anomaly detection.

In addition, we present the distribution of anomaly scores of the proposed approach. The box plot in Fig. 7(a) visualizes the difference in anomaly score distributions between the two groups. The defect group exhibits a significantly higher median and quartile range compared to the normal group, indicating clear separation in both central tendency and dispersion. This heterogeneity in distribution is further highlighted in the kernel density estimation (KDE) plot shown in Fig. 7(b). At the optimal threshold, indicated by a dashed line, the distribution of normal samples is densely concentrated in the low-score region on the left, while defective samples are predominantly located in the high-score region on the right. The minimal overlap between the two distributions demonstrates the strong discriminative capability of the proposed model. This separability strongly supports the previously presented quantitative results, including the high AUROC value and the confusion matrix with low FP and FN values. Consequently, Fig. 7 verifies that the proposed method consistently assigns low anomaly scores to normal samples and high anomaly scores to defective samples, thereby forming a distinct decision boundary that effectively separates the two groups.

### 3.3 Qualitative Reconstruction and Anomaly Visualization

In this subsection, we qualitatively evaluate the proposed model's effectiveness in distinguishing normal from defective samples in the latent space and reconstructing their representations. To assess the qualitative characteristics of the latent representation, we visualize latent vectors from the proposed model. Figure 8 presents the UMAP visualizations projected from the 128-dimensional latent space features of $E_2$ in the proposed model after training. It should be noted that these visualizations are obtained by projecting high-dimensional feature representations into a low-dimensional (2D and 3D) space for visualization purposes, which inevitably.

**Fig. 10** UMAP clusters and corresponding cluster-wise Z-score statistics of image variability



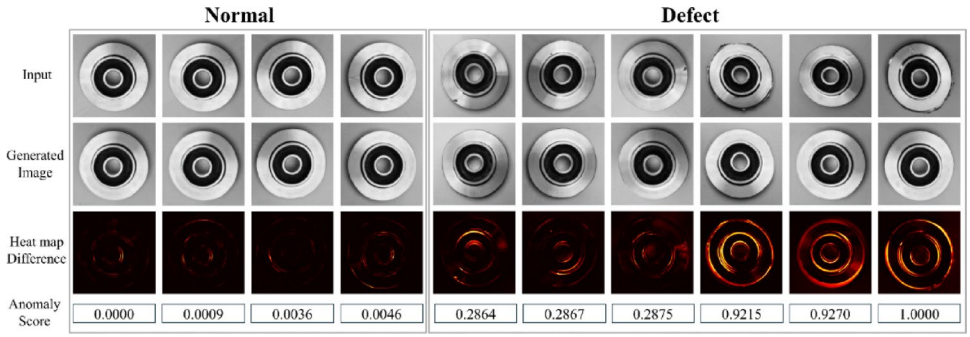



**Fig. 11** Qualitative comparison of reconstruction results from the dual-encoder framework for normal and defective samples, showing input, generated image, heat map difference, and corresponding anomaly score

causes information loss during feature dimensionality reduction. In Fig. 1(b), the original data exhibit a spiral-shaped entanglement, resulting in ambiguous separation between normal and defective samples. In contrast, the latent features extracted from $E_2$ exhibit a relatively more structured and separable distribution between normal and defective samples. Figure 9 is further visualized to determine the evolution of UMAP visualization.

as training progresses. As shown in the figure, we find that the visualized feature representations evolve from a high entangled manifold in the early phases of training to a more gradually disentangled structure. These visualizations indicate that the proposed model progressively learns to discriminate between normal and defective samples.

Figure 10 further visualizes cluster-wise characteristics for defective samples. We obtain Z-score values representing widely utilized image variability measures [34] (global image brightness, background intensity,

edge sharpness, surface intensity variation), for samples belonging to different clusters. From the results, we observe that each cluster emerges from a distinct combination of diverse visual variability factors. This observation is particularly important for real-world industrial applications, where measurement environments are not always consistent due to variations in illumination, surface conditions, and acquisition settings. Although this study primarily focuses on separating normal and defective samples, the presence of multiple.

clusters suggests that the proposed method implicitly captures variations related to measurement conditions. This indicates the potential of the proposed approach to further discriminate differences in acquisition environments beyond defect-related characteristics.

Qualitative characteristics are further demonstrated by analyzing the generated results and comparing the distinctions between normal and defective samples. Figure 11 provides a qualitative comparison of generated results for normal and defective samples. For normal cases, the generated images closely match the inputs, with weak responses in the difference heat maps and low anomaly scores near $10^{-3}$. In contrast, the defective cases show incomplete reconstruction of the defect regions, leading to high heat map differences along defect.

boundaries and increased anomaly scores ranging from 0.28 to 1.00. These results show that the proposed framework performs consistent reconstruction on the data manifold for normal samples, while reconstruction discrepancies for defective samples are amplified as distinct errors. Based on these results, the proposed dual-encoder framework with hybrid learning objectives, namely reconstruction, perceptual consistency, and latent-space alignment, achieve clear separation between normal and defective representations.

### 3.4 Advantages and Limitations of Proposed Approach

We first highlight the potential advantages of the proposed approach. The main contribution of this study can be summarized in two perspectives: (1) the development of a dual-encoder-based generative anomaly detection method for casting surface defect detection, and (2) enhanced interpretability in model's decision making process. First, the proposed approach is designed to define the presence of defects using a dual-encoder generator architecture trained exclusively on normal data, offering an effective alternative to conventional rule-based or label-dependent supervised methods. This enables real-time quality inspection on casting production.

lines, reducing verification time during process transitions and setup stages, and improves operational efficiency. Second, the proposed framework ensures transparency and reliability of predictions through multi-level interpretability, visualizing spatial bases of discrimination. By utilizing reconstruction residual heatmaps and anomaly score distributions, the framework can qualitatively explain the location of casting defects (e.g., surface pores and microcracks), thereby improving the reliability of the data-driven method in real-world application. Overall, our findings suggest that the proposed framework can effectively identify defects using only normal data and enhance both predictive performance and interpretability.

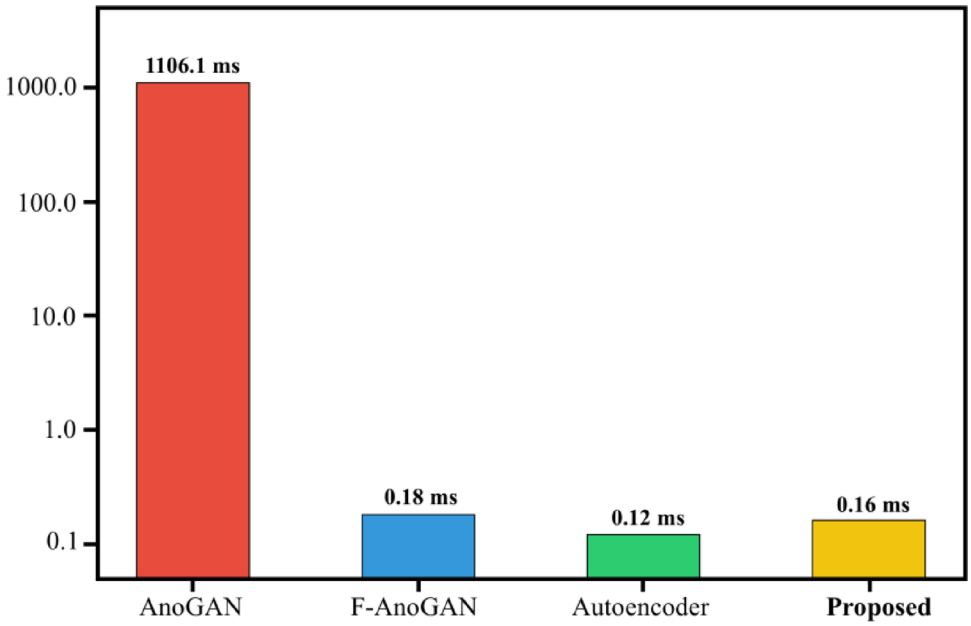



**Fig. 12** Comparative results of the average inference time per test image

We further discuss the scalability and computational efficiency of the proposed method. First, regarding the scalability, it should be noted that the proposed method is designed in a manner that allows it to be scalable to higher-resolution images and larger datasets, owing to its methodological design. In this study, all sub-networks $(G, D, E_1, E_2)$ are designed with architecturally symmetric principles based on encoder-decoder structures. This design facilitates inference on images with resolutions different from the considered $256 \times 256$ (e.g., $128 \times 128$ or $512 \times 512$). In addition, since the proposed method considers multi-level correspondences not only at the image level but also at the latent and feature levels, it can effectively leverage hidden-level feature information even when the input image resolution increases. Furthermore, as this study focuses on unsupervised anomaly detection instead of multi-class classification, issues related to class imbalance and extensive labeling costs that are often encountered in large industrial datasets are largely alleviated. Regarding the computational efficiency, Fig. 12 visualizes a comparative experiment on the inference time of the proposed and baseline models. In this study, the trainable parameters of all models were consistently controlled to ensure a fair evaluation ($\sim 17.2$M). As shown in the figure, all compared methods except AnoGAN are capable of real-time inference for detecting normal or defective samples. The main difference between AnoGAN and the other methods lies in the transition from an iterative latent space search to a deterministic, neural network-based feed-forward mapping. While AnoGAN relies on an optimization-based approach that requires numerous backpropagation steps during the inference stage, other deterministic approaches map the input image directly onto the latent manifold. A key characteristic of the proposed method is the construction of a dual-path architecture, which improves the representation capacity for multi-level features without sacrificing real-time inference efficiency. Such an architectural transition is expected to be more effective in real-world industrial application scenarios, where inference efficiency is a critical requirement.

On the other hand, the proposed approach still has limitations in terms of generalization, which suggests directions for future research. Given that the model is trained on normal data collected under specific casting process conditions, its direct applicability to different product types, production lines, or measurement environments remains limited. Therefore, additional retraining or data adaptation techniques may be required in the following cases. If the target application involves casting products different from those examined in this study, it is necessary to collect additional representative normal samples and perform lightweight fine-tuning to align the model with the new data distribution. If physical measurement conditions such as viewing angle or sensor configuration differ from those in this study, the distribution of normal data may change, which could increase the false positive rate of the model. To ensure robustness under altered conditions, additional preprocessing techniques should be considered to maintain consistency in the measured data representation.

## 4 Conclusions

This study presents a deep learning-based generative anomaly detection approach for unsupervised detection of casting surfaces. To address the challenges in defect detection caused by complex and fine-grained variations of defect cases, we propose a dual-encoder generative anomaly detection framework trained solely on normal data. The proposed model integrates a generator, a critic, and two encoders to learn the intrinsic representation of normal surfaces through adversarial learning and enforce consistency among input, reconstruction, and latent representations. This integration enables the formulation of a unified anomaly score that combines reconstruction error, critic feature matching, and latent consistency. We demonstrate that the proposed framework effectively distinguishes normal and defective samples under diverse defect cases, achieving an AUROC of 0.998 and an accuracy of 99.1%. Furthermore, the experimental results indicate that the model attains balanced performance across precision, recall, and F1 score. Reconstruction residual heatmaps, anomaly-score distributions, and UMAP projections confirm that normal and defective samples.

are clearly separated in score and latent spaces. These results highlight the potential of the framework to improve the effectiveness and applicability of unsupervised anomaly detection, accelerating reliable defect diagnosis and automated quality inspection across diverse manufacturing applications.

**Acknowledgements** This work was supported by the Chung-Ang University Research Grants in 2024. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00342929).

**Funding** Open Access funding enabled and organized by Chung-Ang University

**Data Availability** The datasets used in this work are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Murakami, Y., & Endo, M. (1994). Effects of defects, inclusions and inhomogeneities on fatigue strength. *International Journal Of Fatigue*, *16*, 163–182.
2. Beretta, S., & Romano, S. (2017). A comparison of fatigue strength sensitivity to defects for materials manufactured by am or traditional processes. *International Journal Of Fatigue*, *94*, 178–191.
3. Han, M., Zheng, H., Gao, Y., Zhou, Z., & Liu, X. (2024). Detection of near-surface defects in viscoelastic material based on focused ultrasound thermal effects. *Ultra-sonics*, *143*, 107415.
4. Zhang, B., Li, Y., & Bai, Q. (2017). Defect formation mechanisms in selective laser melting: a review. *Chin J Mech Eng*, *30*, 515–527.
5. Gong, H., Rafi, K., Gu, H., Starr, T., & Stucker, B. (2014). Analysis of defect generation in ti–6al–4v parts made using powder bed fusion additive manufacturing processes. *Addit Manuf*, *1*, 87–98.
6. Chelladurai, C., Mohan, N., Hariharashayee, D., Manikandan, S., & Sivaperumal, P. (2021). Analyzing the casting defects in small scale casting industry. *Mater Today: Proc*, *37*, 386–394.
7. Gunasegaram, D. R., Farnsworth, D., & Nguyen, T. T. (2009). Identification of critical factors affecting shrinkage porosity in permanent mold casting using numerical simulations based on design of experiments. *Journal Of Materials Processing Technology*, *209*, 1209–1219.
8. Cao, X., Wallace, W., Immarigeon, J. P., & Poon, C. (2003). Research and progress in laser welding of wrought aluminum alloys. ii. metallurgical microstructures, defects, and mechanical properties. *Materials And Manufacturing Processes*, *18*, 23–49.
9. Kah, P., Rajan, R., Martikainen, J., & Suoranta, R. (2015). Investigation of weld defects in friction-stir welding and fusion welding of aluminium alloys. *Int J Mech Mater Eng*, *10*, 26.
10. Zhang, Chen, & Poirier (2000). Effect of solidification cooling rate on the fatigue life of a356. 2-t6 cast aluminium alloy. *Fatigue And Fracture Of Engineering Materials And Structures*, *23*, 417–423.
11. Tai, J. L., Sultan, M. T. H., Łukaszewicz, A., Józwik, J., Oksiuta, Z., & Shahar, F. S. (2025). Recent trends in non-destructive testing approaches for composite materials: A review of successful implementations. *Materials*, *18*, 3146.
12. Dwivedi, S. K., Vishwakarma, M., & Soni, A. (2018). Advances and researches on non destructive testing: A review. *Mater Today: Proc*, *5*, 3690–3698.
13. See, J. E. (2015). Visual inspection reliability for precision manufactured parts. *Human Factors*, *57*, 1427–1442.
14. Wang, W., Wong, Y., & Hong, G. S. (2005). Flank wear measurement by successive image analysis. *Computers In Industry*, *56*, 816–830.
15. Ren, Z., Fang, F., Yan, N., & Wu, Y. (2022). State of the art in defect detection based on machine vision. *Int J Precis Eng Manuf - Green*, *9*, 661–691.
16. Manish, R., Venkatesh, A., & Ashok, S. D. (2018). Machine vision based image processing techniques for surface finish and defect inspection in a grinding process. *Mater Today: Proc*, *5*, 12792–12802.
17. Lin, J., Yao, Y., Ma, L., & Wang, Y. (2018). Detection of a casting defect tracked by deep convolution neural network. *International Journal Of Advanced Manufacturing Technology*, *97*, 573–581.
18. Gupta, R., Anand, V., Gupta, S., & Koundal, D. (2023). Deep learning model for defect analysis in industry using casting images. *Expert Systems With Applications*, *232*, 120758.
19. Chamberland, O., Reckzin, M., & Hashim, H. A. (2023). An autoencoder with convolutional neural network for surface defect detection on cast components. *J Fail Anal*, *23*, 1633–1644.
20. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Int Conf Mach Learn PMLR*, *70*, 214–223.

21. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Adv Neural Inf Process* 30, 1-11.
22. Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal Of Thoracic Oncology : Official Publication Of The International Association For The Study Of Lung Cancer*, *5*, 1315–1316.
23. Dabhi, R. (2019). Casting product image data for quality inspection https://www.kaggle.com/datasets/ravirajsinh45/real-life-industrial-dataset-of-casting-product.
24. Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). Deepant: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access : Practical Innovations, Open Solutions*, *7*, 1991–2005.
25. Iqbal, E., Khan, S. U., Javed, S., Moyo, B., Zweiri, Y., & Abdulrahman, Y. (2024). Multi-scale feature reconstruction network for industrial anomaly detection. *Knowledge-Based Syst*, *305*, 112650.
26. Du, X., Li, B., Liu, T., Ding, Y., Jin, L., & Zhao, Z. (2025). Seglgad: Local-to-global industrial anomaly detection with visual segmentation model. *Optics & Laser Technology*, *186*, 112633.
27. Ghojogh, B., Crowley, M., Karray, F., & Ghodsi, A. (2023). Uniform manifold approximation and projection (umap). *Elements of dimensionality reduction and manifold learning* (pp. 479–497). Springer Nature Switzerland AG
28. Adam, K. D. B. J. (2014). A method for stochastic optimization, arXiv preprint arXiv:1412.6980 1412.
29. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv Neural Inf Process* 30, 1-12.
30. Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the youden index and its associated cutoff point. *Biometrical Journal. Biometrische Zeitschrift*, *47*, 458–472.
31. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, CoRR abs/1703.05921 URL: http://arxiv.org/abs/1703.05921. arXiv:1703.05921.
32. Schlegl, T., Seeböck, P., Waldstein, S. M., & Langs, G. (2019). Schmidt-Erfurth, f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, *54*, 30–44.
33. Chow, J. K., Su, Z., Wu, J., Tan, P. S., Mao, X., & Wang, Y. H. (2020). Anomaly detection of defects on concrete structures with the convolutional autoencoder. *Advanced Engineering Informatics*, *45*, 101105.
34. Yeganeh, H., & Wang, Z. (2012). Objective quality assessment of tone-mapped images. *Ieee Transactions On Image Processing*, *22*, 657–667.


**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Seongmin Kwak** is an graduate researcher in the School of Mechanical Engineering at Chung-Ang University, Seoul, Republic of Korea. He is currently affiliated with the Industrial AI Laboratory, where he focuses on the AI development to mechanical and manufacturing systems. His research interests include deep generative models and physics-informed machine learning for solving complex engineering problems.



**Sooyoung Lee** is currently an assistant professor and a principal investigator of the Industrial AI Laboratory at Chung-Ang University, Seoul, Republic of Korea. He received a B.S. degree from Chung-Ang University in 2019 and received his Ph.D. degree from Pohang University of Science and Technology (POSTECH) in 2023. He was an Honorary Research Associate for Sustainable Smart Manufacturing using AI with the University of Wisconsin-Madison, WI, USA, supported by the High-Potential Individuals Global Training Program of International Joint Research. He was also a Senior Research developing large-scale AI-based next-era semiconductor manufacturing technologies at Future Technology Institute, SK Hynix. His research interests span AI-enabled accelerated design, smart manufacturing and intelligent informatics for a variety of engineering systems and industrial applications.