

Research Article

Open Access



# Application of self-supervised learning in steel surface defect detection

Shiyu Hu, Xudong Ma, Yuqi Zhang, Wei Xu\*

State Key Laboratory of Digital Steel, Northeastern University, Shenyang 110819, Liaoning, China.

\*Correspondence to: Prof. Wei Xu, State Key Laboratory of Digital Steel, Northeastern University, NO. 3-11, Wenhua Road, Shenyang 110819, Liaoning, China. E-mail: xuwei@ral.neu.edu.cn

**How to cite this article:** Hu, S.; Ma, X.; Zhang, Y.; Xu, W. Application of self-supervised learning in steel surface defect detection. *J. Mater. Inf.* 2025, 5, 44. <https://dx.doi.org/10.20517/jmi.2025.21>

**Received:** 1 Apr 2025 **First Decision:** 10 Jun 2025 **Revised:** 24 Jun 2025 **Accepted:** 8 Jul 2025 **Published:** 22 Jul 2025

**Academic Editor:** Qian Ma **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

## Abstract

In scientific research, effective utilization of unlabeled data has become pivotal, as exemplified by AlphaFold2, which won the 2024 Nobel Prize. Pioneering this paradigm shift, we develop a universal self-supervised learning methodology for detecting surface defects in steel materials. By harnessing unlabeled data, our approach significantly reduces the dependence for manual annotation and enhances scalability while training robust models capable of generalizing across defect types. Using a Faster R-CNN framework, we achieved a mean average precision (mAP) of 0.385 and a mAP at IoU = 0.5 (mAP<sub>50</sub>) of 0.768 on the NEU-DET steel defects dataset. These results demonstrate both the efficacy of our self-supervised strategy and its potential as a framework for developing image detection systems with minimal labeled data requirements in surface defect identification.

**Keywords:** Unlabelled data, self-supervised learning, deep learning, steel materials, image detection

## INTRODUCTION

In recent years, deep learning (DL) has revolutionized various fields by significantly enhancing the accuracy and capabilities of data analysis<sup>[1]</sup>. One prominent application is in computer vision, where DL algorithms have become indispensable tools, driving advancements in areas ranging from image recognition to complex pattern detection<sup>[2-4]</sup>.



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Existing computer vision models exhibit persistent limitations in industrial defect detection applications, particularly concerning prediction accuracy. Numerous studies highlight fundamental challenges including inadequate small-object detection capabilities - especially problematic for discerning subtle metal surface defects against complex backgrounds<sup>[5-7]</sup> - current approaches demonstrate significant vulnerability to environmental variations (e.g., uneven lighting, surface reflections)<sup>[8]</sup>. And specialized models frequently demonstrate material-specific bias that compromises cross-domain generalizability<sup>[9]</sup>. Model-specific constraints further exacerbate these limitations: The “You Only Look Once” (YOLO) series (e.g., YOLOv5 and YOLOv7) demonstrate compromised small-object sensitivity despite their efficiency advantages<sup>[10,11]</sup> while the region-based convolutional neural network (R-CNN) series (e.g., Faster R-CNN and Mask R-CNN) incur prohibitive computational costs that hinder industrial adoption despite high accuracy<sup>[12]</sup>. Even specialized solutions such as improved random forests (91% steel defect accuracy)<sup>[13]</sup>, and Mask R-CNN-based rail defect identification network enhancing rail safety through precise defect localization<sup>[14]</sup>, while methods specifically optimized for NEU-DET - including enhanced Faster R-CNN variants<sup>[15,16]</sup>, Region of Interest (ROI)-pooling-based steel defect detectors<sup>[17]</sup>, and the YOLO-DSC algorithm, which significantly improves detection speed on the NEU-DET dataset<sup>[18]</sup> - demonstrate limited cross-dataset generalizability due to their calibration to particular defect distributions and imaging conditions. Despite these advancements in accuracy and efficiency, supervised learning models face critical limitations in real-world industrial applications due to their computational demands and reliance on extensive labeled data, which refers to information that has been annotated with one or more labels to provide context or meaning. Their performance degrades significantly under variable imaging conditions<sup>[19]</sup>, while the need for large volumes of labeled training data poses significant challenges<sup>[20]</sup>, particularly in industrial contexts. Acquiring sufficient labeled data - especially for specialized tasks such as manufacturing defect detection - is often impractical due to constraints of time, cost, and expertise. Moreover, the labor-intensive nature of data labeling complicates the deployment of supervised learning models in practical scenarios.

To overcome these limitations, the scientific community has increasingly focused on leveraging unlabeled data without any annotations. Self-supervised learning<sup>[21]</sup> has emerged as a promising paradigm by utilizing unlabeled data to learn intrinsic features and patterns, thereby reducing dependency on manual annotation. The efficacy of self-supervised learning has been demonstrated across diverse fields<sup>[22-24]</sup>, notably in AlphaFold2’s Nobel Prize-winning application of predicting protein structures with high accuracy<sup>[25]</sup>. Given the complexities and data constraints in industrial defect detection, self-supervised learning presents a highly promising solution. In industries such as aerospace, automotive, and manufacturing, the ability to detect surface defects in metallic materials is vital for ensuring product reliability and safety<sup>[26]</sup>.

Researchers have increasingly investigated the potential of self-supervised learning frameworks for detecting surface defects in unlabeled image datasets<sup>[27-29]</sup>. To develop a more practical defect inspection system, the self-supervised efficient defect detector (SEDD)<sup>[30]</sup> was introduced. This detector combines self-supervised learning with image segmentation, utilizing an enhanced single-responsiveness self-supervised strategy to achieve competitive performance without requiring annotated defective samples. The model has been evaluated on three representative datasets, consistently demonstrating superior average precision (AP) performance. Compared to traditional CNN models, integrating self-supervised learning principles with adaptive learning rates - adjusted based on loss and weight - can significantly improve detection outcomes<sup>[31]</sup>. However, this approach was validated only on a small dataset, highlighting the need for further research on larger datasets to substantiate the efficacy of self-supervised learning methods. Importantly, leveraging extensive collections of unlabeled images from relevant scenarios allows for pre-training self-supervised models on upstream tasks, followed by fine-tuning for downstream target domains or tasks, aligning with conventional transfer learning practices<sup>[32]</sup>. Combining self-supervised learning with transfer

learning has proven effective in addressing metal defect detection challenges in datasets with limited labeling<sup>[33]</sup>. Self-supervised pre-training models are typically built on contrastive learning frameworks<sup>[34,35]</sup>, which train models to differentiate data samples by comparing similarities and differences between data points. Representative methods include momentum contrast (MOCO)<sup>[36]</sup>, simple framework for contrastive learning of visual representations (SimCLR)<sup>[37]</sup>, and simple Siamese network (SimSiam)<sup>[38]</sup>. MOCO optimizes contrastive learning by dynamically managing a large number of negative samples using a dictionary queue, though its complex architecture demands significant computational resources. SimCLR enhances performance through large batch sizes and advanced data augmentation techniques to generate negative sample pairs but similarly requires substantial computational resources and carefully designed augmentation strategies. In contrast, SimSiam stands out for its simplicity and efficiency. It eliminates the need for negative samples or momentum encoders and prevents feature collapse via stop-gradient operations and symmetric predictor designs, resulting in a streamlined architecture that is easier to implement and scale. SimSiam has achieved state-of-the-art performance across multiple benchmark datasets.

In this study, we introduce a streamlined self-supervised defect detection framework and devise a weight transfer scheme to pre-train the comparative learning SimSiam model on a comprehensive dataset of unlabeled images to learn their inherent features. Subsequently, we employ the learned weights of the model to serve as a feature extractor within Faster R-CNN for application on a constrained dataset of labeled steel surface defect images to evaluate detection performance. This study examines the factors influencing detection accuracy and confirms the efficacy of self-supervised learning approaches in steel surface defect detection. The principal contributions of this paper are as follows:

- (1) This paper introduces a novel self-supervised learning approach specifically tailored for steel surface defect detection. It utilizes unlabeled data to train a model capable of generalizing effectively to various types of defects with a high degree of accuracy. The approach markedly diminishes dependence on manual labeling, which is labor-intensive and costly, and offers a scalable solution for microstructural image classification and localization in surface defect identification.
- (2) We propose a streamlined self-supervised learning framework for steel surface defect detection that reduces model complexity, enhances interpretability and reliability, negates the need for extensive labeled data, and abbreviates detection time.
- (3) Extensive experiments demonstrate that our approach achieves superior results compared to a baseline model using random weights and ImageNet pre-trained ResNet18 weights on the publicly available downstream defect dataset NEU-DET. Our study underscores the viability of self-supervised methods in this domain and lays the groundwork for more advanced defect detection techniques.

The subsequent sections of this paper are organized as follows: The “Methodology” section details the proposed methodology, including a comprehensive description of dataset creation and the self-supervised learning framework. The “Results” section delineates the experimental results and analysis. The “Discussion” section discusses the limitations of the study and proposes future research directions. Finally, the “Conclusions” section encapsulates the findings of the paper.

## MATERIALS AND METHODS

### Dataset acquisition and data pre-processing

In this study, a strip steel surface defect image dataset<sup>[39]</sup> provided by Northeastern University (NEU) was used. The dataset contains six major categories of surface defects associated with hot-rolled strips: slag (RS), indentation (Pa), crack (Cr), speckle (PS), inclusions (In), and scratches (Sc). Examples of six defect types are shown in Figure 1A. 300 samples were taken from each category, resulting in a comprehensive dataset of 1,800 grayscale images. Each image in the NEU dataset has a resolution of  $200 \times 200$  pixels. In the defect detection task, the NEU-DET dataset also provides bounding box annotations to delineate the defect categories and their spatial locations in each image. The bounding boxes in the NEU-DET dataset were exclusively utilized during fine-tuning as annotation labels. The dataset is available on the web ([http://faculty.neu.edu.cn/songkechen/zh\\_CN/zdylm/263270/list/index.htm](http://faculty.neu.edu.cn/songkechen/zh_CN/zdylm/263270/list/index.htm)).

To further expand the unlabeled dataset for self-supervised learning, we acquired the dataset from the Severstal Steel Defect Detection competition hosted on the Kaggle platform. The dataset is available on the web (<https://www.kaggle.com/competitions/severstal-steel-defect-detection/data>). Original training images ( $1,600 \times 256$  pixels) were cropped along the length direction into four sub-images of equal dimensions. This process generated augmented samples while preserving defect features. The merged dataset (hereafter referred to as the SSDD dataset) consists of 20,272 images of surface defects on strips, each  $400 \times 256$  pixels in size. Examples of these defects are shown in Figure 1B.

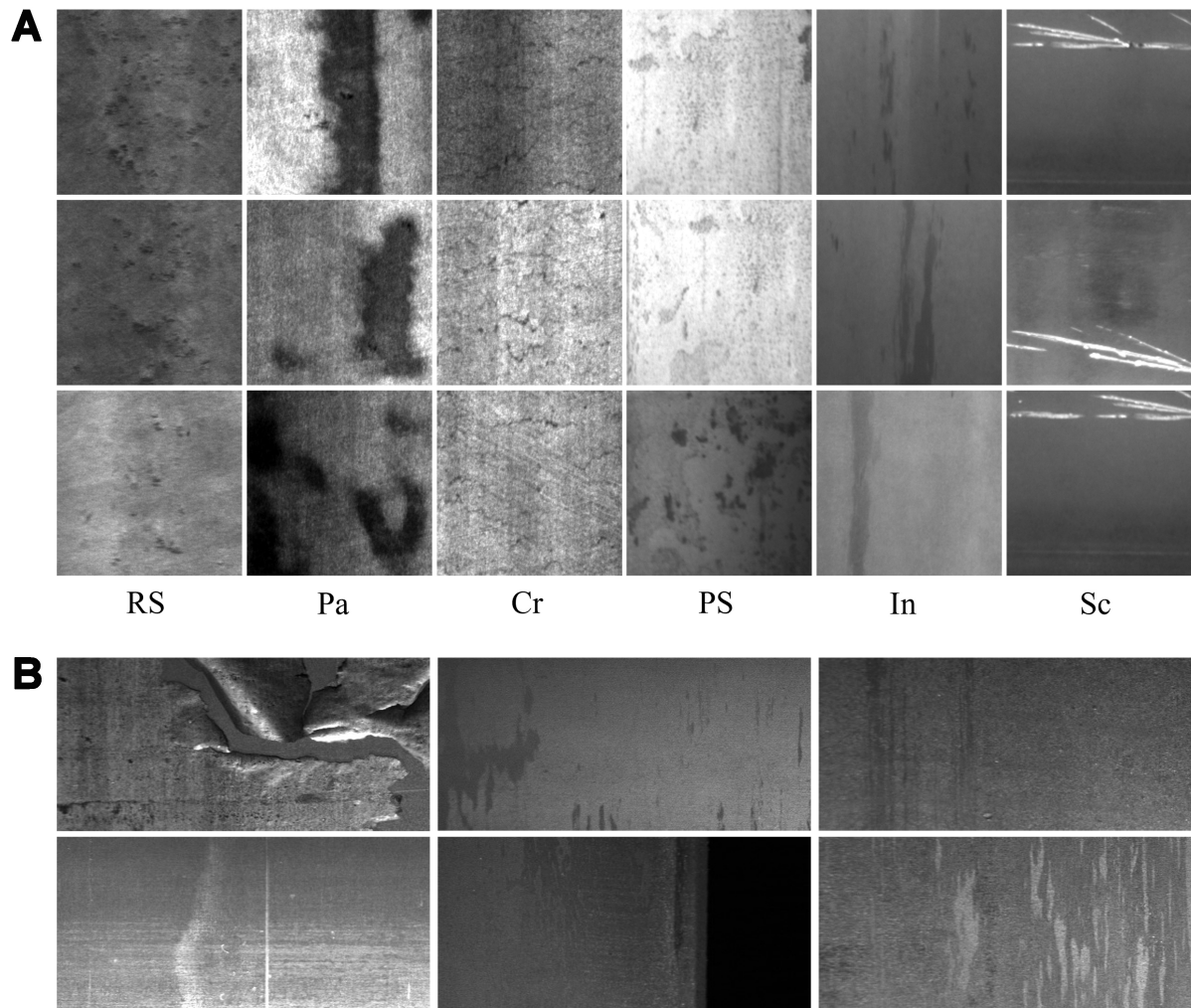
Following the specified protocol, we integrate the NEU dataset with the SSDD dataset to form a new pre-training dataset named SSDD + NEU.

### Self-supervised learning framework

Self-supervised learning has emerged as a promising approach for advancing machine learning. Unlike supervised learning, which requires labeled data, self-supervised learning generates supervisory signals directly from unannotated visual data, enabling the acquisition of generalizable and transferable representations without human intervention<sup>[40]</sup>. In computer vision, both transfer learning and self-supervised learning are employed for model pre-training<sup>[41]</sup>. Critically, self-supervised learning facilitates pre-training on large-scale datasets of unlabeled images, substantially expanding data resources. Figure 2 illustrates our workflow for self-supervised learning-enhanced material defect detection.

The algorithm comprises a pre-training stage followed by a downstream defect detection task. During self-supervised pre-training, the model learns from unlabeled images of steel surface defects. This phase focuses not on direct defect detection but on learning generic image representations. The resulting weights encapsulate transferable features, providing a foundation for downstream tasks. In transfer learning, these weights serve as feature extractors for steel defect detection. Typically, only the final layers require fine-tuning when adapting to this task: deeper layers retain general features, while output-proximal layers necessitate task-specific adaptation. By fine-tuning on limited labeled defect data, the model acquires defect-specific characteristics. During detection, the model processes new steel surface images to localize and classify defects. Leveraging rich pre-trained features and task-adapted fine-tuning enables effective defect detection with minimal labeled data. The self-supervised pre-training employs an online augmentation pipeline adapted from established contrastive learning methodologies. For each input image, two stochastically augmented views are generated via random resized cropping, color jittering, random grayscale conversion, gaussian blurring, and horizontal flipping. This multi-transformation strategy ensures robust feature learning while preserving defect discriminability. Critically, independent augmentation sampling per view maintains diversity and prevents trivial solutions during representation learning.



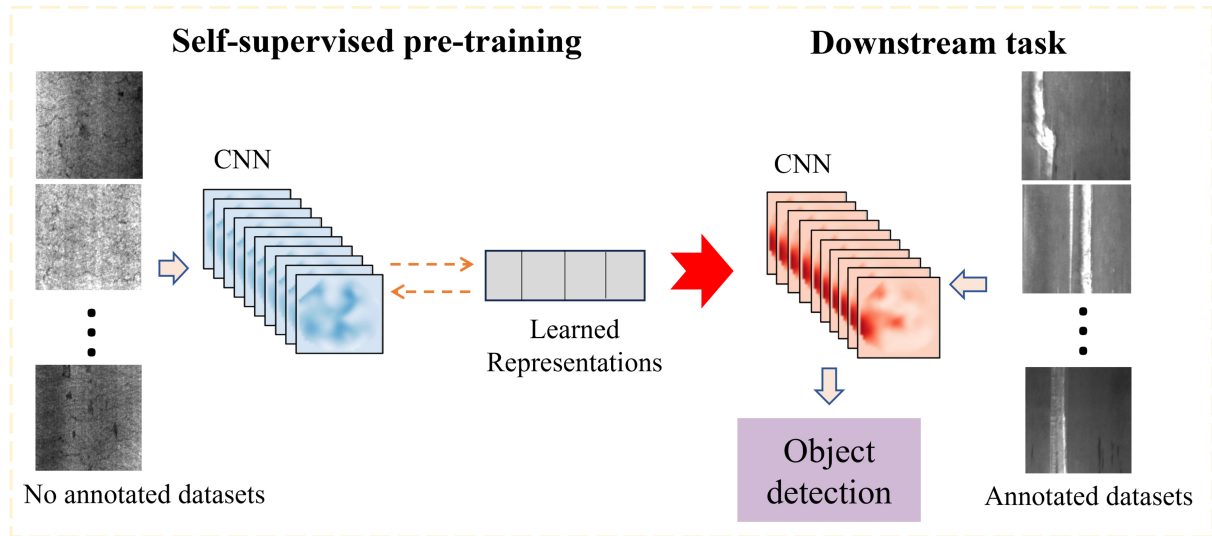


**Figure 1.** Presentation of defect samples from steel surface defect datasets. (A) Six representative defect types from the NEU dataset; (B) Defect examples from the SSDD dataset.

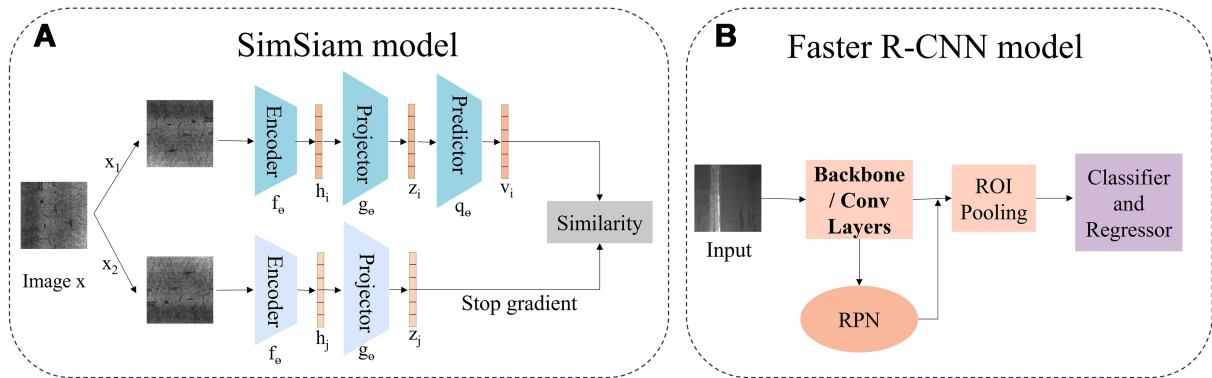
### SimSiam models and object detection with Faster R-CNN

In the self-supervised pre-training phase, we adopt SimSiam<sup>[38]</sup> for its architectural efficiency and deployment advantages. Unlike MOCO - which maintains a dynamic negative sample dictionary requiring substantial memory - or SimCLR - which depends on computationally intensive large batches - SimSiam eliminates these overheads via its stop-gradient mechanism and symmetric predictors. This approach achieves robust feature learning without negative samples or momentum encoders while preventing feature collapse. Figure 3A illustrates the SimSiam architecture. The model processes two augmented views ( $x_1, x_2$ ) of an input image  $x$ . Each view passes through an encoder  $f$  [backbone network + projection multi-layer perceptron (MLP)  $h$ ] to extract feature representations. The encoder output is further condensed by a projector before entering a predictor. The model then computes similarity between the two feature representations - a core self-supervised learning mechanism - enabling unsupervised feature acquisition without external annotations.

(1) Optimizer: stochastic gradient descent (SGD) is employed for pre-training. The learning rate is determined using



**Figure 2.** Flowchart: object detection process utilizing self-supervised learning.



**Figure 3.** Model architecture. (A) SimSiam model; (B) Faster R-CNN model. R-CNN: Region-based convolutional neural network.

$$\text{Learning Rate} = \frac{\text{lr} \times \text{Batch Size}}{256} \quad (\text{lr}_0 = 0.05) \quad (1)$$

The learning rate follows a cosine decay schedule, which is mathematically expressed as:

$$\text{Learning Rate} (t) = \text{lr} \times \cos \frac{t \times \pi}{T} \quad (2)$$

where  $t$  is the current epoch, and  $T$  is the total number of epochs. Additionally, the optimizer includes a weight decay of 0.0001 and a momentum of 0.90. The default batch size is 512 and batch normalization (BN) is implemented.

(2) Projection MLP: Each fully connected layer within the projected MLP component of the coding network is succeeded by BN. The fully connected output layer does not use rectified linear unit (ReLU) activation. The hidden layer of the fully connected network has a dimension of 2,048 and the MLP comprises three fully connected layers.

(3) Prediction MLP: BN is also applied to the prediction MLP; however, the output fully connected layer is not followed by BN or ReLU activation. This MLP comprises two fully connected layers, with the first layer having input and output dimensions of 2,048, and the second layer having an output dimension of 512.

The NEU dataset was subjected to pre-training across a range of 100-800 epochs, employing a batch size of 64 and a learning rate of 0.1. This process yielded eight distinct sets of pre-trained weights, designated as NEU-100e to NEU-800e. Concurrently, the SSDD dataset underwent pre-training for 100-400 epochs, yielding four unique sets of pre-trained weights, designated as SSDD-100e to SSDD-400e, with a batch size of 128 and a learning rate of 0.2. Furthermore, the amalgamated SSDD + NEU dataset was pre-trained for 100 and 200 epochs, yielding two additional sets of pre-trained weights, termed SSDD + NEU-100e and SSDD + NEU-200e. Subsequently, the aforementioned pre-trained weights are applied to the downstream tasks for object detection.

For object detection, we selected Faster R-CNN<sup>[42]</sup> [Figure 3B] - a two-stage framework that generates region proposals before classification/refinement. Its multi-scale detection capability and proven accuracy suit defect detection tasks prioritizing precision over speed.

We implement Faster R-CNN with ResNet18<sup>[43]</sup> [Figure 4] as the backbone. To preserve generic features while adapting to defects:

Stage 0 freezes the initial feature extractors: conv1 (7 × 7 convolution), bn1 (BN), ReLU activation, and maxpool (3 × 3 max pooling).

Stage 1 freezes the first residual group: layer1 (containing four 3 × 3 convolutions with two skip connections).

This allows layers 2-4 to fine-tune defect-specific patterns, leveraging transfer learning from pre-trained weights.

In Faster R-CNN, the region proposal network (RPN) processes the backbone-generated feature map to filter anchor boxes and generate region proposals. The RPN classifies regions as foreground (object-containing) or background but cannot identify object categories. During the second stage, positive proposals and corresponding feature map regions undergo ROI Pooling for dimensional standardization before entering the ROI Head. Here, convolutional and fully connected layers classify objects and refine bounding box coordinates to produce final detections.

### Evaluation index

Quantitative evaluation of steel surface defect detection methods employed two primary metrics: mean average precision (mAP) and mAP at intersection over union (IoU) threshold 0.5 (mAP<sub>50</sub>).

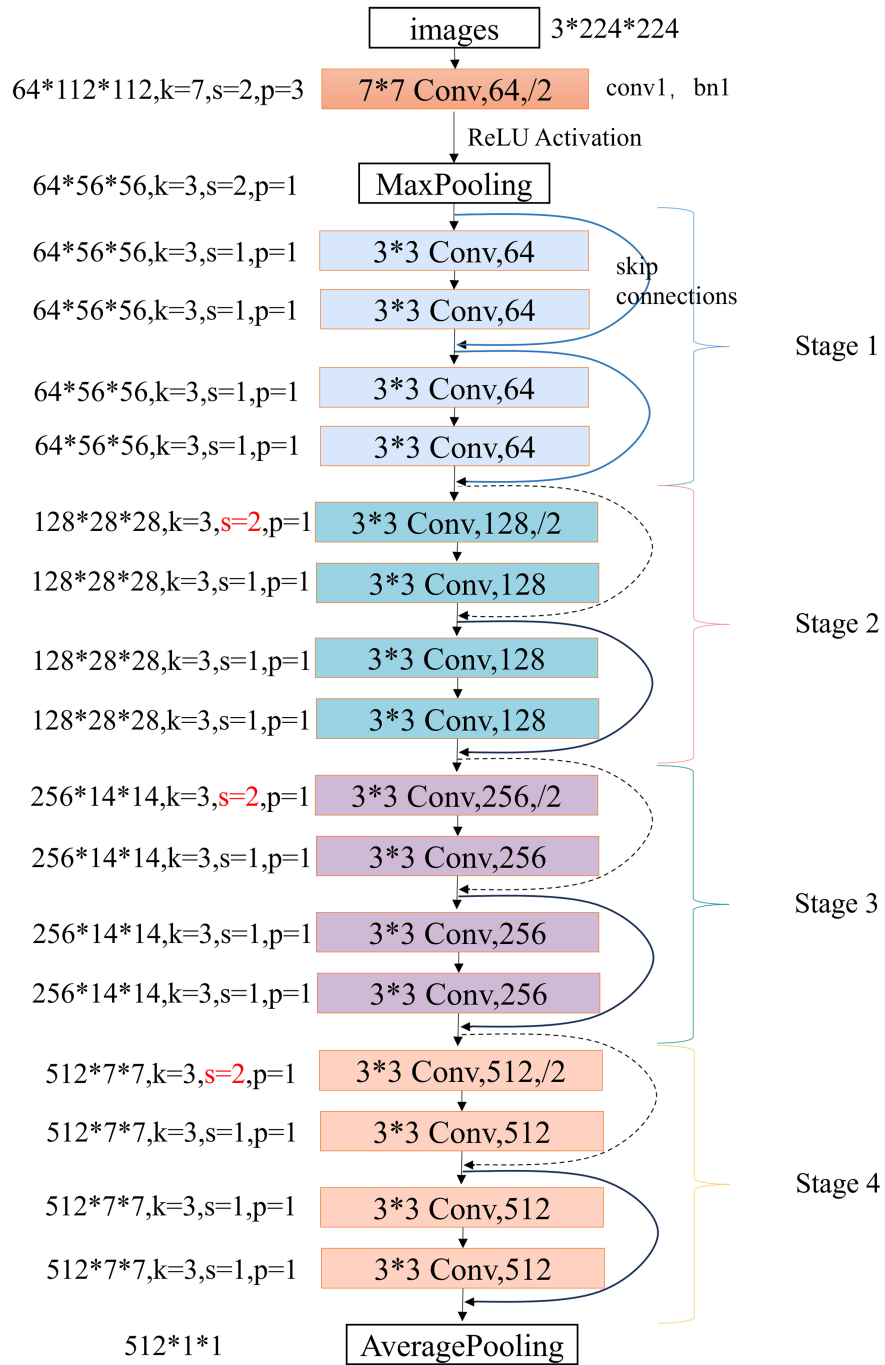
IoU measures bounding box alignment between predictions and ground truth, defined as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (3)$$

mAP derivation:

$$\text{Precision (P)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4)$$

$$\text{Recall (R)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (5)$$



**Figure 4.** Schematic representation of the ResNet18 architecture.

$AP^{[44]}$  is area under the Precision-Recall curve (P: y-axis, R: x-axis) per category. The mAP is mean of AP across all categories.

## RESULTS AND DISCUSSION

### Results

#### *Benchmark test on NEU-DET*

To evaluate the effectiveness of the self-supervised learning approach for steel surface defect detection, we established baselines on the NEU-DET dataset (six defect categories; 1,800 total images: 1,080 training, 360 validation, 360 testing; [Table 1](#)) using Faster R-CNN with random weights and weights from an ImageNet pre-trained ResNet18. These baselines provide comparative benchmarks for subsequent self-supervised learning evaluations.

Faster R-CNN was trained for 24 epochs on the NEU-DET training split, and the final accuracy is shown in [Figure 5](#). The object detection accuracy began to stabilize after 18 epochs. After 24 epochs of training with randomly initialized weights, the achieved mAP and mAP<sub>50</sub> values were 0.0880 and 0.2800, respectively. Conversely, after 24 epochs with weights from the ImageNet pre-trained ResNet18 model, the corresponding mAP and mAP<sub>50</sub> values were 0.3800 and 0.7730, respectively.

#### *Detailed analysis of object detection on NEU-DET*

##### “Frozen\_stages” on object detection

We trained a contrastive representation model via self-supervision on two unlabeled datasets (NEU, SSDD) until loss convergence. The resulting weights were transferred to steel defect detection, with fine-tuning and validation performed on NEU-DET. During detection, varying frozen\_stages (0 vs. 1) revealed superior mAP at frozen\_stages = 1 [[Figure 6](#)]. This optimal configuration was consequently adopted.

##### Number of pre-training epochs on object detection

To determine whether downstream detection accuracy correlates with pre-training epoch count, we transferred different pre-trained models to the object detection task following pre-training on the NEU and SSDD datasets. We evaluated the effect of different pre-training epochs on the object detection performance using the NEU-DET dataset. [Figure 7A](#) depicts the mAP for object detection across 24 epochs with the use of pre-training weights NEU-100e to NEU-800e. These results indicate that the accuracy of object detection initially increased with the number of pre-training epochs, then experienced a slight decline, and ultimately stabilized. As shown in [Figure 7B](#), the six types of pre-trained models from SSDD, when transferred to the object detection task, produced similar results, with the peak mAP reaching 0.3850, corresponding to the SSDD-200e pre-trained model.

##### Construction of the pre-training dataset

This section analyzes how pre-training datasets influence defect detection performance on NEU-DET. We transferred weights from three protocols - NEU, SSDD, and SSDD + NEU (trained across various epochs) - to the detection task, recording mAP and mAP<sub>50</sub> in [Table 2](#). On the SSDD dataset, the model attained its peak mAP of 0.3850 and mAP<sub>50</sub> of 0.7680 at 200 epochs. Beyond this point, as pre-training progressed to 300 and 400 epochs, both metrics exhibited a decline. In contrast, when trained on the SSDD + NEU dataset, the model's performance remained consistently robust across varying pre-training durations, sustaining mAP values above 0.3780 and mAP<sub>50</sub> scores above 0.7420. Notably, models initialized with NEU-derived pre-trained weights consistently underperformed compared to those pre-trained on SSDD or SSDD + NEU datasets.

Self-supervised learning viability for steel defect detection was assessed by transferring pre-trained weights to NEU-DET and comparing against benchmarks. As shown in [Figure 8A](#), the model initialized with random weights exhibited the lowest performance, with a mAP of 0.0880. In contrast, SSDD-200e weights



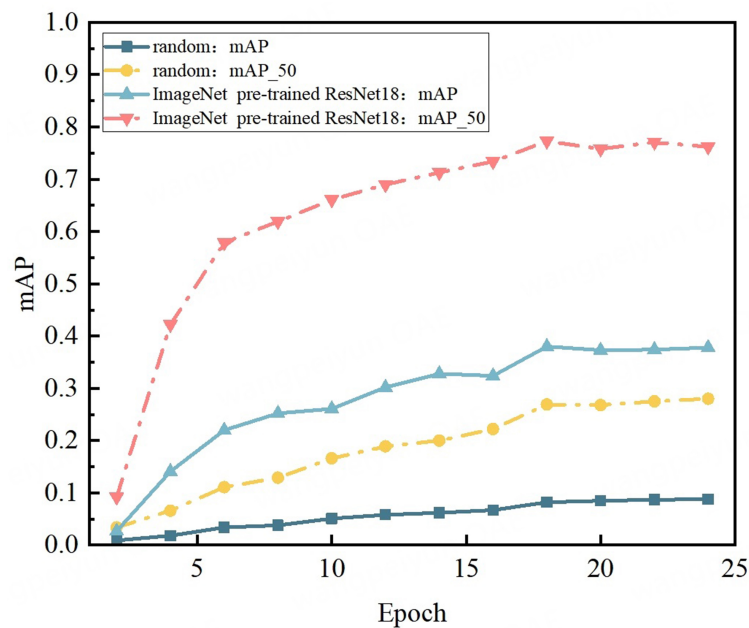
**Table 1. Details of the NEU-DET dataset**

Split	Images	Annotations	Categories
Train	1,080 items	2,488 items	6 items
Test	360 items	855 items	6 items
Val	360 items	846 items	6 items

**Table 2. Performance metrics of the NEU-DET dataset: mAP and mAP\_50**

Performance metrics	NEU		SSDD				SSDD + NEU	
	100e	400e	100e	200e	300e	400e	100e	200e
mAP	0.3160	0.3700	0.3790	<b>0.3850</b>	0.3790	0.3780	0.3810	0.3810
mAP_50	0.6910	0.7550	0.7630	<b>0.7680</b>	0.7420	0.7550	0.7540	0.7520

The defect detection results corresponding to the SSDD-200e weight are the best, with the optimal values emphasized in bold. mAP: Mean average precision.



**Figure 5.** Comparative analysis of mAP and mAP\_50 for object detection on the NEU-DET dataset with random initialization and ImageNet pre-trained ResNet18. mAP: Mean average precision.

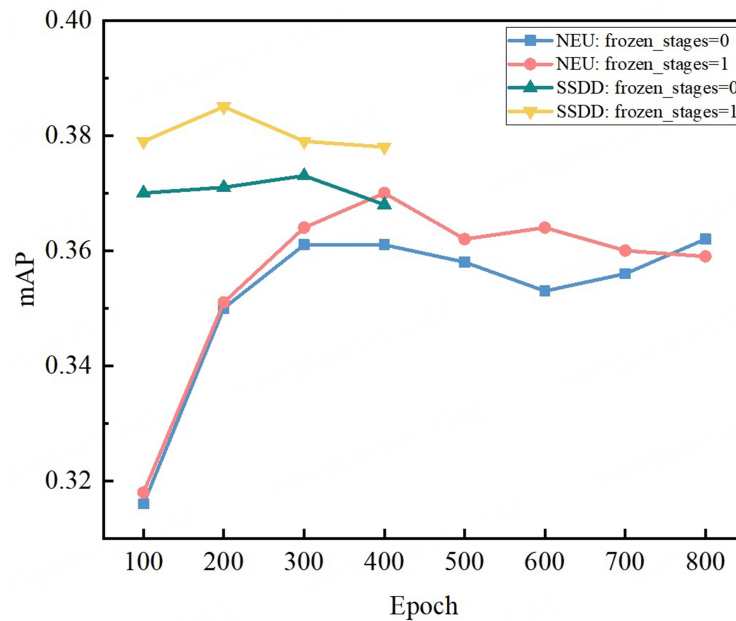
achieved peak mAP (0.3850), exceeding ImageNet-pre-trained ResNet18 (0.3800). SSDD + NEU weights showed strong performance with mAP of 0.3810 for both 100e and 200e.

**Figure 8B** compares mAP\_50 across initialization methods. All self-supervised weights outperformed random initialization, with SSDD-200e achieving peak mAP\_50 (0.7680) - approaching ImageNet-pre-trained performance (0.7730).

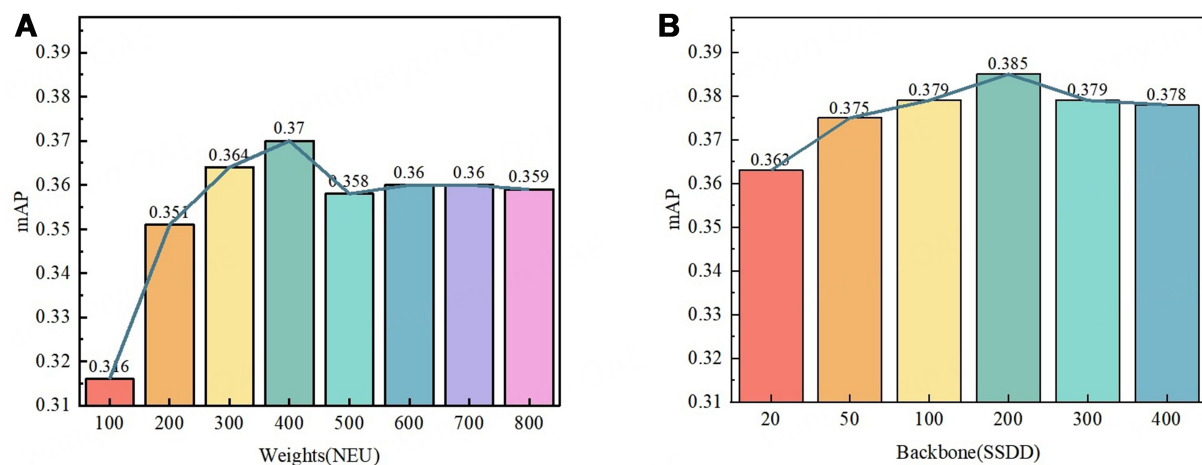
#### Detection visualization

In the experiment using NEU-400e pre-trained weights on the NEU-DET dataset for defect detection, the results demonstrated the model's effectiveness and reliability. **Figure 9** presents six defect cases (crazing,





**Figure 6.** mAP comparison across “frozen\_stages” configurations. mAP: Mean average precision.



**Figure 7.** Detection detection performance on NEU-DET with varying pre-trained weights. (A) mAP using NEU-100e to 800e weights; (B) mAP using SSDD-20e to 400e weights. mAP: Mean average precision.

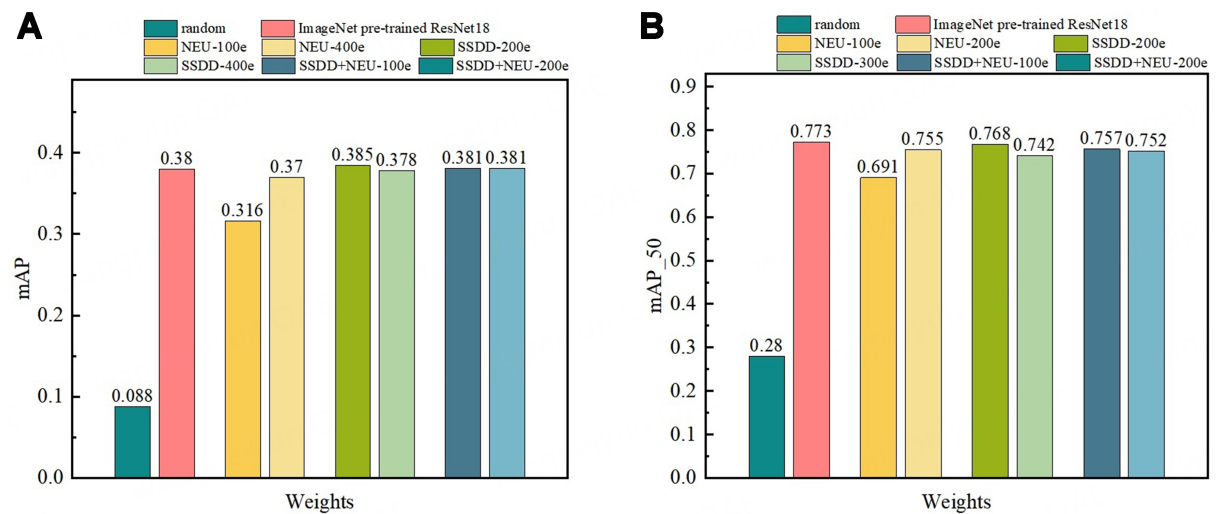
rolled-in scale, inclusion, scratches, patches, and pitted surface), with the left side showing the final detection outcomes at an 85% threshold and the right side displaying the corresponding defect detection confidence levels. Each type of defect shown in the figures is well-detected with high confidence.

Notably, Figure 9D shows simultaneous detection of patches (98.5%) and scratches (99.5%) without cross-interference, confirming multi-defect recognition capability in complex scenarios.

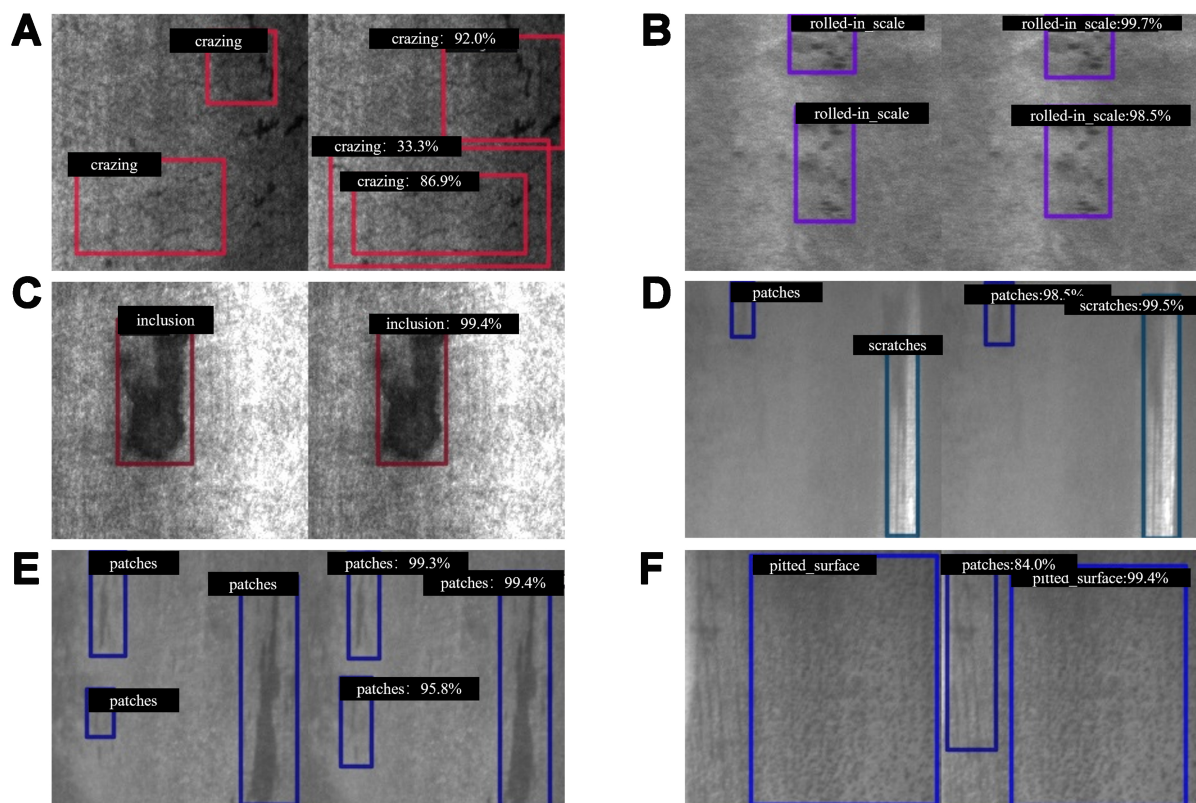
## Discussion

### *Performance evaluation of self-supervised learning on defect detection tasks*

Utilizing a comparative learning framework, our self-supervised learning model extracts meaningful



**Figure 8.** NEU-DET detection performance with varied pre-trained weights vs. benchmarks. (A) Comparative evaluation of mAP for object detection across NEU, SSDD, and SSDD + NEU pre-trained weights in conjunction with benchmark test results; (B) Comparative evaluation of mAP<sub>50</sub> for object detection across NEU, SSDD, and SSDD + NEU pre-trained weights in conjunction with benchmark test results. mAP: Mean average precision.



**Figure 9.** Visual examples of detected defects on NEU-DET. (A) Crazing; (B) Rolled-in\_scale; (C) Inclusion; (D) Scratches; (E) Patches; (F) Pitted\_surface. Left column: ground truth annotations; Right column: model predictions with confidence scores.

representations from unlabeled image data. Experimental results demonstrate that employing pre-trained weights from NEU and SSDD for defect detection on the NEU-DET dataset yields an approximately slight

mAP improvement when freezing the first layer compared to full-model fine-tuning (frozen\_stages = 0). However, the freezing strategy requires dynamic adjustment based on data scale and material diversity.

This study investigated the impact of varying pre-trained epochs during the self-supervised pre-training phase on object detection performance on the NEU-DET dataset. As shown in Figure 7, there is a significant correlation between pre-trained epochs and detection accuracy for both the NEU and SSDD datasets. The mAP initially rises but then plateaus and slightly declines after reaching an optimal number of epochs. This implies that increasing pre-training epochs beyond a certain point does not significantly enhance detection capability and may even cause slight performance degradation. These findings highlight the importance of selecting an appropriate number of pre-training epochs to optimize model performance. It is worth noting that the NEU dataset consists of 1,800 images, while the SSDD dataset contains 20,272 images. This difference in dataset sizes explains why the NEU dataset required pre-training epochs ranging from 100e to 800e, whereas the SSDD dataset only required epochs from 100e to 400e. The larger SSDD dataset provides more diverse samples for the model to learn from, thus requiring relatively fewer epochs to achieve satisfactory performance. In contrast, the smaller NEU dataset necessitates more epochs to ensure the model can adequately learn the features needed for effective defect detection.

For the NEU dataset, as the number of pre-training epochs increased from 100e to 400e, both mAP and mAP<sub>50</sub> showed an upward trend, indicating that the model gradually learned more effective features from the NEU dataset, thereby improving its ability to detect steel surface defects. However, after a certain number of epochs, the increase in mAP and mAP<sub>50</sub> slowed down and began to stabilize, suggesting that the model's performance may reach a saturation point with further pre-training epochs. Regarding the SSDD dataset, the model achieved the highest mAP of 0.3850 and mAP<sub>50</sub> of 0.7630 at 200e. When the pre-training epochs continued to increase to 300e and 400e, both mAP and mAP<sub>50</sub> exhibited a downward trend. This might be due to overfitting to the SSDD dataset as the number of pre-training epochs increased beyond the optimal range, reducing the model's generalization ability for steel surface defect detection. As for the SSDD + NEU dataset, the model demonstrated relatively stable performance across different pre-training epochs, maintaining mAP values above 0.3780 and mAP<sub>50</sub> values above 0.7420. This indicates that combining the SSDD and NEU datasets for pre-training helps enhance the model's robustness and generalization ability. The SSDD + NEU-100e configuration achieved a mAP of 0.3810 and a mAP<sub>50</sub> of 0.7540, while the SSDD + NEU-200e configuration obtained a mAP of 0.3810. Although the performance was slightly lower than that of the SSDD-200e configuration, it still outperformed the baseline model using ImageNet pre-trained ResNet18 weights (mAP of 0.3800). This suggests that the fusion of multiple datasets can provide the model with more diverse features, enabling better adaptation to the complexity of steel surface defect detection tasks.

Overall, the choice of pre-training dataset and the number of pre-training epochs significantly influence the model's performance. The SSDD-200e configuration achieved the highest mAP value, while the SSDD + NEU configuration exhibited stable and relatively superior performance, highlighting the advantages of combining multiple datasets for pre-training. These findings offer valuable insights for future research in optimizing pre-training strategies to enhance the performance of self-supervised learning models in steel surface defect detection.

Our method achieves mAP<sub>50</sub> = 0.7680 on NEU-DET, surpassing SSDD-Net<sup>[45]</sup> by 4%, and YOLOv7<sup>[46]</sup> (mAP<sub>50</sub> = 0.7290). Transfer learning with our weights consistently outperforms random initialization, confirming self-supervised efficacy for steel defect detection.

Self-supervised learning methods have demonstrated impressive accuracy on the defect detection task using the NEU-DET dataset. To further demonstrate the superiority of the prediction accuracy of the SimSiam model, we also present the prediction results of another self-supervised model (SimCLR) on this defect detection task in [Supplementary Figure 1](#). The results show that the mAP and mAP<sub>50</sub> prediction values of the SimSiam model are consistently higher than those of SimCLR.

Cross-dataset transferability demonstrates that learned features generalize across metal types, capturing universal surface characteristics [[Supplementary Table 1](#)]. Scratch and patch defect categories exhibit robust detection performance, achieving peak mAP<sub>50</sub> values of 0.9780 and 0.9320, respectively, attributable to their distinctive morphological characteristics. Conversely, crack defects demonstrate significantly diminished accuracy with maximum mAP<sub>50</sub> of only 0.4720, likely resulting from subtle visual manifestations or inadequate training sample representation. Regarding weight initialization strategies, SSDD and SSDD + NEU pre-trained weights show superior domain-specific adaptation, consistently outperforming both random initialization and NEU pre-trained configurations, while ImageNet-pre-trained ResNet18 maintains competitive detection capability across most defect categories.

#### *Limitations of self-supervised learning in object detection applications*

This study reveals that self-supervised pre-trained weights transferred to downstream defect detection tasks consistently underperform ImageNet-supervised ResNet18 baselines in mAP<sub>50</sub> metrics. While self-supervised approaches achieve competitive accuracy, their inability to surpass large-scale supervised pre-training underscores fundamental limitations in feature generalizability. We attribute this gap to two interrelated factors: (1) constrained diversity in domain-specific unlabeled datasets (e.g., steel defect imagery), which restricts comprehensive feature representation learning; and (2) inherent architectural discontinuities between contrastive pre-training (e.g., SimSiam) and task-specific fine-tuning, inducing feature subspace mismatches that undermine transfer efficacy.

Practical implementation faces additional challenges: Computational demands for industrial-scale datasets necessitate prohibitive pre-training durations. Current data augmentation pipelines, though effective for texture-based defects, fail to capture the full spectral and topological variance of real-world metallic degradation phenomena.

Future research should focus on three cooperation pathways: hybrid semi-supervised frameworks that leveraging limited labeled exemplars alongside abundant unlabeled data to bridge representation gaps; and meta-transfer learning for few-shot adaptation to novel defect morphologies. These advances would establish resource-efficient pipelines adaptable to the dynamic feature landscapes of industrial metal inspection without compromising detection robustness.

## CONCLUSIONS

Drawing inspiration from groundbreaking applications of self-supervised learning in scientific research, such as AlphaFold2 - which leverages unlabeled data to solve complex problems - our study underscores the potential of self-supervised methods as viable alternatives or supplements to traditional supervised learning techniques in steel surface defect detection. The innovative use of unlabeled data significantly reduces reliance on manual annotation while enhancing scalability, making it an attractive approach for various surface defect identification tasks.

In this study, we present a self-supervised learning method for detecting surface defects in steel materials. Our method achieves high accuracy in target detection on the NEU-DET steel defects dataset, without reliance on a large amount of labeled data, with mAP and mAP\_50 values of 0.3850 and 0.7680, respectively, which demonstrates the effectiveness of the method. The above discussion highlights the value of self-supervised learning for detecting surface defects in steels. While further improvements are possible in the mAP\_50 scores obtained in this study, our results demonstrate the potential of these methods as potential alternatives or supplements to traditional supervised learning methods. Future research efforts ought to concentrate on refining the pre-training process, exploring the integration of various datasets, and developing sophisticated data enhancement techniques to enhance the capabilities of self-supervised models in this domain. This study not only advances the state of the art in steel surface defect detection but also provides guidance for constructing robust image analysis models with minimal reliance on labeled data across related fields.

## DECLARATIONS

### Authors' contributions

Made substantial contributions to conception and design of this review, writing and editing: Hu, S.; Zhang, Y.; Xu, W.

Made substantial contributions to collation of literature, figures preparation, and writing: Hu, S.; Ma, X.; Zhang, Y.; Xu, W.

Performed data analysis, discussion and writing review: Hu, S.; Ma, X.; Zhang, Y.; Xu, W.

Provided administrative, technical, and material support: Zhang, Y.; Xu, W.

### Availability of data and materials

The original contributions presented in this study are included in the article/[Supplementary Materials](#). Further inquiries can be directed to the corresponding author(s).

### Financial support and sponsorship

The research was supported by the National Key Research and Development Program of China (No. 2022YFB3707500) and the National Natural Science Foundation of China (No. 52304392).

### Conflicts of interest

All authors declared that there are no conflicts of interest.

### Ethical approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Copyright

© The Author(s) 2025.

## REFERENCES

1. Ren, B.; Wang, C.; Zhang, Y.; Wei, X.; Xu, W. Industrial big data analysis strategy based on automatic data classification and interpretable knowledge graph. *J. Mater. Inf.* **2025**, *5*, 2. [DOI](#)
2. Zhou, C.; Lu, Z.; Lv, Z.; et al. Metal surface defect detection based on improved YOLOv5. *Sci. Rep.* **2023**, *13*, 20803. [DOI](#)
3. Park, J. K.; Kwon, B. K.; Park, J. H.; Kang, D. J. Machine learning-based imaging system for surface defect inspection. *Int. J. Precis. Eng. Manuf. Green. Tech.* **2016**, *3*, 303-10. [DOI](#)
4. Han, S.; Wang, C.; Zhang, Y.; Xu, W.; Di, H. Employing deep learning in non-parametric inverse visualization of elastic-plastic



- mechanisms in dual-phase steels. *Mater. Genome. Eng. Adv.* **2024**, *2*, e29. DOI
5. Qiao, Z.; Shi, D.; Yi, X.; Shi, Y.; Zhang, Y.; Liu, Y. UEFPN: unified and enhanced feature pyramid networks for small object detection. *ACM. Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1-21. DOI
  6. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert. Syst. Appl.* **2021**, *172*, 114602. DOI
  7. Zhu, X.; Wang, Q.; Zhang, B.; Sun, Z.; Yu, J.; Qian, S. An improved feature enhancement CenterNet model for small object defect detection on metal surfaces. *Adv. Theory. Simul.* **2024**, *7*, 2301230. DOI
  8. Yu, J.; Cheng, X.; Li, Q. Surface defect detection of steel strips based on anchor-free network with channel attention and bidirectional feature fusion. *IEEE. Trans. Instrum. Meas.* **2022**, *71*, 1-10. DOI
  9. He, Q.; Li, Z.; Yang, W. LMFE-RDD: a road damage detector with a lightweight multi-feature extraction network. *Multimed. Syst.* **2024**, *30*, 176. DOI
  10. Sun, L.; Cai, Z.; Liang, K.; Wang, Y.; Zeng, W.; Yan, X. An intelligent system for high-density small target pest identification and infestation level determination based on an improved YOLOv5 model. *Expert. Syst. Appl.* **2024**, *239*, 122190. DOI
  11. Yuan, Y.; Wu, Y.; Zhao, L.; Chen, H.; Zhang, Y. Multiple object detection and tracking from drone videos based on GM-YOLO and multi-tracker. *Image. Vis. Comput.* **2024**, *143*, 104951. DOI
  12. Zhu, Y.; Ai, Z.; Yan, J.; Li, S.; Yang, G.; Yu, T. NATCA YOLO-based small object detection for aerial images. *Information* **2024**, *15*, 414. DOI
  13. Wang, Y.; Xia, H.; Yuan, X.; Li, L.; Sun, B. Distributed defect recognition on steel surfaces using an improved random forest algorithm with optimal multi-feature-set fusion. *Multimed. Tools. Appl.* **2017**, *77*, 16741-70. DOI
  14. Wang, H.; Li, M.; Wan, Z. Rail surface defect detection based on improved Mask R-CNN. *Comput. Electr. Eng.* **2022**, *102*, 108269. DOI
  15. Liu, L. J.; Zhang, Y.; Karimi, H. R. Resilient machine learning for steel surface defect detection based on lightweight convolution. *Int. J. Adv. Manuf. Technol.* **2024**, *134*, 4639-50. DOI
  16. Shi, X.; Zhou, S.; Tai, Y.; Wang, J.; Wu, S.; Liu, J. An improved faster R-CNN for steel surface defect detection. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, Shanghai, China. Sep 26-28, 2022. IEEE; 2022. p. 1-5. DOI
  17. Akhyar, F.; Liu, Y.; Hsu, C. Y.; Shih, T. K.; Lin, C. Y. FDD: a deep learning-based steel defect detectors. *Int. J. Adv. Manuf. Technol.* **2023**, *126*, 1093-107. DOI
  18. Hong, Y.; Wang, Z.; Wu, W.; et al. Steel surface defect detection based on denoising diffusion implicit models with data augmentation. In *2024 8th International Conference on Imaging, Signal Processing and Communications (ICISPC)*, Fukuoka, Japan. Jul 19-21, 2024. IEEE; 2024. pp. 15-9. DOI
  19. Kadam, S. Advancements in image detection: a comprehensive approach to object localization and classification using deep learning techniques. *Int. J. Multidiscip. Res.* **2024**, *6*, 27133. DOI
  20. Fang, Z.; Roy, K.; Xu, J.; Dai, Y.; Paul, B.; Lim, J. B. P. A novel machine learning method to investigate the web crippling behaviour of perforated roll-formed aluminium alloy unflipped channels under interior-two flange loading. *J. Build. Eng.* **2022**, *51*, 104261. DOI
  21. Balestriero, R.; Ibrahim, M.; Sobal, V.; et al. A cookbook of self-supervised learning. *arXiv* **2023**, arXiv:2304.12210. <https://doi.org/10.48550/arXiv.2304.12210>. (accessed 10 Jul 2025)
  22. Wang, Y.; Li, T.; Zong, H.; et al. Self-supervised probabilistic models for exploring shape memory alloys. *npj. Comput. Mater.* **2024**, *10*, 185. DOI
  23. Magar, R.; Wang, Y.; Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *npj. Comput. Mater.* **2022**, *8*, 231. DOI
  24. Fu, N.; Wei, L.; Hu, J. Physics-guided dual self-supervised learning for structure-based material property prediction. *J. Phys. Chem. Lett.* **2024**, *15*, 2841-50. DOI
  25. Zhang, S.; Wang, W. Y.; Wang, X.; et al. Large language models enabled intelligent microstructure optimization and defects classification of welded titanium alloys. *J. Mater. Inf.* **2024**, *4*, 34. DOI
  26. Kim, S.; Ryu, S. Effect of surface and internal defects on the mechanical properties of metallic glasses. *Sci. Rep.* **2017**, *7*, 13472. DOI
  27. Masci, J.; Meier, U.; Ciresan, D.; Schmidhuber, J.; Fricout, G. Steel defect classification with max-pooling convolutional neural networks. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia. Jun 10-15, 2012. IEEE; 2012. p. 1-6. DOI
  28. Tian, R.; Jia, M. DCC-CenterNet: a rapid detection method for steel surface defects. *Measurement* **2022**, *187*, 110211. DOI
  29. Zabin, M.; Kabir, A. N. B.; Kabir, M. K.; Choi, H. J.; Uddin, J. Contrastive self-supervised representation learning framework for metal surface defect detection. *J. Big. Data.* **2023**, *10*, 145. DOI
  30. Xu, R.; Hao, R.; Huang, B. Efficient surface defect detection using self-supervised learning strategy and segmentation network. *Adv. Eng. Inform.* **2022**, *52*, 101566. DOI
  31. Zhang, S.; Zhang, Q.; Gu, J.; Su, L.; Li, K.; Pecht, M. Visual inspection of steel surface defects based on domain adaptation and adaptive convolutional neural network. *Mech. Syst. Signal. Process.* **2021**, *153*, 107541. DOI
  32. Sirotkin, K.; Escudero-Viñolo, M.; Carballeira, P.; García-Martín, Á. Improved transferability of self-supervised learning models through batch normalization finetuning. *Appl. Intell.* **2024**, *54*, 11281-94. DOI
  33. Geng, X.; Wang, F.; Wu, H. H.; et al. Data-driven and artificial intelligence accelerated steel material research and intelligent manufacturing technology. *Mater. Genome. Eng. Adv.* **2023**, *1*, e10. DOI



34. Bachman, P.; Hjelm, R. D.; Buchwalter, W. Learning representations by maximizing mutual information across views. *arXiv* **2019**, arXiv:1906.00910. <https://doi.org/10.48550/arXiv.1906.00910>. (accessed 10 Jul 2025)
35. Pöppelbaum, J.; Chadha, G. S.; Schwung, A. Contrastive learning based self-supervised time-series analysis. *Appl. Soft. Comput.* **2022**, *117*, 108397. [DOI](#)
36. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA. Jun 13-19, 2020. IEEE; 2020. pp. 9726-35. [DOI](#)
37. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR; 2020. pp. 1597-607. <https://proceedings.mlr.press/v119/chen20j.html>. (accessed 10 Jul 2025)
38. Chen, X.; He, K. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA. Jun 20-25, 2021. IEEE; 2021. pp. 15745-53. [DOI](#)
39. Song, K.; Yan, Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, *285*, 858-64. [DOI](#)
40. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H. A survey on self-supervised learning: algorithms, applications, and future trends. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2024**, *46*, 9052-71. [DOI](#)
41. Zhao, Z.; Alzubaidi, L.; Zhang, J.; Duan, Y.; Gu, Y. A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations. *Expert. Syst. Appl.* **2024**, *242*, 122807. [DOI](#)
42. Huang, J.; Rathod, V.; Sun, C.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv* **2016**, arXiv:1611.10012. <https://doi.org/10.48550/arXiv.1611.10012>. (accessed 10 Jul 2025)
43. Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA. Jul 21-26, 2017. IEEE; 2017. pp. 936-44. [DOI](#)
44. Lin, T. Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO: common objects in context. *arXiv* **2014**, arXiv:1405.0312. <https://doi.org/10.48550/arXiv.1405.0312>. (accessed 10 Jul 2025)
45. Li, Z.; Wei, X.; Jiang, X. SSDD-Net: a lightweight and efficient deep learning model for steel surface defect detection. In *Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023*, Xiamen, China. Oct 13-15, 2023. Springer-Verlag; 2023. pp. 237-48. [DOI](#)
46. Li, M.; Wei, L.; Zheng, B. Steel surface defect detection based on improved YOLOv7. In *2024 4th International Conference on Computer, Control and Robotics (ICCCR)*, Shanghai, China. Apr 19-21, 2024. IEEE; 2024. pp. 51-5. [DOI](#)