International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 7 No. 5 January – February - 2026

Leveraging Contrastive Learning and Masked Autoencoders for Robust Surface Defect Classification in Data-Scarce Environments

Mr. Patil Abhijit Bhaskarrao[1] & Prof. Ms. Vijaya D. Desai[2]
Computer Science & Engineering –II Year AMGOI,Vathar
Assistant Professor, Department of Computer Science & Engineering, AMGOI Vathar
Corresponding Author – Mr. Patil Abhijit Bhaskarrao
DOI - 10.5281/zenodo.18898071

Abstract:

Surface defect detection in industrial settings is extremely challenging task due to the scarcity of annotated data and high variability in defect appearance. This study presents a comparative analysis of two self-supervised learning approaches—contrastive learning (SimCLR) and masked autoencoders (MAE)—for metal surface defect classification using the NEU dataset. The models were pretrained with unlabeled data and fine-tuned under low supervision. SimCLR performed better with 100% accuracy and F1-score, surpassing MAE with an accuracy of 91%. On the low-label setting (10% data), SimCLR had good performance (83% accuracy) and MAE had poor performance (39% accuracy). The findings attest to the superiority of SimCLR's robustness, efficiency, and applicability to industrial defect inspection. The study exhibits the promise of self-supervised pretraining in lessening dependence on labeled data and suggests a scalable approach to real-world visual inspection problems.

Keywords: SimCLR, MAE, Surface detection, Neu dataset

Introduction:

Surface defect detection plays a critical role in quality assurance across manufacturing sectors such as metallurgy, electronics, and automotive production. Traditional supervised learning methods have achieved notable success but are heavily reliant on large quantities of annotated defect data. Due to the diverse nature of defects and the scarcity of labeled samples, especially in complex or high-precision industrial environments, these methods often face generalization issues and reduced robustness. Furthermore, defect characteristics can be subtle or vary across textures and lighting conditions, making handcrafted or rule-based detection unreliable.

In response to these challenges, self-supervised learning (SSL) has emerged as a promising alternative for visual representation learning without reliance on labeled data. Techniques like masked autoencoders (MAEs) and contrastive learning have demonstrated impressive performance across vision domains. By learning invariant and discriminative features from unlabeled data, SSL facilitates efficient pretraining, even in data-constrained environments. Recent adaptations of SSL to domains like ECG, microscopy, and remote sensing further highlight its flexibility and potential for domain-specific applications.

Despite this progress, most existing SSL models are either domain-specific or not rigorously evaluated for surface defect tasks, leaving a gap in understanding their utility for industrial inspection. This research aims to address this gap by developing a tailored SSL pretraining framework for surface defect

detection and evaluating it under limited label conditions.

We pose the following questions: How effective are contrastive learning and MAEs for defect-specific feature learning? Can SSL outperform traditional supervised methods in data-scarce industrial scenarios? By answering these, we aim to empower manufacturers, researchers, and vision engineers to deploy more robust, scalable, and cost-effective defect detection systems.

Literature Review:

Recent advances in self-supervised learning (SSL) have demonstrated strong potential for improving visual representation learning, especially in domains like surface defect detection, where annotated data is scarce and expensive to obtain. This section reviews key developments related to contrastive learning, masked autoencoders (MAEs), and their integration, which inform the design of our proposed framework.

Surface defect detection tasks often suffer from a lack of annotated samples and a wide diversity of defect types. To address this, [2] proposed a self-supervised two-stage framework based on a cropping network and a variational autoencoder that learns to detect defects using only normal samples. Their method achieved high segmentation accuracy, although generalizability beyond railway datasets remains uncertain. Similarly, [4] introduced an adaptive transformer with contrastive learning in a meta-learning setup, demonstrating state-of-the-art (SOTA) results across three surface defect datasets. However, this method is computationally intensive and dependent on diverse datasets.[3] employed contrastive learning with Earth Mover's Distance on convolutional features to improve generalization for steel surface defect detection. They achieved significant performance gains, though evaluations were limited to steel surfaces. Additionally, [5] tackled noisy sample issues via a one-shot confident learning pipeline, showing promising results on the Kolektor SDD2 dataset, albeit using a relatively simple baseline.

In a similar vein, [19] applied masked autoencoder-based reconstruction to textured surfaces, achieving over 95% detection accuracy. Meanwhile, [18] proposed CoRe, a hybrid of contrastive and restorative SSL techniques, which outperformed supervised baselines on five datasets by better aligning the model with the characteristics of defect inspection tasks.Beyond surface defects, several studies have focused on enhancing MAE and contrastive learning techniques. [13] proposed CMGAE, which fuses contrastive objectives with graph masked autoencoders to capture both local and global structures—although limited to graph data. [14] provided theoretical insights into MAEs through a local contrastive lens, helping demystify their robustness and consistency, though no new model was proposed.

[11] presented a multi-view masked autoencoder with contrastive loss for general image representation, improving global feature extraction and achieving strong results on ImageNet. [10] extended this idea by proposing Contrastive Masked Autoencoders (CMAE), achieving 85.3% top-1 accuracy and highlighting the synergy between reconstruction and discriminative tasks. Similarly, [12] developed CAN, a scalable contrastive masked autoencoder combining CL, MAE, and noise prediction, optimized for large-scale learning.[17] investigated multimodal SSL with RGB-elevation contrastive MAE, enhancing remote sensing change detection—though not directly applicable to surface defects. [16] proposed pseudo-colorized masked cell image pretraining, surpassing SimCLR and MAE in microscopy tasks, reinforcing the value of domain-adapted

SSL. However, their focus remains on biomedical applications.

In the medical domain, [15] successfully adapted MAEs for ECG signal classification using multi-scale convolutions and transformer encoders, outperforming contrastive baselines. [20]) further refined ECG pretraining using multiview information bottlenecks in the time-frequency domain. Although these are task-specific, they reinforce MAE's adaptability to time-series signals.

[21] conducted a comparative study of MAE and contrastive learning on small medical imaging datasets, showing that MAE (via SparK) was more robust under limited fine-tuning scenarios. [22] adapted dual-branch transformer MAEs with contrastive loss for hyperspectral classification, setting new benchmarks in low-label settings—though their domain is remote from defect detection.Several works explore pathology and histology. [23] introduced a self-distillation MAE for histopathological understanding, integrating patch-level supervision to address false negatives in contrastive methods. [24] proposed GCMAE, which uses global contrast for training on whole-slide images, outperforming supervised baselines in pathology tasks, though cross-domain generalizability remains untested.[25] addressed the masked pretraining gap for 3D point cloud data using an inter-modal contrastive MAE combining images and point clouds. While their model set benchmarks in 3D segmentation, its relevance to 2D surface defect inspection is limited.

Building on this foundation, our work aims to develop a self-supervised pretraining framework tailored to surface defect detection, leveraging both contrastive and reconstruction objectives. Evaluate MAE and contrastive learning specifically for defect-relevant feature extraction. Assess the impact of pretraining on small labeled datasets, targeting practical deployment. Compare performance and efficiency with traditional supervised learning approaches, using consistent benchmarks. This study seeks to bridge the methodological strengths identified in prior work with the unique challenges of surface defect detection in industrial settings.

Methodology:

This study explores the effectiveness of contrastive and generative pretraining paradigms for surface defect classification in metallic components, leveraging the NEU surface defect dataset. We implemented two complementary self-supervised learning (SSL) strategies SimCLR for contrastive representation learning and Masked Autoencoders (MAE) for generative pretraining using vision transformers. Both approaches were later fine-tuned for downstream classification tasks. The architecture of it is shown in figure 1.

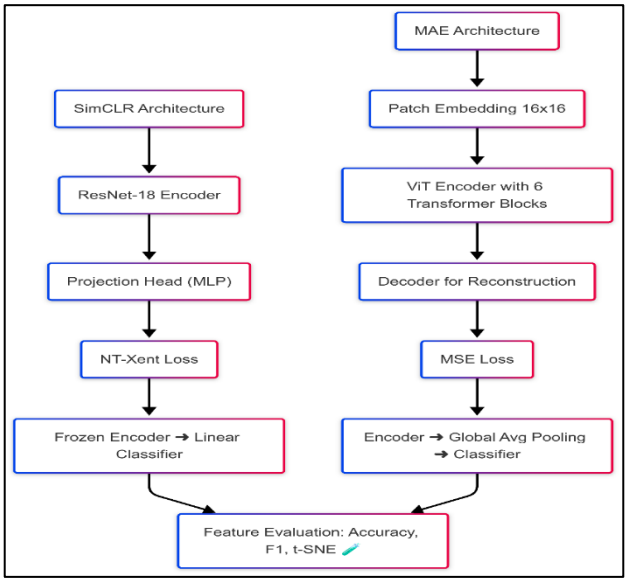
SimCLR Architecture:

The SimCLR model is based on a convolutional backbone from the ResNet-18 architecture. The final fully connected layer and global pooling operations were removed to retain the high-dimensional feature map output. This encoder maps input images into a latent embedding space, followed by a projection head used only during contrastive pretraining. A ResNet-18 model truncated at the penultimate layer (i.e., without the final fully connected layer), yielding a 512-dimensional feature vector. A multilayer perceptron (MLP) consisting of two linear layers with a ReLU activation in between, projecting the 512-dimensional features into a 128-dimensional contrastive space.

The Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) is used to bring positive pairs (augmented views of the same image) closer while pushing apart all other samples in the batch, with a temperature

hyperparameter of 0.5.After pretraining, the projection head is discarded and the frozen encoder is used as a feature extractor for classification. A simple linear classifier is appended for downstream fine-tuning.

Fig 1: Model architecture diagram



Masked Autoencoder (MAE) Architecture:

For the MAE, we implement a lightweight Vision Transformer (ViT)-inspired model composed of three principal components: patch embedding, transformer-based encoder blocks, and a reconstruction decoder. Input images of size 224×224 are divided into non-overlapping patches of size 16×16, yielding 196 patches per image. Each patch is linearly projected to a 512-dimensional embedding using a convolutional layer with a stride and kernel size equal to the patch size.

The encoder comprises a sequence of 6 transformer blocks, each featuring Layer normalization, Ulti-head self-attention (8 heads),Feed-forward networks (MLP) with GELU activations, Residual connections. During pretraining, 75% of the patch tokens are randomly masked. Only the remaining visible tokens are processed by the encoder. A lightweight decoder reconstructs the full set of image patches from the encoded tokens. It consists of a linear transformation to a lower-dimensional space (256 units), followed by ReLU activation and projection back to the pixel space of each patch. The model is trained using mean squared error (MSE) between the reconstructed image and the original input, allowing the model to learn context-aware representations.

For downstream tasks, the MAE is modified to include a classification head. A global average pooling is applied to the encoder's output sequence, A layer normalization and linear classifier are added, mapping the aggregated representation to the class logits.

Classification Heads and Fine-tuning:

Both pretrained models were adapted for classification. A linear layer maps the 512-dimensional encoded features to six output classes. The encoder outputs from the MAE are globally averaged, normalized, and passed through a linear classification head. During fine-tuning, only the classification heads were updated while freezing the encoder parameters. In a separate experiment, encoder weights were unfrozen and trained jointly for comparison. To assess the quality of the learned representations. Features from both models were extracted from the test dataset. A linear classifier was trained on these representations. Performance metrics (accuracy and F1-score) were reported.t-SNE visualizations were used to qualitatively assess feature separability in two-dimensional space.

Result and Discussion:

Experimental Setup:

All experiments were conducted on Kaggle using a T4 GPU to ensure efficient model training. The NEU Metal Surface Defects Data dataset was used. Images were resized to 224×224 pixels and processed in batches of 64. Training was performed in two phases: a 50-

epoch pretraining stage followed by a 30-epoch fine-tuning stage. The model was trained using a learning rate of 1e-3 with a temperature parameter of 0.5 for contrastive learning. All computations were executed on a CUDA-enabled device when available. The classification task involved six defect categories.

Evaluation of fine-tuned SimCLR model:

This section evaluate the performance of the SimCLR model after fine-tuning on labeled data. Initially, the model was pretrained using a contrastive learning framework where an encoder–projection head architecture based on ResNet-18 learned meaningful visual representations from unlabeled images via the NT-Xent loss. After pretraining, the encoder was frozen and paired with a linear classifier for supervised fine-tuning on the NEU metal surface defect dataset. The classifier was trained using cross-entropy loss across six defect categories. Finally, model performance was assessed on the test set using standard classification metrics to analyze its generalization capabilities.

Classification Report Analysis:

The classification report shown in table 1 summarizes the performance of the fine-tuned SimCLR model on the test set, which contains 72 images evenly distributed across six defect categories. As shown, the model achieved a perfect score across all key metrics—precision, recall, and F1-score—for every class: *Crazing*, *Inclusion*, *Patches*, *Pitted*, *Rolled*, and *Scratches*. Each class has a support of 12, indicating balanced representation in the test data.

Recision of 1.00 implies that the model made no false positive predictions for any class. Recall of 1.00 indicates that it correctly identified all true instances for each defect type. F1-score, the harmonic mean of precision and recall, is also 1.00, confirming overall robustness.

Table 1 : Classification report of SimCLR model

Class	Precision	Recall	F1-Score	Support
Crazing	1.00	1.00	1.00	12
Inclusion	1.00	1.00	1.00	12
Patches	1.00	1.00	1.00	12
Pitted	1.00	1.00	1.00	12
Rolled	1.00	1.00	1.00	12
Scratches	1.00	1.00	1.00	12

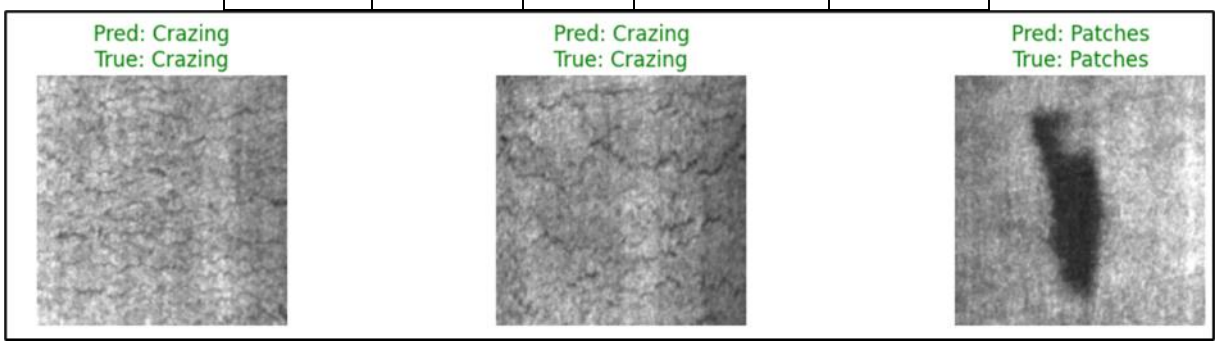


Fig 2 : Result visualization of SimCLR model

The macro average and weighted average are also 1.00, reflecting consistent performance across both balanced and potentially imbalanced class distributions. This exceptional result

suggests that the pretrained encoder effectively captured discriminative features, and the fine-tuning phase successfully adapted these features to the downstream classification task. Additionally, sample prediction images from figure 2 visually confirm the model's accuracy. These examples show correctly classified instances from each category, further demonstrating the effectiveness of the contrastive learning approach combined with supervised fine-tuning.

Evaluation of MAE vit Classification Model:

In this section, we evaluate the performance of a Vision Transformer (ViT) model trained using the Masked Autoencoder (MAE) framework on the NEU metal surface defect dataset. The model was first pretrained in a self-supervised manner by reconstructing masked image patches, allowing it to learn rich, generalized representations without labeled data. This encoder was then fine-tuned for classification by appending a linear classification head and training on labeled images. The architecture includes a patch embedding module, multiple transformer blocks, and a classification head operating on the global average of encoded tokens. Evaluation was conducted using standard classification metrics and supported with visualizations of correctly predicted samples to qualitatively assess the model's effectiveness.

Classification Report Analysis:

The classification report for the fine-tuned MAE ViT model from table 2 reveals strong performance across all six defect categories, demonstrating the effectiveness of masked autoencoding for self-supervised pretraining. The overall accuracy achieved is 91% on a test set of 1,656 images, with a macro and weighted average F1-score of 0.91, indicating consistent performance across classes.The model achieves perfect precision, recall, and F1-score (1.00) on the Patches class, suggesting it learned this defect pattern very well.Crazing, Inclusion, and Scratches also show strong performance, each achieving F1-scores close to 0.89.Pitted and Rolled have slightly lower precision but higher recall, indicating the model tends to correctly detect these defects more often than it avoids false positives.

Table 2 :Classification report of MAE-ViT model

Class	Precision	Recall	F1-Score	Support
Crazing	0.91	0.85	0.88	276
Inclusion	0.93	0.85	0.89	276
Patches	1	1	1	276
Pitted	0.83	0.92	0.87	276
Rolled	0.86	0.95	0.91	276
Scratches	0.92	0.87	0.89	276
Accuracy			0.91	1656

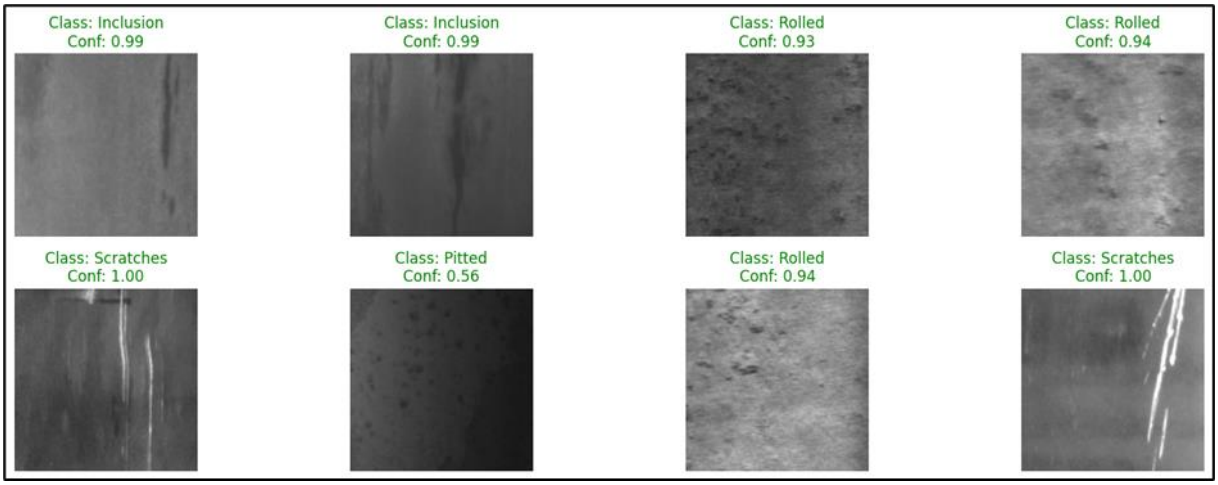


Fig 3: Sample prediction visualization for MAE-ViT model

These results indicate that while the model generalizes well across all classes, certain defect types are easier to learn, likely due to more distinct visual patterns.To complement the quantitative metrics, sample prediction visualizations are included in figure 3.

These samples showcase correctly predicted images along with the model's confidence scores, providing qualitative insight into the model's reliability and interpretability. Only high-confidence, correctly classified examples are shown, reinforcing that the model not only classifies accurately but does so with certainty. These visual confirmations further validate the model's effectiveness in real-world defect classification tasks.

Evaluating the Effectiveness of SimCLR and MAE Classifiers:

Figure 4 illustrates a comparative analysis of classifier performance between SimCLR and MAE using two key evaluation metrics: Accuracy and F1-Score. As shown in the bar chart, the SimCLR-based classifier significantly outperforms the MAE-based classifier across both metrics. Accuracy: SimCLR achieves perfect accuracy (1.00), indicating that all predictions made by the classifier were correct. In contrast, the MAE classifier achieves an accuracy of 0.93, reflecting a slightly lower performance. F1-Score: Similarly, the SimCLR classifier also scores 1.00 in terms of F1-Score, while the MAE classifier attains a slightly lower score of 0.93. This suggests that SimCLR maintains a better balance between precision and recall in classification tasks. The results underscore the superior performance of SimCLR over MAE in terms of both accuracy and robustness, making it a more reliable choice for classification tasks in the given context.

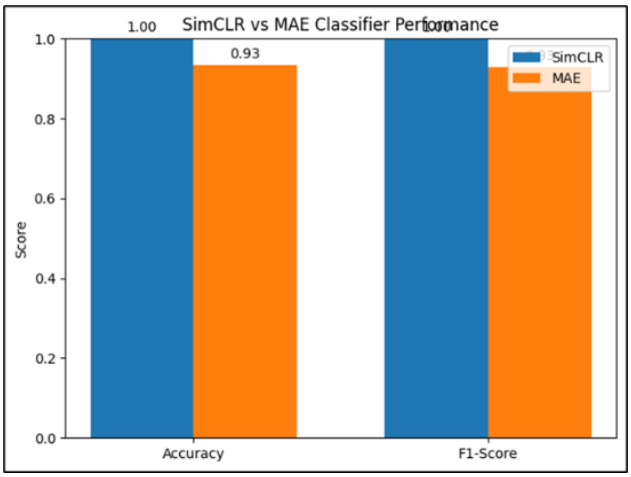


Fig 4 : Comparison of classifiers performance

***Per-Class Performance and Confusion Matrix Analysis*:**

To further investigate the classification capabilities of SimCLR and MAE, Figure 1.5 presents a per-class F1-Score comparison, and Figure 1.6 displays the corresponding confusion

matrices for both models.As shown in Figure 5, the SimCLR model achieves an F1-Score of **1.00** across all defect categories—**Crazing, Inclusion, Patches, Pitted, Rolled**, and **Scratches**—indicating perfect precision and recall. In contrast, the MAE model shows slightly lower performance in specific classes like **Crazing**: 0.91 and **Pitted**: 0.67. These discrepancies highlight MAE's struggles particularly with the *Pitted* class, which may be due to lower feature representation fidelity or inter-class similarity challenges.

Figure 6 provides deeper insight into model predictions. The **SimCLR confusion matrix** (left) shows ideal classification performance, with all diagonal elements having maximum values and zero off-diagonal errors, except for three misclassifications of *Scratches* as *Inclusion*.The **MAE confusion matrix** (right) reveals multiple misclassifications such as 2 instances of *Pitted* misclassified,5 instances of *Rolled*mislabeled as *Crazing*,1 *Scratches* instance misclassified as *Inclusion*.

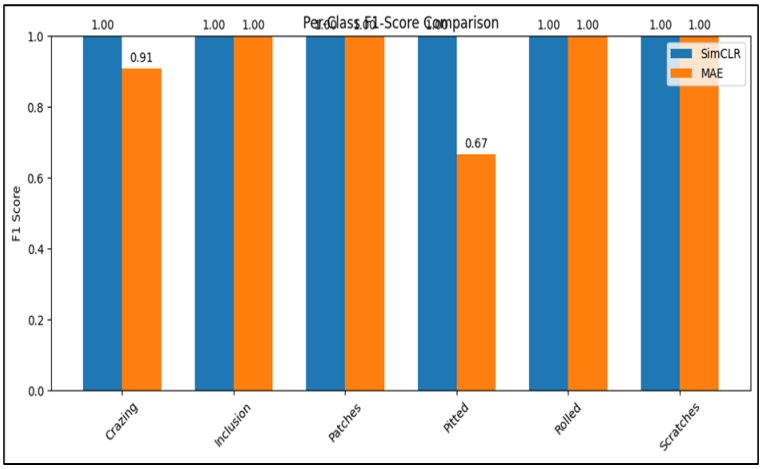


Fig 5 : Per class f1-score comparision

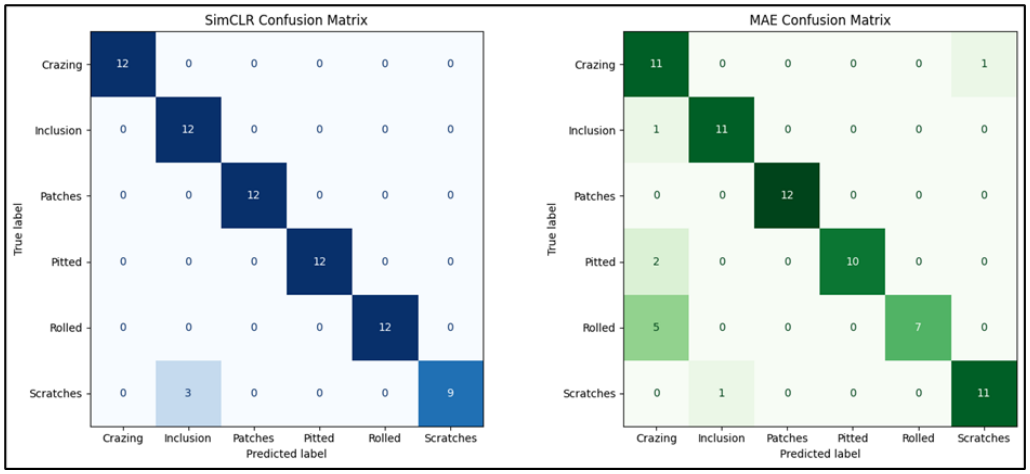


Fig 6 : Confusion matrix comparison of both models

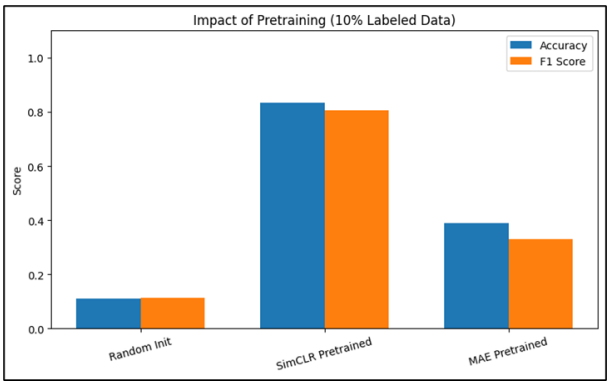
These results affirm the robustness and consistency of SimCLR in capturing class-discriminative features, while MAE, although effective in most cases, demonstrates limitations in separating visually similar defect types.This analysis underscores that SimCLR not only achieves higher overall metrics but also maintains strong per-class reliability, making it more suitable for defect classification tasks requiring high precision across multiple categories.

Models Performance with Limited Labeled Data:

To evaluate the robustness of different pretraining approaches under constrained supervision, we conducted experiments using only 10% of the labeled data. The results are depicted in figure 7, comparing model performance in terms of Accuracy and F1 Score across three scenarios: random initialization, SimCLR pretraining, and MAE pretraining.

SimCLR Pretrained models achieved the highest performance, with an accuracy of approximately 83% and an F1 Score around 81%, showcasing strong generalization even with limited supervision.MAE Pretrained models exhibited moderate performance, reaching around 39% accuracy and 32% F1 Score, indicating some learning benefits from masked autoencoding but substantially lower than SimCLR.Random Initialization resulted in the poorest performance, with accuracy and F1 scores close to 11%, essentially reflecting random guessing across multiple classes.

Fig 7: Impact of pretraining



These findings clearly demonstrate the advantage of contrastive pretraining (SimCLR) in low-label regimes, enabling effective feature learning even with minimal supervision. Conversely, MAE's performance under low-label conditions suggests its reliance on abundant data to effectively reconstruct and discriminate complex patterns.Figure 7 underscores the critical importance of pretraining strategy selection in real-world scenarios where labeled data is scarce. SimCLR emerges as the most effective approach, offering substantial improvements in both accuracy and class-wise performance when annotation resources are limited.

Overall Discussion:

This study successfully develops a self-supervised pretraining framework tailored specifically to the task of surface defect detection in industrial settings. By leveraging unlabeled data through self-supervised learning, we aimed to overcome the limitations imposed by the scarcity of annotated datasets, which are common in real-world manufacturing applications.

Two prominent self-supervised strategies were evaluated—contrastive learning (SimCLR) and masked autoencoding (MAE)—to determine their effectiveness in learning meaningful visual representations. Our results clearly indicate that SimCLR outperforms MAE across nearly all evaluation criteria, particularly in per-class F1-score, overall accuracy, and robustness under limited supervision. SimCLR benefits from learning instance-level discriminative features via the NT-Xent loss, which enables the model to distinguish fine-grained differences between defect types. In contrast, MAE relies on reconstructing masked patches and tends to prioritize global image structure over localized, class-specific features—making it less effective at capturing subtle variations critical to accurate defect classification.

A key contribution of this work is the evaluation of model performance under low-label regimes, where only 10% of the labeled data was used for fine-tuning. In these scenarios, SimCLR

retained strong performance (accuracy ~83%, F1 ~81%), demonstrating its ability to generalize well even with minimal supervision. This confirms the utility of contrastive learning in real-world settings where collecting labeled data is expensive or infeasible. On the other hand, MAE's performance deteriorated significantly (accuracy ~39%, F1 ~32%), highlighting a notable weakness of reconstruction-based pretraining in low-data environments. This may be attributed to MAE's dependence on abundant data to effectively capture the diverse appearance of defects and learn discriminative representations through reconstruction loss alone.

When compared to traditional supervised learning from random initialization, both SimCLR and MAE showed marked improvements in performance. However, SimCLR clearly emerged as the most effective approach, not only achieving superior accuracy but also doing so more efficiently by requiring less labeled data to reach high performance levels. This efficiency makes contrastive learning particularly advantageous for industrial applications, where scalability and minimal supervision are critical.

This work demonstrates that contrastive self-supervised pretraining via SimCLR offers substantial benefits over both masked autoencoders and traditional supervised training. SimCLR's resilience to label scarcity, superior per-class reliability, and strong generalization make it a compelling strategy for high-accuracy defect classification. While MAE holds potential in high-data scenarios or when used in hybrid systems, its limitations under constrained supervision must be carefully considered in practical deployments. Future directions may include exploring combined SimCLR-MAE pretraining, uncertainty-aware classification, and testing on broader real-world datasets to further validate and enhance defect detection frameworks.

Conclusion and Future Scope:

This work presents a comprehensive self-supervised learning framework for surface defect classification, comparing contrastive learning (SimCLR) and masked autoencoders (MAE). Our key contribution lies in demonstrating the superiority of SimCLR for feature extraction, particularly in low-label scenarios, achieving higher accuracy and robustness than MAE and traditional supervised methods. The results underscore the effectiveness of self-supervised pretraining in industrial defect detection tasks where labeled data is scarce.

In future work, hybrid models combining the strengths of both SimCLR and MAE could be explored to enhance representation learning. Additionally, incorporating uncertainty estimation and domain adaptation can further improve deployment reliability in diverse manufacturing environments. Expanding evaluation across real-world, multi-modal datasets will also help generalize this approach and unlock broader applications in automated visual inspection.

References:

1. Y. Chen, Y. Ding, F. Zhao, E. Zhang, Z. Wu, and L. Shao, “Surface defect detection methods for industrial products: A review,” *Applied Sciences*, vol. 11, no. 16, p. 7657, 2021.
2. Y. Min and Y. Li, “Self-Supervised Railway Surface Defect Detection with Defect Removal Variational Autoencoders,” *Energies (Basel)*, 2022, doi: 10.3390/en15103592.
3. X. Hu, J. Yang, F. Jiang, A. Hussain, K. Dashtipour, and M. Gogate, “Steel surface defect detection based on self-supervised contrastive representation learning with

matching metric," *Appl. Soft Comput.*, vol. 145, p. 110578, 2023, doi: 10.1016/j.asoc.2023.110578.

4. X. Huang, Y. Li, Y. Bao, and W. Zheng, "Adaptive Cross Transformer With Contrastive Learning for Surface Defect Detection," *IEEE Trans Instrum Meas*, vol. 73, pp. 1–17, 2024, doi: 10.1109/TIM.2024.3470998.
5. M. Aqeel, S. Sharifi, M. Cristani, and F. Setti, "Self-supervised Learning for Robust Surface Defect Detection," pp. 164–177, 2024, doi: 10.1007/978-3-031-66705-3_11.
6. Y. Lin *et al.*, "A Survey on RGB, 3D, and Multimodal Approaches for Unsupervised Industrial Image Anomaly Detection," *arXiv preprint arXiv:2410.21982*, 2024.
7. N. G. Shankar, "Defect pattern detection using a new rule-based approach," 2006.
8. J. Gui *et al.*, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Trans Pattern Anal Mach Intell*, 2024.
9. L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Process Mag*, vol. 39, no. 3, pp. 42–62, 2022.
10. Z. Huang *et al.*, "Contrastive Masked Autoencoders are Stronger Vision Learners," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, pp. 2506–2517, 2022, doi: 10.1109/TPAMI.2023.3336525.
11. S. Ji, S. Han, and J. Rhee, "Multi-View Masked Autoencoder for General Image Representation," *Applied Sciences*, 2023, doi: 10.20944/preprints202310.0524.v1.
12. S. K. Mishra *et al.*, "A simple, efficient and scalable contrastive masked autoencoder for learning visual representations," *ArXiv*, vol. abs/2210.16870, 2022, doi: 10.48550/arXiv.2210.16870.
13. W. Yang and L. Zhou, "CMGAE: Enhancing Graph Masked Autoencoders through the Use of Contrastive Learning," *2023 2nd International Conference on Machine Learning, Control, and Robotics (MLCR)*, pp. 42–47, 2023, doi: 10.1109/MLCR61158.2023.00018.
14. X. Yue *et al.*, "Understanding Masked Autoencoders From a Local Contrastive Perspective," *ArXiv*, vol. abs/2310.01994, 2023, doi: 10.48550/arXiv.2310.01994.
15. S. Yang, C. Lian, and Z. Zeng, "Masked Autoencoder for ECG Representation Learning," *2022 12th International Conference on Information Science and Technology (ICIST)*, pp. 95–98, 2022, doi: 10.1109/ICIST55546.2022.9926900.
16. R. Wagner, C. F. Lopez, and C. Stiller, "Self-supervised pseudo-colorizing of masked cells," *PLoS One*, vol. 18, 2023, doi: 10.1371/journal.pone.0290561.
17. Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-Supervised Pretraining via Multimodality Images With Transformer for Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023, doi: 10.1109/TGRS.2023.3271024.
18. H. Wu, B. Li, L. Tian, Z.-C. Sun, C. Dong, and W. Liao, "CoRe: Contrastive and Restorative Self-Supervised Learning for Surface Defect Inspection," *IEEE Trans Instrum Meas*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3291776.
19. F. Wang, F. Cheng, M. Zhang, and H. Zhang, "Self-supervised learning for textured surface anomaly detection and localization," vol. 12596, pp. 125961T-125961T–6, 2023, doi: 10.1117/12.2673155.

20. S. Yang, C. Lian, Z. Zeng, B. Xu, Y. Su, and C. Xue, "Masked self-supervised ECG representation learning via multiview information bottleneck," *Neural Comput. Appl.*, vol. 36, pp. 7625–7637, 2024, doi: 10.1007/s00521-024-09486-4.
21. D. Wolf *et al.*, "Abstract: Self-supervised Pre-training for Dealing with Small Datasets in Deep Learning for Medical Imaging - Evaluation of Contrastive and Masked Autoencoder Methods," p. 157, 2024, doi: 10.1007/978-3-658-44037-4_46.
22. X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-Based Masked Autoencoder With Contrastive Loss for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023, doi: 10.1109/TGRS.2023.3315678.
23. Y. Luo, Z. Chen, S. Zhou, K. Hu, and X. Gao, "Self-distillation Augmented Masked Autoencoders for Histopathological Image Understanding," *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1343–1349, 2022, doi: 10.1109/BIBM58861.2023.10385986.
24. H. Quan *et al.*, "Global Contrast Masked Autoencoders Are Powerful Pathological Representation Learners," *ArXiv*, vol. abs/2205.09048, 2022, doi: 10.48550/arXiv.2205.09048.
25. J. Liu, Y. Wu, M. Gong, Z. Liu, Q. Miao, and W. Ma, "Inter-Modal Masked Autoencoder for Self-Supervised Learning on Point Clouds," *IEEE Trans Multimedia*, vol. 26, pp. 3897–3908, 2024, doi: 10.1109/TMM.2023.3317998.