Automated Detection of Defects on Metal Surfaces using Deep Learning Techniques and Vision Transformers

Arwa Zakaria Khaled Alorbany*, Toqa Alaa Awad*, Mariam Mahmoud Mohamed Diab*,Mostafa Ahmed Atef Kotb*, Walid Gomaa*,†
*Egypt-Japan University of Science and Technology, Alexandria, Egypt.
†Faculty of Engineering, Alexandria University, Alexandria, Egypt.
{arwa.zakaria, toqa.alaa, mariam.diab, mostafa.atef, walid.gomaa}@ejust.edu.eg

***Abstract*—Metal manufacturing and reshaping industries often yield products with inevitable defects, which can lead to serious operational issues. Manual defect detection is a time-consuming and unreliable process, necessitating the development of an automated monitoring tool. This research presents a novel approach to address this problem, employing deep learning techniques and vision transformers. The proposed model aims to achieve two primary objectives: defect classification and defect localization. Defect classification enables the identification of the underlying problem, while defect localization facilitates precise treatment of each identified defect.**

To accomplish this, our model leverages a pre-trained Vision Transformer (ViT) on the "Imagenet" dataset, exploiting the attention mechanism for effective feature extraction. Subsequently, the model bifurcates into two distinct pathways. The first pathway employs a fully connected layer followed by a "Softmax" output layer for defect classification. The second pathway utilizes a fully connected layer followed by an output layer that predicts the bounding box coordinates for each defect. The developed model focuses on achieving high classification accuracy, as well as minimizing the Mean Square Error (MSE) and Mean Average Error (MAE) for defect localization. Through extensive experimentation and evaluation, the model demonstrates promising results in automating defect detection on metal surfaces. The proposed approach holds significant potential for improving operational efficiency and reducing costly errors in metal manufacturing industries.

***Index Terms*—CNN, ViT, InceptionV3, GC10-DET, NEU-DET**

Contents

List of Figures

I. Introduction

The manufacturing and reshaping of metal surfaces are vital processes in various industries such as automotive, aerospace, and construction. However, the outputs of these processes often contain defects that can have detrimental effects on the performance and reliability of the final products. These defects can include cracks, dents, scratches, and other surface irregularities, which not only compromise the structural integrity of the metal but also pose significant challenges in terms of quality control and product usability. Detecting and addressing these defects is crucial to ensure the production of high-quality metal products and prevent costly failures.

Traditionally, the detection of defects on metal surfaces has relied heavily on manual inspection, where human experts visually examine the surfaces for any abnormalities. This approach, while valuable, is highly time-consuming, labor-intensive, and subjective. It is prone to errors and lacks consistency, especially when dealing with large-scale manufacturing processes that produce high volumes of metal components. Furthermore, manual inspection may not be capable of detecting subtle defects that are not easily visible to the human eye. As a result, there is a pressing need for automated defect detection systems that can accurately and efficiently identify and classify defects on metal surfaces.

Over the years, researchers have made significant progress in developing automated defect detection techniques for metal surfaces. Traditional computer vision methods, such as edge detection, thresholding, and image segmentation, have been widely explored. These methods often rely on handcrafted features and rule-based algorithms to identify defects based on predefined criteria. While these approaches have achieved some success in detecting certain types of defects, they are limited in their ability to handle complex and varied defect patterns. They heavily rely on the expertise of domain-specific engineers and lack the ability to adapt to new defect types or variations.

In recent years, with the advent of deep learning techniques and the availability of large-scale annotated datasets, there has been a shift towards employing neural networks for automated defect detection. Convolutional Neural Networks (CNNs) have shown remarkable performance in various computer vision tasks, including image classification, object detection, and semantic segmentation. CNNs have the ability to automatically learn discriminative features from raw input data, making them well-suited for defect detection on metal surfaces. Several studies have successfully applied CNNs to detect defects on metal surfaces, achieving high accuracy and demonstrating the potential of deep learning in this domain.

While CNNs have shown promise, they have limitations that hinder their effectiveness in defect detection. CNNs typically rely on local receptive fields and hierarchical feature extraction, which may not adequately capture long-range dependencies and global context in images. This limitation becomes especially critical when dealing with complex defect patterns that span a significant portion of the metal surface. Additionally, CNNs require large amounts of labeled training data, which can be challenging and time-consuming to acquire for specific defect types or rare occurrences. To address these limitations, this research paper proposes the use of ViTs for automated detection of defects on metal surfaces. Vision Transformers, originally introduced for natural image classification, have gained attention for their ability to capture global context and long-range dependencies through the use of self-attention mechanisms. This makes them well-suited for capturing intricate defect patterns on metal surfaces that may extend across the image.

The main objective of this research is twofold: defect classification and defect localization. Defect classification aims to accurately identify the type and nature of each defect, enabling prompt problem-solving measures to be taken. Defect localization aims to precisely delineate the boundaries of each defect, allowing for targeted treatment and repair.

To achieve these objectives, the proposed approach leverages the power of pre-trained ViTs on large-scale image datasets, such as Imagenet. By utilizing transfer learning, the model can benefit from the learned representations of ViTs, which capture rich visual features. This pre-training stage enables the model to effectively extract meaningful defect-related features from raw metal surface images.
In conclusion, the detection of defects on metal surfaces is a critical task in manufacturing and reshaping industries. Manual inspection methods are time-consuming, subjective, and limited in their ability to detect complex defects. This research paper proposes the use of Vision Transformers and deep learning techniques to automate defect detection on metal surfaces.

By using the power of ViTs and transfer learning, the model aims to achieve accurate defect classification and precise defect localization. The research contributes to the advancement of automated defect detection systems, enhancing product quality, and reducing costly errors in metal manufacturing industries.

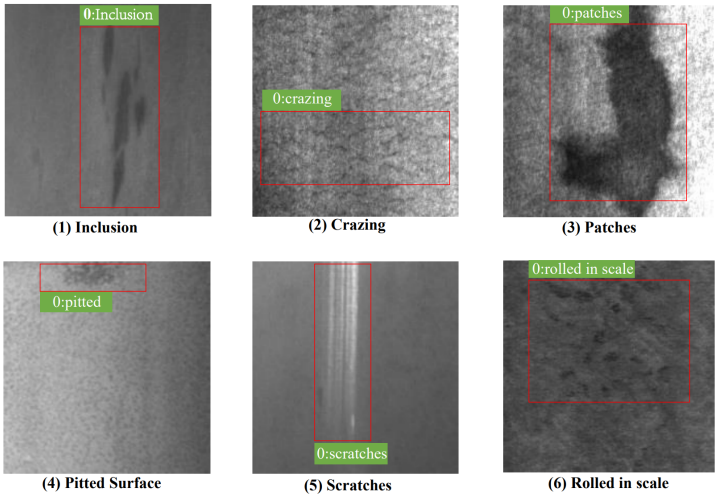


Fig. 1: NEU-DET Six Classes

II. Related Work

The detection of defects on metal surfaces using deep learning techniques and vision transformers is an active area of research, and several studies have been conducted in this domain. In this section, we present a review of the existing work, highlighting the key findings and approaches used, while identifying the research gap that this paper aims to address. The work done in this research is based on two metal surface datasets: NEU-DET [1] and GC10-DET[2]. NEU-DET is a dataset that was collected by Kechen Song and Yunhui Yan at the Northeastern University. It contains six kinds, shown in Figure 1, of typical surface defects of the hot-rolled steel strip, which are rolled in scale, patches, crazing, pitted surface, inclusion, and scratches. The database includes 1,800 grayscale images; with an average of 300 samples per each of six different kinds of typical surface defects.

GC10-DET is a surface defect dataset collected in a real industry, and first-time mentioned by Xiaoming Lv in his paper "Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network" [2]. It contains ten types, shown in figure 2, of surface defects, which are punching hole, welding line, crescent gap, water spot, oil spot, silk spot, inclusion, rolled pit, crease, and waist folding. The collected defects are on the surface of the steel sheet. The dataset includes 3570 gray-scale images.

One prominent approach in defect detection is the utilization of Convolutional Neural Networks (CNNs). Wang et al. in 2021[3] proposed a CNN-based method for the automatic detection of surface defects on steel plates. Their model achieved high accuracy by learning discriminative features from image patches. Similarly, Li et al. in 2022 [4] employed a CNN architecture to classify defects on aluminum alloy surfaces, achieving satisfactory results. However, these CNN-based methods primarily focused on defect classification and lacked precise defect localization, which is essential for effective treatment and repair. To address the limitation of defect localization, researchers have explored object detection algorithms. Tang et al. in 2019 [5] proposed a defect detection approach using Faster R-CNN, a popular object detection framework. They achieved accurate defect localization by predicting bounding boxes for each defect instance on metal surfaces. However, their method relied on handcrafted features, which may limit its ability to capture complex defect patterns. Furthermore, the study did not explore the use of vision transformers, which have shown promise in capturing long-range dependencies and global context.

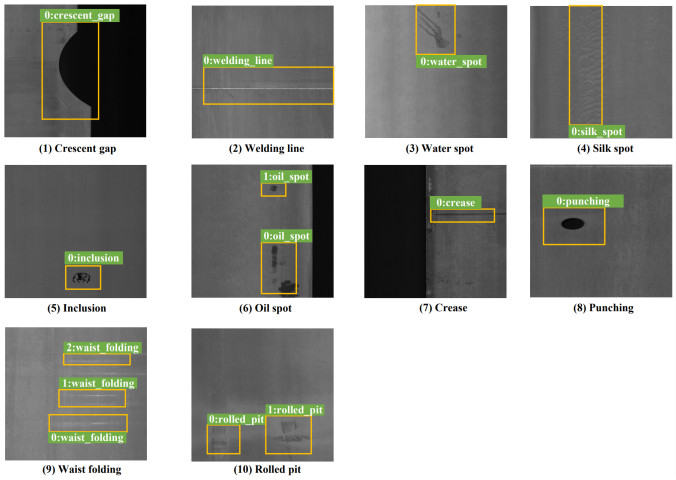


Fig. 2: GC10-DET Ten Classes

In recent years, vision transformers have emerged as powerful models for image analysis tasks. Dosovitskiy et al. in 2020, [6] introduced the Vision Transformer (ViT) model, demonstrating its effectiveness in large-scale image classification. The self-attention mechanism of ViTs allows for the modeling of global context, enabling the capture of intricate defect patterns on metal surfaces. Although ViTs have been extensively studied in the context of image classification, their application to defect detection on metal surfaces is relatively unexplored.

To bridge this research gap, this paper proposes the use of pre-trained ViTs for automated defect detection on metal surfaces. By leveraging the attention mechanism and transfer learning, the model aims to capture global context and learn meaningful defect-related features. This approach combines the advantages of ViTs in capturing long-range dependencies with the ability to detect and classify defects accurately. Furthermore, the proposed model incorporates defect localization through the prediction of bounding box coordinates for each defect, facilitating precise treatment and repair strategies.

While some previous research has investigated defect detection on metal surfaces using CNNs and object detection algorithms, there is a lack of exploration regarding the application of vision transformers in this domain. The proposed research aims to fill this gap by investigating the effectiveness of ViTs for defect detection and localization on metal surfaces. By leveraging the strengths of pre-trained ViTs and transfer learning, the model seeks to achieve accurate defect classification and precise defect localization, contributing to the advancement of automated defect detection systems in metal manufacturing industries.

In summary, the existing research in the field of defect detection on metal surfaces has primarily focused on CNN-based methods and object detection algorithms. However, there is a lack of exploration regarding the utilization of vision transformers for this task. This paper aims to address this research gap by proposing a novel approach that combines the power of vision transformers with deep learning techniques for automated defect detection and localization on metal surfaces.

III. Methodology

This section presents the methodology employed in the research, including the data collection process, the development of different architectures, and the evaluation metrics used. All our work can be found on this repository on GitHub: https://shorturl.at/vy169.

A. Data Collection and Renaming

The data used in this study was collected from two datasets, namely GC10-DET and NEU-DET. From GC10-DET, the following classes were selected: punching hole, welding line, crescent gap, water spot, oil spot, silk spot, inclusion, crease, and waist folding. The rolled pit class was excluded due to its noisy and insufficient content. From NEU-DET, the classes rolled in scale, inclusion, and scratches were chosen. The inclusion class contains data from both datasets.

This selection was made to focus on metal surface defects within the scope of the research and to achieve maximum accuracy while minimizing confusion by excluding classes with insufficient data. These decisions were made after applying trial models and debugging techniques to ensure that including these classes would not be beneficial and may lead to more harm than good. Afterwards, the dataset passed on a renaming procedure, where all the images were named in the following manner “Defect Name Image Number”. The XML files that contain the labels for each image were also named in the same manner to ease access and collaboration.

Furthermore, a smaller dataset was created by the research team to simulate diverse environments and metal surfaces. This dataset was intended to be used for fine-tuning the model to improve its generalization capabilities. Work on this dataset is currently ongoing.

B. Classification Architecture

The initial step in developing the current model involved creating an architecture with the sole objective of classifying the entire image into one of the 11 defect classes present in the modified dataset.

Two classification architectures were implemented. The first one consisted of a basic CNN architecture with several convolution layers followed by a ”flatten” layer and fully connected dense layers. The final dense layer employed a SoftMax activation function with 11 nodes for classification. The model architecture can be found in Figure 3. This straightforward model was designed solely for classification purposes.

To enhance model complexity and improve accuracy, another classification architecture was implemented. In this approach, the initial convolution layers were replaced with the pre-trained InceptionV3 model, which served as a feature extractor. As the InceptionV3 model was trained on different data, the last 10 layers were fine-tuned to achieve better results.

C. Single Defect Detection

To incorporate defect localization into the classification models, the next step involved assuming that each image in the dataset contained only one defect and attempting to localize it. The model inputs consisted of five numbers: the defect class number and the bounding box dimensions (xmin, xmax, ymin, ymax).

The model architecture, as shown in Figure 4, included a set of common convolution layers for feature extraction. The extracted features were then forwarded through two separate paths: one for classification, which ended with a dense layer using a SoftMax activation function with 11 nodes, and the other for localization, which ended with a dense layer without an activation function and had four nodes for predicting the dimensions of the bounding box.

To improve results, the convolution layers were removed, and the InceptionV3 Neural Network was utilized instead. The same architecture was retained for the two paths.

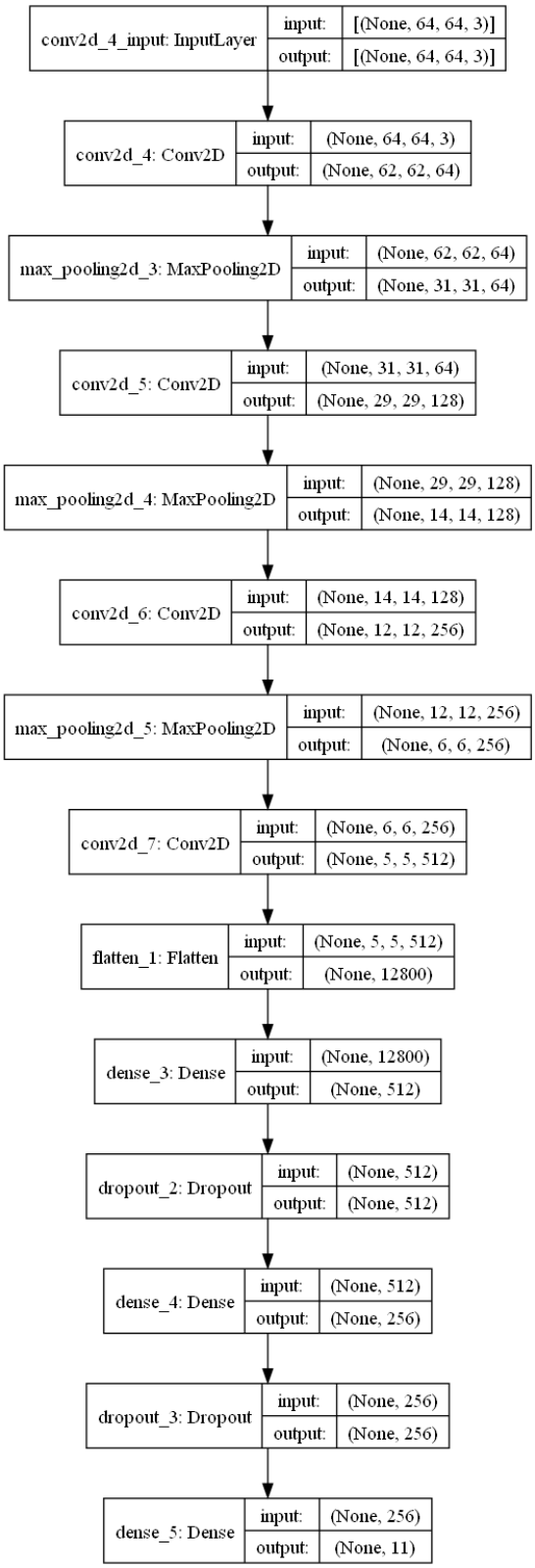


Fig. 3: Classification Model Architecture

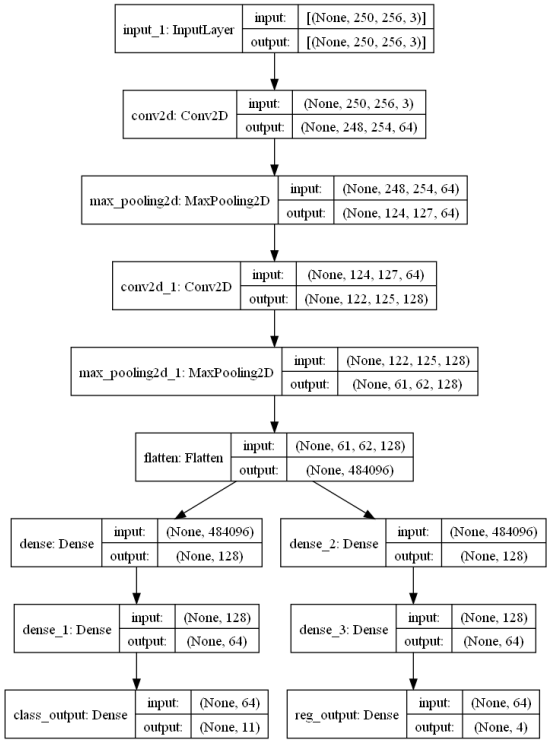


Fig. 4: Single Defect Model Architecture

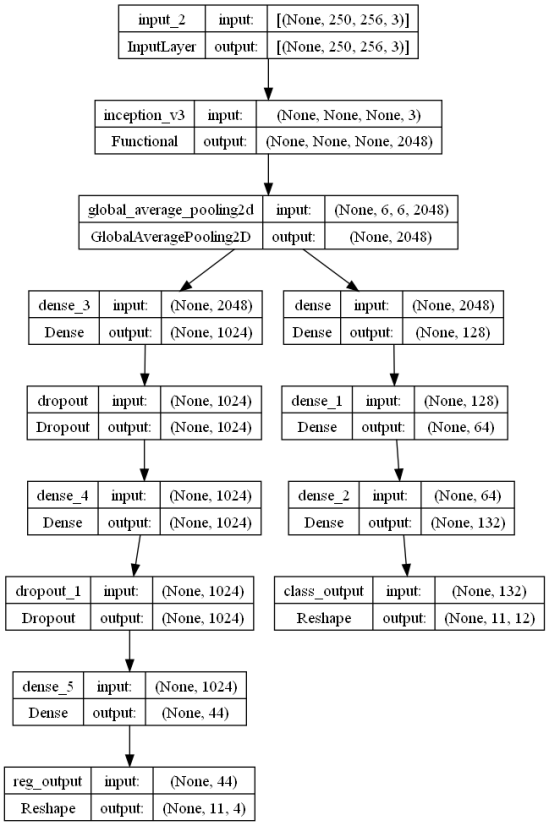


Fig. 5: Multi-Defect Model Architecture

D. Multi Defect Detection

To fulfill the main objective of the model, an architecture for multi-defect detection was introduced. The data required a different format to handle images that contained multiple defects with various interconnecting areas. The proposed approach involved padding the input data with unrelated data to reach the maximum number of defects in one image, similar to zero-padding in images. This was done to make the data usable by the model.

For classification data, the format was achieved by padding with a twelfth class, indicating the absence of a defect in the selected area. For localization data, the format was achieved by padding the four dimensions of the bounding box with zeros, indicating the absence of any defect in that area.

The model architecture inherited from the previous edition was modified only by changing the shapes of the output layers. The softmax layer was adjusted to have 11x12 nodes, representing the maximum number of defects per image in the proposed dataset. The regression layer had 11x4 nodes to predict the dimensions of the boundary box for each defect.

E. Vision Transformers Defect Detection

The final step involved incorporating vision transformers into the architecture. The Vision Transformer was placed at the beginning of the model to replace the InceptionV3 Neural Network. The pre-trained ViT on ImageNet served as an attention-conscious feature extractor, extracting the most relevant features for the two paths mentioned earlier. The transformer facilitated better results through self-attention mechanisms. Although the vision transformer was not directly integrated into the model itself, all the image data was passed through it to extract features. These features were then outputted as tensors, which served as the new input for the model architecture. The architecture began with a flatten layer followed by the two aforementioned paths.

F. Evaluation Metrics

The evaluation metrics used in this study were standard in the field. For the classification and single defect detection models, the loss function employed was Categorical Cross Entropy. In the multi-defect

detection models, a modified version of the function was used to calculate accuracy more precisely. The metrics reported for classification included loss and accuracy. For the regression parts used in localization, the loss function was Mean Squared Error (MSE), and Mean Absolute Error (MAE) was also reported.

By following this methodology, a systematic approach was employed in terms of data collection, model development, and evaluation. The different architectures and techniques utilized aimed to achieve accurate classification and precise localization of defects on metal surfaces.

IV. Results

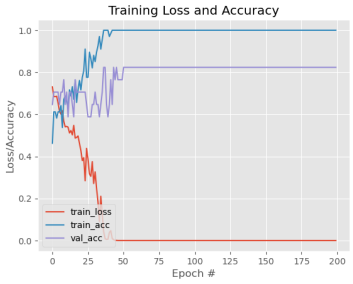
In this section, the results of the mentioned architectures will be presented.

A. Two Classes Classification

This was one of the first trial models trying to explore the dataset. It achieved relatively good accuracy with different couples of classes. The resulted graphs can be shown in the following figures.

(a) Silk Spot-Water Spot Classification Results

(b) Welding Line-Crescent Gap Classification Results

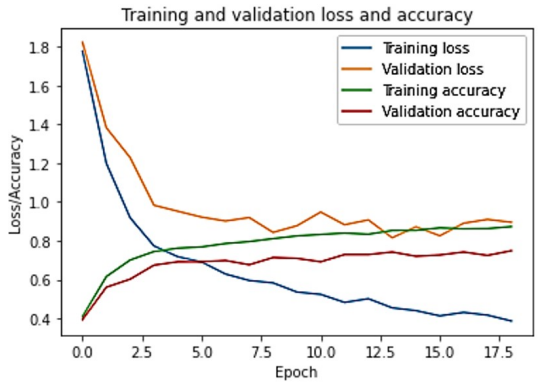


(c) Punching Hole-Waist Folding Classification Results

The results were pretty useful as a start in exploring the datasets and benefited as to collect an idea about what the development process will be. The accuracies were high and overfitting was acceptable.

B. GC10-DET Classification

The second step that was taken was to classify the whole dataset of GC10-DET. This was before excluding the unwanted class The noise was recognizable due to the confusion from the unwanted class. The accuracy wasn't high enough and the gap between the two accuracies, training and validation, was wide.



(d) GC10-DET Classification using InceptionV3

(e) GC10-DET Classification using Custom CNN

C. Collected Dataset Classification

The next step was customizing the new dataset with 11 classes from both datasets. The accuracy became a lot better than the previous noisy dataset. Although, the gap between the accuracy and the validation accuracy still exists, which indicates a deeper problem.

Fig. 6: Collected Dataset Classification Results

D. Single Defect Detection

The following results are after adding the localization feature for the single defect detection.The accuracies can converge to better results with more epochs, but the gap between the validation and training is getting wider in both the accuracy and the MAE.

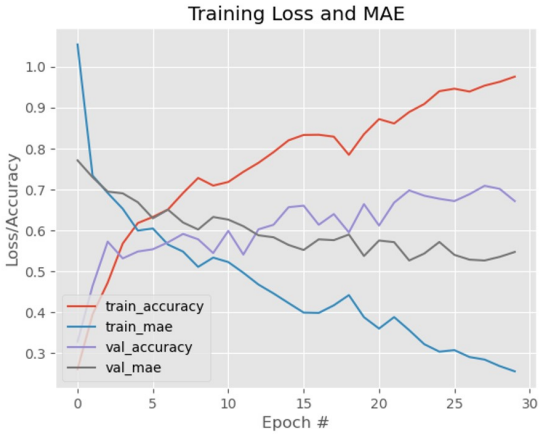


Fig. 7: Single Defect Detection Results using InceptionV3

E. Multi Defect Detection

The following results are the pre-final step, where the multi-defect detection model was implemented.We can find that the accuracies and the MSE are very good and acceptable which achieves our target.

Fig. 8: Multi-Defect Detection using InceptionV3

F. ViT Multi-Defect Detection

The last step was to convert all the images into features using the pre-trained Vision Transformer on ImageNet and passing them as inputs for the model.The noise and sudden divergence are way less, and the accuracy graph is stable. The overfitting nearly doesn't exist.

Fig. 9: ViT Multi-Defect Detection Results

V. Discussion

In this section, we will analyze the previous results and discuss the development process leading to the final outcomes.

A. Two Classes Classification

The results obtained from the classification of two classes together indicate that the dataset contains sufficient data for each of these two classes. The occurrence of overfitting was minimal and within an acceptable range. Our objective was to validate

the dataset, and we successfully achieved this goal using these simple architectures.

B. GC10-DET Classification

The results of the models for GC10-DET classification were not as expected. In the architecture utilizing the custom CNN, the training accuracy was relatively low, and the occurrence of overfitting was remarkably high. Even when we enhanced the model by incorporating the InceptionV3 Neural Network, the overall improvement was negligible. These outcomes indicated that there was an issue with the dataset itself.

After conducting an analysis of the dataset's images and employing confusion matrices to identify the class with the most errors, we discovered that the rolled pit class was the cause of these results. Two possible approaches were considered: augmenting the data through techniques such as rotation, shifting, and applying effects, or eliminating the class entirely. We decided to remove the class and introduced two additional classes from the NEU-DET dataset.

The revised dataset comprised 11 classes along with their corresponding annotations, and the inclusion class represented the combination of classes from both datasets. This inclusion of data increased the sample size for the inclusion defect, resulting in improved detection accuracy.

C. Collected Dataset Classification

Upon collecting the dataset, we evaluated the classification model on this new dataset to assess the results. The performance was significantly better than the previous run, as the loss value was significantly lower, and the training accuracy demonstrated notable improvement. However, there were still indications of overfitting, necessitating further debugging efforts. After experimenting with different architectures and pre-trained neural networks, we reached the conclusion that the issue lies with the fundamental assumption itself.

The notion that each entire image represents a single class, despite potentially containing features from different classes, confuses the model. It identifies repeated features but encounters differing labels, which leads to accepting the displayed results and progressing to the next step. It is evident that further advancements cannot be achieved using the same flawed approach.

D. Single Defect Detection

The results from the single defect detection approach exhibited some noise and overfitting. However, this can be attributed to the model architecture, which is designed to process the whole image features once for classification and once for regression. The model does not establish a correlation between the outputs of the two paths, which is crucial for single defect detection.

This approach closely resembles the previously flawed methodology, where the classification path remains largely unchanged. It continues to process an image containing features from different classes and assigns a single label that varies each time. Despite these limitations, the results prompt us to accept the current output and proceed to the next step.

E. Multi Defect Detection

This stage represents a significant advancement, with most of the issues addressed. The model is now capable of treating each identified feature differently and classifying them individually. The classification metrics of the model exhibit strong performance, with limited instances of overfitting and relatively high accuracies.

The regression metrics of the model also demonstrate satisfactory results, as the mean squared error (MSE) and mean absolute error (MAE) values for both training and validation sets are low. This indicates accurate measurements and predictions of defect locations. The only concern is a slight increase in validation loss during the classification process, suggesting that the model lacks complete stability and may require further refinement.

F. ViT Multi-Defect Detection

The final results obtained from the vision transformer exhibit greater stability and lack sudden divergences. The MSE and MAE values for both training and validation sets do not indicate any signs of overfitting, indicating the successful operation of self-attention mechanisms and the attainment of the desired objectives. The loss graphs show improved stability and manageability.

In summary, this study explored the problem of metal defect detection on surfaces using deep learning techniques. The analysis of the results revealed the importance of dataset quality and model

architecture in achieving accurate classification and detection. By addressing issues such as dataset refinement, incorporating multiple defect classes, and leveraging advanced architectures like vision transformers, significant improvements were observed in classification and localization performance. However, challenges related to overfitting and model stability were encountered, indicating areas for further research and refinement.

VI. Conclusion

Automated defect detection on metal surfaces is vital for industries like automotive, aerospace, and construction. Manual inspection methods are slow and subjective, calling for automated systems. This study proposes using Vision Transformers (ViTs) to overcome limitations of traditional methods.
ViTs, with their attention mechanisms, can capture complex defect patterns effectively. The research focuses on defect classification and localization, using pre-trained ViTs and transfer learning. By automating defect detection, the approach aims to improve product quality and reduce errors in metal manufacturing. The study addresses a research gap in applying ViTs to metal surface defect detection, contributing to the field. The methodology involves data collection, architecture development, and evaluation using metrics like accuracy and loss. Promising results demonstrate accurate defect classification and precise defect localization. This research advances automated defect detection, benefiting multiple industries.

In contribution to the field of automated defect detection on metal surfaces by employing deep learning techniques and vision transformers, our methodology offers a promising approach for addressing the challenges posed by metal defects in manufacturing and reshaping industries. However, there is still room for improvement, particularly in addressing overfitting and model stability issues. Future research can focus on refining the methodology, exploring additional architectures, and expanding the application of deep learning techniques in industrial quality control.

Ultimately, this research paves the way for more effective defect detection, ensuring the production of high-quality metal products and reducing operational challenges in various industries.

VII. References

References

[1] K. Song and Y. Yan, *NEU surface defect database," Northeastern University*, Conference Name, Northeastern University, 2021. [Online]. Available: http://faculty.neu.edu.cn/songkc/en/zdylm/263265/list/index.htm [Accessed 1 December 2022]

[2] X. Lv, F. Duan, J.-j. Jiang, X. Fu,and L. Gan, *Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network*, Journal Name, vol. 20, no. 6, pp. 1560, 2020.

[3] S. Wang, X. Xia, L. Ye, and B. Yang, *Automatic Detection and Classification of Steel Surface Defect Using Deep Convolutional Neural Networks*, Metals,vol. 11, no. 3, p. 388, 2021.

[4] Z. Li, B. Li, H. Ni, F. Ren, S. Lv and X. Kang, *An Effective Surface Defect Classification Method Based on RepVGG with CBAM Attention Mechanism (RepVGG-CBAM) for Aluminum Profiles*, Metals,vol. 12, no. 11, p. 1809, 2022.

[5] B. Wei, K. Hao, X.-S. Tang and L. Ren, *Fabric Defect Detection Based on Faster RCNN: Proceedings of the Artificial Intelligence on Fashion and Textiles*, in Advances in Intelligent Systems and Computing, Hong Kong, Springer, 2019, pp. 45-51.

[6] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner and X. Zhai, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*,in ICLR, Vienna, 2021.

[7] Y. Bao, K. Song, J. Liu, Y. Wang, Y. Yan, H. Yu, and X. Li, *Triplet-Graph Reasoning Network for Few-Shot Metal Generic Surface Defect Segmentation*, IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021.

[8] K. Song, and Y. Yan, *A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects*, Applied Surface Science, vol. 285, no. B, pp. 858-864, 2013.

[9] Y. He, K. Song, Q. Meng ,and Y. Yan, *An End-to-End Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Fea-*

tures, IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 4, pp. 1493-1504, 2020.