# VISUAL DEFECT CLASSIFICATION IN STEEL SURFACES USING TRANSFORMER ARCHITECTURE

**K.Andal[1], S.Sathiya[2], P.Sivaraj[3]**

**Article History: Received:** 13.02.2023 **Revised:** 28.03.2023 **Accepted:** 13.05.2023

## Abstract

Defect classification in a typical surface using automated visual inspection (AVI) tool for planar materials is an important task often implemented after flaw detection, and it serves as a necessary prerequisite for achieving the on-line quality inspection of finished goods. In the industrial environment of manufacturing flat steels, this detection and classification of defects is incredibly difficult due to different appearances, unclear intraclass and interclass variations.This study shall propose a classification approach using transformer architecture for classification of defects present in the steel surfaces. A sequence of vectors is created by dividing animage into fixed-size patches, linearly embedding each one, adding position embeddings, and then feeding the assembled vectors to a conventional Transformer encoder. The traditional method of performing classification involves including an extra learnable "classification token" in the sequence.By giving the network attention, it is possible to learn the connections between the image patches. This can be accomplished either in combination with a Convolutional Neural Network(CNN) model or by substituting a few of its elements. These network architectures can be used for image classification tasks.

**Keywords:** automated visual inspection, transformer architecture, conventional Transformer encoder, network attention, Convolutional Neural Network

[1]Research Scholar, Department of Computer Science and Engineering, Annamalai University
[2]Assistant Professor, Department of Computer Science and Engineering, Annamalai University
[3]Associate Professor, Department of Manufacturing Engineering, Annamalai University

Email: [1]sanandal86@gmail.com, [2]Sathiya.sep05@gmail.com, [3]cemajorsiva@gmail.com

**DOI: 10.31838/ecb/2023.12.s3.314**

## 1. Introduction

Many businesses are interested in using machine vision-based surface inspection technology to automate inspection processes and vastly enhance overall product quality [1]. The rolled steel strip business is a classic example of a sector implementing these sophisticated inspection techniques. To produce a product with ever-fewer surface flaws, production lines must be visually inspected in real-time. Even while manual inspection is very accurate when dealing with a small number of samples, it is slow and prone to errors brought on by fatigue in today's high-speed setups. Due to the numerous additional costs that emerge, this strategy is not only overly expensive but also inappropriate. One of the most researched topics in quality control is automated machine vision inspection, which is quick and reliable and frequently produces results that are adequate.The tools needed to meet the demands for speed, resolution, and categorization of current production lines are provided by improvements in camera technology, acquisition hardware, and machine learning algorithms. Steel defect categorization and identification remain a challenging challenge, even with the best tools and most cutting-edge algorithms [3]. It is challenging to make further advancements based on expert knowledge contained in geometrical and shape-based properties. Particularly in contemporary dynamic processes where production swiftly switches from one product to another, an effective inspection algorithm should learn adaptively in response to the changing data distribution.

A deep learning model that uses attentional mechanisms to differentially weigh the importance of each component of the input sequence of data is referred to as a transformer in machine learning [2]. Transformers in machine learning are made up of numerous layers of self-attention. In comparison to convolutional neural networks (CNN), Vision Transformer provides impressive results while using significantly less computational resources for pre-training. Transformer architecture exhibit a generally less inductive bias than convolutional neural networks (CNN), which increases reliance on model regularization or data augmentation while training on smaller datasets.When compared to convolutional neural networks (CNN), a transformer architecture performs well while requiring significantly fewer computational resources for pre-training. While training on lesser datasets, Vision Transformer (ViT) [4] shows a generally reduced inductive bias compared to convolutional neural networks (CNN), which increases reliance on model regularization or data augmentation.

Transformer divides the input images into visual tokens while CNN uses pixel arrays. By dividing animage into fixed-size patches, accurately embedding each one, and including positional embedding as an input to the transformer encoder, the visual transformer can change an image. Transformer models also surpass CNNs in terms of accuracy and computing efficiency by a factor of approximately four. Transformer's self-attention layer makes it feasible to embed data globally throughout the entire image [5]. In order to recreate the structure of the image, the model also learns from training data to encode the relative locations of the image patches. The objective of this research is to develop an efficient and reliable method for classifying defects present in steel surfaces images using vision transformer architecture.

**Literature Survey**

Although there are only few literature available on steel defect detection, the issue can be seen in the perspective of computer vision's extensive research into defect detection in textured materials. The

filter-bank paradigm is used in traditional approaches to feature extraction. Each image is convolved using a collection of two-dimensional filters, the structure and support of which are derived from prior task knowledge, and the filter responses are then fed to conventional dense connected classification layers.Deep convolutional neural networks (DCNNs) have shown remarkable performance as a defect classification tool in recent years and have been employed in various manufacturing industries. For a DCNN to operate well, there must be an enough amount of training data collected. Yet, because some flaws only sometimes occur in the metal manufacturing industries, it is challenging to collect enough data. The generalization performance of the DCNN-based classification method is decreased as a result of the imbalanced data issue.In order to address this issue, the study in [6] suggests a new convolutional variationalautoencoder (CVAE) and deep CNN-based defect classification algorithm. Enough faulty data is produced using the CVAE-based data production technology to train the classification model. It is suggested that a conditional CVAE (CCVAE) be used to produce images for each type of defect in a single CVAE model. Also, they suggest a classifier with strong generalization performance that uses data from the CCVAE and is based on a DCNN. The work proposed in [7] has analyzed the idea of classifying defects using residual neural networks. The ResNet50 neural network-based classifier was employed as a base feature extractor. Based on test data, the model is capable of classifying images of flat surfaces with damage into one of three categories with an overall accuracy of 96.91%.

For quick and precise steel surface defect classification, [8] proposed a small-but-powerful convolutional neural network (CNN) model that focuses on learning lowlevel features and integrates several receptive fields. As the foundation of their architecture, they used the pre-trained SqueezeNet. On a diversity-enhanced testing dataset of steel surface defects that incorporates severe non-uniform illumination, image capturing noise, and motion blur, high accuracy detection can be achieved with only a limited number of defect-specific training samples.Transformers do not generalize well when trained on insufficient quantities of data because they lack several of the inductive biases present in CNNs, such as translation equivariance and localization. If the models are trained on a larger dataset, the situation does, however, change. Large-scale training was proven to be superior against inductive bias. When pre-trained at a large enough scale and applied to applications requiring fewer datapoints, Vision Transformer (ViT) shall achieves great results.

Transformers were first suggested in [9] for machine translation, and they have since emerged as the most advanced approach for many NLP applications. Transformer-based models are frequently pre-trained on massive corpora. For example, BERT [10] employs a de-noising self-supervised pre-training task, whereas the GPT line of work uses language modelling as its pre-training task [11].

Deep neural network-based methods are frequently employed for the examination of steel surfaces. A max-pooling convolutional neural network method for supervised steel defect classification is presented by the authors of one study [13]. An error rate of 7% is found on a classification task using seven defects gathered from a genuine production line. One strategy for diagnosing steel faults using a deep structured neural network, such as a convolutional neural network with class activation maps, is suggested by the authors of another study [14].

## Transformer Architecture

The input image is restructured in to smaller $N = \frac{HW}{P^2}$ number of patches in which each patch resolution is $P \times P$

pixels. The input image is applied to the architecture after completing the following operations;

- Image patches were flattened to a vector $X_p^n$ of length $P^2 \times C$ where $n = 1, \dots N$.
- Using a trainable linear projection function $E$, the flattened patches are mapped to $D$ dimensions to generate a sequence of embedded image patches.
- The sequence of embedded image patches are prepended with a learnable class embedding $X_{class}$; where the value of $X_{class}$ represent the output of the classification, $y$.
- At last the patch embeddings are augmented with 1D positional embeddings$E_{pos}$ which adds positional information to the input. This is also learned during training process.

The embedding vectors are represented mathematically as follows;

$$Z_0 = \left[X_{class};\ X_p^1 E; \dots;\ X_p^N E\right] +\ E_{pos}$$

After completing the above mentioned preprocessing steps; the sequence of embedding vectors $Z_0$ is fed as input of the transformer encoder. The encoder comprised of stacked identical layers and the classification layer implements a non-linear activation, Gaussian Error Linear Unit (GELU). The encoder component of the original Transformer architecture is employed by the transformer architecture that is used for image classification. The encoder receives a sequence of embedded image patches as input, together with positional information and a learnable class embedding prepended to the sequence. The learnable class embedding value is sent to a classification layer coupled to the encoder's output, which uses it to produce a classification output depending on its state.In CNNs, each layer of the model utilizes the spatial information, two-dimensional neighborhood structure, and translation invariance. In the transformer architecture used for image classification, the self-attention layers are global, but only the MLP layers are local and translational invariant. In the beginning of the model, the image is divided into patches, and at the time of fine-tuning, the position embeddings are appended to account for images of varied resolutions. Other than that, all spatial interactions between the patches must be learned from start because the position embeddings at initialization time include no information about the 2D positions of the patches [12].

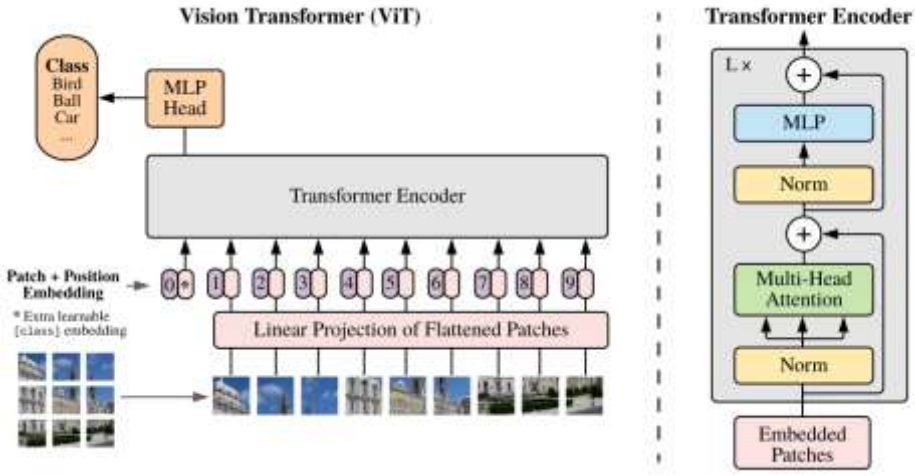


Fig. 1 Schematic view of Transformer Architecture [12]

## 2. Experiments and Results

For the experimental analysis, the 12,568 imagesavailable under common steel surface defect dataset provided byKaggleis used. Real-world industrial application scenarios have exploited this dataset [15]. There are four different class categories in this dataset: Class 1, Class 2, Class 3, and Class 4. According to the study presented in [16], Class 1 has conditions with pitted surfaces, Class 2 has conditions with crazing, Class 3 has conditions with scratches, and Class 4 has conditions with patches. Figure 7 provides a comprehensive explanation of the defectconditions and nature. Also, we used an Nvidia Tesla P4 to train the dataset, with split ratio of 80% for training, 10% for validation, and 10% for testing.

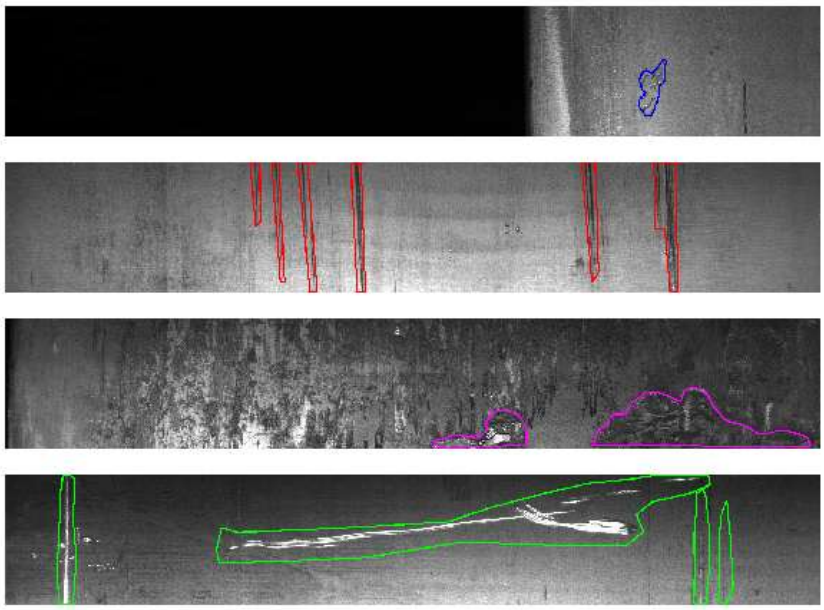

Fig. 2 Sample images with different category of defects [17]

Through a series of transformations, an input picture of shape (height, width, channels) is encapsulated into a vector of shape (n+1, d) in the first stage. The first input to the stacked transformer encoders is the outcome, z0. The second element of the architecture is represented by the L stacked encoders. Each transformer receives as input features expressed as a (n+1, d) tensor, and creates an output that has the identical dimension. Using a stack of L transformer encoders, the network learns additional feature representations from the embedded patches in the second step. The encoder component incorporates a multi-headed attention (MHA) method and a 2-layer MLP, with layer normalization and residual connectivity in between.Layer normalization helps to stabilize hidden state complexities and to minimize the training time. Scaling each training example's mean and standard deviation is how it is done (as opposed to the batch norm where this is done per feature). A scaling factor and a shifting factor, both of which can be learned during training, are multiplied by the generated features. In order to address the issue of vanishing gradients in extremely deep architectures, residual connections provide gradients with alternate routes. The learnable weights in this part lie inside the MHA method and the MLP weights.

The transformer model was trained with a batch size of 32, learning rate of 3e-5, and with a weight decay rate of 0.01. The Adam optimizer was used with a weight decay rate implementation as suggested in [18]. Adam would gain a lot with a planned learning rate multiplier. The fact

that Adam is an adaptive gradient method and as such modifies the learning rate for each component does not eliminate the potential to considerably improve its performance by utilizing a global learning rate amplifier, scheduled, such as cosine annealing. By separating the weight decay from the gradient-based update, the regularization in Adam is intended to be improved. With a 15% relative improvement in test error, the work presented in [18] demonstrated through a thorough analysis that Adam generalizes significantly better with decoupled weight decay than with L2 regularization.In comparison to the original ViT models, [19] DeiT (data-efficient image transformers) are trained transformers for image classification that are more effective feature extractor. They require significantly less data and computer power.The DeiT-Tiny model has 5M parameters and 12 layers in MLP and has an embedding dimension of 192. The results of the experimental analysis are evaluated using the conventional metrics such as precision, accuracy, and Recall. The performance of the proposed transformer based classifier is compared in terms of the training accuracy with the ResNET-50 and presented in Fig. 3

**Table 1. Performance Evaluation of DeiT-Tiny based transformer and ResNET-50 model**

| Model | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| DeiT-Tiny Transformer | 87.8 | 86.8 | 87.30 | 88.4 |
| ResNET-50 | 85.4 | 84.6 | 84.99 | 85.2 |

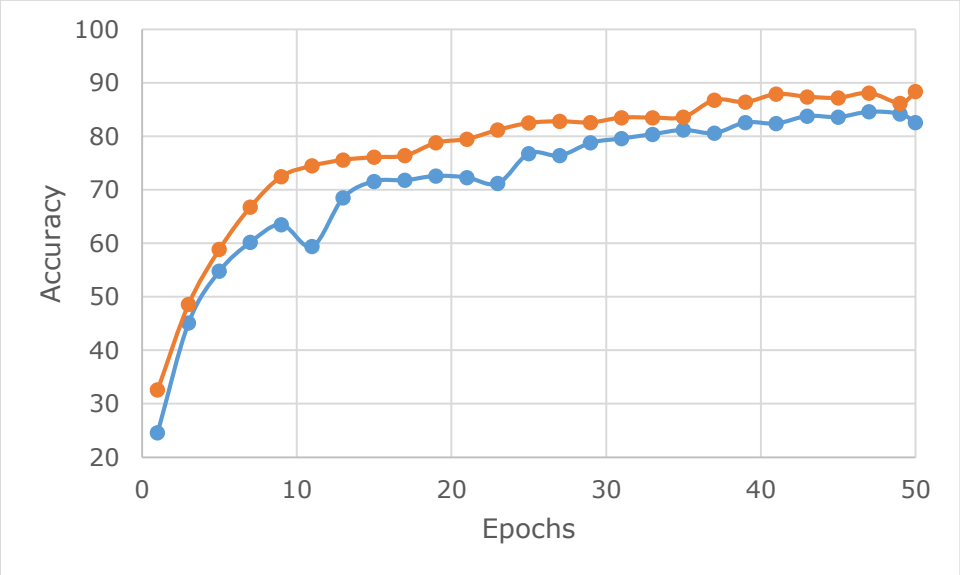


Fig. 3 Training Accuracy Comparison

## 3. Conclusion

The objective of this study is to evaluate the performance of the transformer based classifier against the most widely used convolutional neural network based classification process. Stacks of encoder blocks constitutes a transformer architecture. These blocks are multilayer networks made up of simple linear layers, feedforward networks, and self-attention layers. In terms of training time and computational requirements the base transformer model requires more resources and time. But the DeiT-Tiny model used for feature extraction requires less resources during the training period. Vision Transformers have produced

encouraging outcomes despite being relatively new architectural designs. They have inspired a tonne of scientific attention, hence many research need to be carried to simplify the computational requirements and improve their performance. This study explored the efficiency of the transformer architecture in comparison with the pre-trained CNN model.

## 4. Reference

**1.** Aggarwal, Akarsh, and Manoj Kumar. "Image surface texture analysis and classification using deep learning." Multimedia Tools and Applications 80 (2021): 1289-1309.

**2.** Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

3. Masci, Jonathan, et al. "Steel defect classification with max-pooling convolutional neural networks." The 2012 international joint conference on neural networks (IJCNN). IEEE, 2012.

4. Zhou, Daquan, et al. "Deepvit: Towards deeper vision transformer." arXiv preprint arXiv:2103.11886 (2021).

5. Zhang, Pengchuan, et al. "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

6. Jong Pil Yun, WoosangCrino Shin, Gyogwon Koo, Min Su Kim, Chungki Lee, Sang Jun Lee, Automated defect inspection system for metal surfaces based on deep learning and data augmentation, Journal of Manufacturing Systems, Volume 55, 2020, Pages 317-324.

7. Konovalenko, Ihor, et al. "Steel surface defect classification using deep residual neural network." Metals 10.6 (2020): 846.

8. Fu, Guizhong, et al. "A deep-learning-based approach for fast and robust steel surface defects classification." Optics and Lasers in Engineering 121 (2019): 397-405.

9. Ashish Vaswani, Noam Shazeer, NikiParmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and IlliaPolosukhin. Attention is all you need. In NIPS, 2017.

10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.

11. Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, PrafullaDhariwal, ArvindNeelakantan, Pranav Shyam, GirishSastry, Amanda Askell, et al. Language models are few-shot learners. arXiv, 2020.

12. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

13. Masci, J.; Meier, U.; Ciresan, D.; Schmidhuber, J.; Fricout, G. Steel defect classification with Max-Pooling Convolutional Neural Networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Brisbane, QLD, Australia, 10–15 June 2012; pp. 1–6.

14. Lee, S.Y.; Tama, B.A.; Moon, S.J.; Lee, S. Steel Surface Defect Diagnostics Using Deep Convolutional Neural Network and Class Activation Map. Appl. Sci. 2019, 9, 5449.

15. Boikov, A.; Payor, V.; Savelev, R.; Kolesnikov, A. Synthetic data generation for steel defect detection and classification using deeplearning. Symmetry 2021,13, 1176.

16. Wang, S.; Xia, X.; Ye, L.; Yang, B. Automatic detection and classification of steel surface defect

using deep convolutional neuralnetworks. Metals 2021,11, 388.

17. Gonçalves, Luan &Guerreiro, Tiago &Pinherio, Samya&Brito, Flávio&Nascimento, Ingrid & Linder, Neiva &Klautau, Aldebaro. (2020). Splitted Neural Networks for Equipment Inspection: Experiments and Resources. 10.14209/SBRT.2020.1570661611.
18. Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
19. Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International conference on machine learning. PMLR, 2021.