Efficient Transformer-based Identification of Defects in Fasteners

Simone Garau[1,2,†], Giorgio Leonardi[1,†], Stefania Montani[1,†], Luigi Portinale[1,*,†], Mario Brumini[2,†], Guido Noce[2,†] and Enrico Ottaviano[2,†]

[1]Computer Science Institute, University of Piemonte Orientale, Viale Teresa Michel 11, Alessandria, 15121, Italy
[2]DIMAC srl, S.S. N.10 per Voghera 93/5/D, Tortona, Alessadria, 15057, Italy

Abstract

In the fastener manufacturing domain, ensuring product quality is of mandatory for the safety and reliability of critical applications. Traditional manual inspection face limitations in speed, repeatability and objectivity, driving the adoption of automated solutions. While mechanical sorting and early artificial vision systems addressed dimensional conformity and structural integrity, surface defects detection remains challenging. Recent technological advancements, particularly the integration of Artificial Intelligence (AI) and Deep Learning (DL) with artificial vision, have enabled precise and reliable detection of surface defects, surpassing human capabilities at lower costs. Technologies like segmentation allow for comprehensive analysis of components at high speeds, addressing long-standing automation needs. However, the broad applicability of current AI/DL algorithms needs dedicated customized training, posing a barrier to widespread adoption. This work investigates the use of a transformer-based architecture specifically for the detection of defects in small metal parts in the fastener sector, aiming to improve accuracy and to support the serial identification of defective parts in industrial fastener quality control.

Keywords

Artificial vision systems, Deep Learning, Quality control, Industrial research

1. Introduction

In the fastener manufacturing sector (which includes production of screws, bolts, rivets, and threaded rods), product quality is fundamental to ensure safety and reliability of the final applications [1]. Even minimal defects, like cracks, burrs, or imperfections in the threading, can compromise the structural integrity of the components, with potentially serious consequences in critical areas such as automotive, aerospace, and construction domains [2]. Historically, quality control in this sector relied on manual inspections [3] which, while effective, had limitations in terms of speed, repeatability, and objectivity. With mass production and the need for increasingly high quality standards, both for safety purposes and to ensure operational continuity of automated assembly systems, adoption of automated solutions for selection and rejection of defects [4] became essential. Initially, this involved mechanical sorting systems

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
†These authors contributed equally.
20034352@studenti.uniupo.it (S. Garau); giorgio.leonardi@uniupo.it (G. Leonardi); stefania.montani@uniupo.it (S. Montani); luigi.portinale@uniupo.it (L. Portinale); mario.brumini@dimacsrl.com (M. Brumini); guido.noce@dimacsrl.com (G. Noce); enrico.ottaviano@dimacsrl.com (E. Ottaviano)

while, starting from the last decade of the last century, machinery equipped with artificial vision systems and non-destructive testing techniques are being employed. Such solutions can verify dimensional conformity of products and structural integrity of materials. On the other hand, surface defects detection of lines, scratches, stains, oxidation and blemishes has historically been limited by the drawbacks of traditional image analysis techniques. Only recently, technological evolution has made it possible to combine artificial vision systems with Artificial Intelligence (AI) and Deep Learning (DL) [5]. This step has enabled solutions for the detection of surface defects with precision and reliability comparable to (or higher than) human visual inspection, and with drastically lower costs. For example, technologies such as segmentation allow for a complete and detailed analysis of each component at high speeds, across a wide range of surface defects, cracks, occlusions, and threading issues; this addresses an automation need expressed by the market for years, without a satisfactory answers yet. The AI/DL algorithms currently available to integrators and machine manufacturers are intended for a broad audience and meet the control needs arising from very different sectors. This involves delicate and customized training, which is an obstacle to the diffusion of AI-equipped machines, requiring specific skills and a validation phase for each individual scenario. The work presented here aims to investigate the use of a transformer-based architecture for the detection of defects in small metal parts, aiming to improve accuracy and execution speed within the fastener industrial environment. The present work has been performed in the industrial environment provided by Dimac srl[1], an Italian company leader in the field of fastener quality control.

2. The approach

In our approach, the identification of defects in small metal parts has been treated as a binary classification task with classes `ok` and `defective`. The classification architecture we have decided to adopt is based on a Vision Transformer (ViT) backbone [6], since ViTs have been reported as a suitable architectural solution in the fastener classification task [7]. A ViT is a transformer [8] specifically designed to work with image data; it decomposes an input image into a set of patches, serializes each patch into a vector and maps it to a reduced dimension through embedding matrix multiplication. The attention mechanism in a ViT repeatedly transforms representation vectors of image patches, incorporating more and more semantic relationships between them. Thanks to the attention map, the model visually highlights the image portions that were deemed as the most important for classification. With respect to Convolutional Neural Networks (CNN) [9], ViTs can capture long-range relationships between image patches; they require more training data, but can progressively improve their performance. CNNs, instead, apply the same filter to the entire image; they are less sensitive to local perturbation, but might not improve after training on a large enough training dataset. Moreover, ViTs are also very robust to input image distortions [10]. In our application we rely on the `google/vit-large-patch16-224-in21k` model, described in [6], and available on the Hugging Face models hub[2]. Such a model is pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224×224. Figure 1 shows the ViT architecture we have adopted.

[1] https://www.dimacsrl.com
[2] https://huggingface.co/

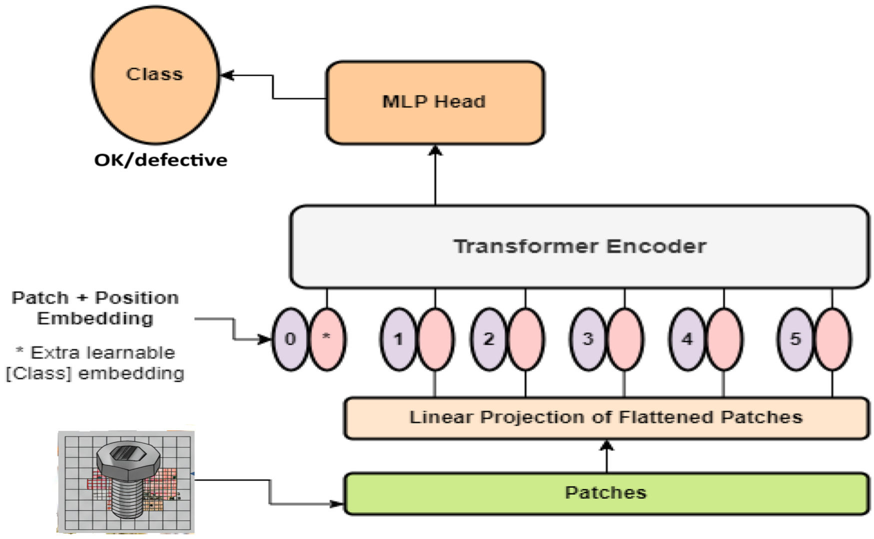

Figure 1: ViT classification architecture

Because of a limited availability of training data, we exploited a pre-trained ViT, by fine-tuning the model in a transfer learning perspective, by allowing the model to adapt to specific patterns in the target data without starting from scratch. Transfer learning is an essential option when a sufficiently large training set is not available in the target domain, and it has been proved to achieve reduced training time and improved accuracy with limited data [11]. To fine-tune the model we used three datasets obtained from Dimac control machineries:

- **D1: Bi-component**: it contains 970 OK (i.e., non defective) and 580 defective images. This dataset was used as-is, without any augmentation, since the classes were fairly balanced.
- **D2: Rounded Pieces**: it is composed by 900 OK and 230 defective images. This dataset is unbalanced, so we applied data augmentation during training using random rotations and horizontal/vertical flips in the images. We also used a weighted sampler to give more importance to the minority class (`defective`).
- **D3: Screws with Thread Focus**: it consists of 1500 OK and 300 defective images. Due to the high imbalance, we applied extensive augmentation to generate a balanced dataset. The augmentation added images with random changes in brightness and contrast, rotations, translations, crops, and resizing.

Figures 2, 4 and 6 show images of defective pieces from dataset D1, D2 and D3, respectively, while figures 3, 5 and 7 show the corresponding attention maps, highlighting the defects, as produced by the attention layers of the ViT encoder. The capability of producing heat maps based on attention (typical of transformers) is very important in order to provide visual insights (usueful for inspection) once a defect in a piece has been discovered. To organize the datasets

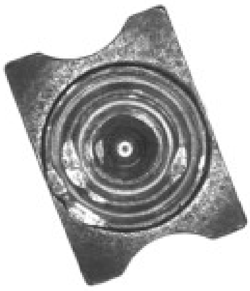

Figure 2: Example of a Bi-component piece with defect (dataset D1).

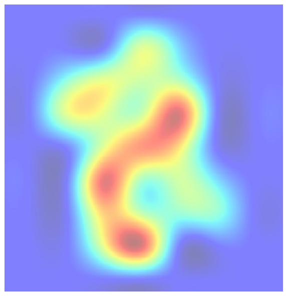

Figure 3: Attention map produced by a ViT on the piece on the left

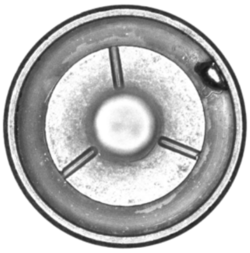

Figure 4: Example of a round piece with defect (dataset D2).

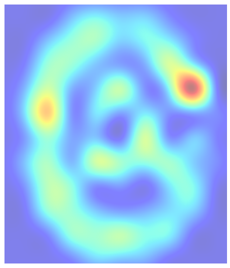

Figure 5: Attention map produced by a ViT on the piece on the left

Figure 6: Example of a screw with defect (dataset D3).

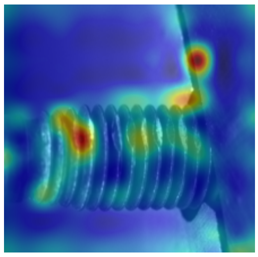

Figure 7: Attention map produced by a ViT on the piece on the left

into training, validation, and test sets, we developed a Python script that automatically creates the required directory structure and distributes the images accordingly. In all datasets we performed a stratified split, ensuring that each set maintains the same class distribution as the original dataset. In some of them, as explained above, we also applied augmentation. The split percentage we used is 70/15/15: 70% for training, 15% for validation, and 15% for testing.

3. Experimental results

To test the classification model we adopted an early-stopping strategy with a patience parameter of 5 for the experiments on the D1 and D2 datasets, and of 8 for D3. The maximum number of

Class	Precision	Recall	F1-score
ok	1.00	1.00	1.00
defective	1.00	1.00	1.00
Accuracy			1.00
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00

Table 1
Classification report for the ViT model on the D1 (Bi-component) dataset

Class	Precision	Recall	F1-score
ok	0.97	0.94	0.96
defective	0.99	0.99	0.99
Accuracy			0.98
Macro Avg	0.98	0.97	0.97
Weighted Avg	0.98	0.98	0.98

Table 2
Classification report for the ViT model on the D2 (Rounded Pieces) dataset

epochs was set to 20, while early stopping was triggered at epoch 10 on all the experiments we have performed. Table 1 shows the classification performance of the fine-tuned ViT model on the D1 (Bi-component) validation dataset. Table 2 shows the results on the D2 (Rounded Pieces) dataset and table 3 shows the classification performance on the D3 (Screws) dataset.

Class	Precision	Recall	F1-score
ok	1.00	1.00	1.00
defective	1.00	1.00	1.00
Accuracy			1.00
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00

Table 3
Classification report for the ViT model on the D3 (Screws) dataset

Tables 1, 2 and 3 show results where accuracy, precision, recall and F1-score are very close to or equal to 1. On one hand, this result is very positive, and testifies that the approach of selecting pre-trained visual transformers and specializing them through fine-tuning allows one to obtain models for defect recognitions which are effective in terms of classification performance, in particular when dealing with limited data. On the other hand, such high performance can point out problems such as overfitting or datasets where the defects are too similar to each other, or a poorly representative sample distribution. Analyzing the images with the DIMAC experts, however, it emerged that the datasets are not affected by these problems; the images are selected to show a range of typical defects in the areas covered by all 3 datasets used for these experiments. A possible explanation given by the industrial experts is that the range of defects in these images, unless wide, is detectable with good precision even by human operators. The superior ability of ViT models in the automatic detection, leads the latter to obtain even higher accuracy, with the advantage of supporting the serial control of the pieces produced by the machines. In the future, we plan to extend these experiments to an even wider range of cases, to further assess the results obtained so far. Table **??** shows training and validation over 10 epochs for the training session on Bi-component images. Table 5 refers to the Round Pieces dataset and table 6 refers to the Screw dataset. The plots of training and validation loss and accuracy on the Bi-component dataset are reported in figures 8 and 9 respectively. The same

Epoch	Train Loss	Train Acc.	Val Loss	Val Acc.
1	0.1131	0.9640	0.0480	0.9828
2	0.0172	0.9935	0.0008	1.0000
3	0.0055	0.9982	0.0004	1.0000
4	0.0003	1.0000	0.0002	1.0000
5	0.0002	1.0000	0.0002	1.0000
6	0.0001	1.0000	0.0001	1.0000
7	0.0001	1.0000	0.0001	1.0000
8	0.0001	1.0000	0.0001	1.0000
9	0.0001	1.0000	0.0001	1.0000
10	0.0001	1.0000	0.0001	1.0000

Table 4
Training and validation metrics over the Bi-component dataset.

Epoch	Train Loss	Train Acc.	Val Loss	Val Acc.
1	0.5149	0.7270	0.1934	0.9471
2	0.1071	0.9610	0.3024	0.9118
3	0.0927	0.9648	0.0688	0.9824
4	0.0644	0.9761	0.0933	0.9588
5	0.1348	0.9535	0.1218	0.9588
6	0.0427	0.9836	0.0751	0.9824
7	0.0082	0.9987	0.0583	0.9765
8	0.0282	0.9937	0.0566	0.9706
9	0.0258	0.9887	0.2586	0.9294
10	0.0238	0.9925	0.0975	0.9706
11	0.0238	0.9925	0.0578	0.9824
12	0.0033	1.0000	0.0692	0.9706
13	0.0170	0.9925	0.1363	0.9529

Table 5
Training and validation metrics over the Rounded Pieces dataset.

metrics are reported in figures 10 and 11 on the Rounded Pieces dataset and in figures 12 and 13 on the Screws dataset. The plots related to all these datasets confirm that the fine tuning phase on the datasets provided by DIMAC enables the models to specialize and to learn quickly, reaching optimum performance after few epochs.

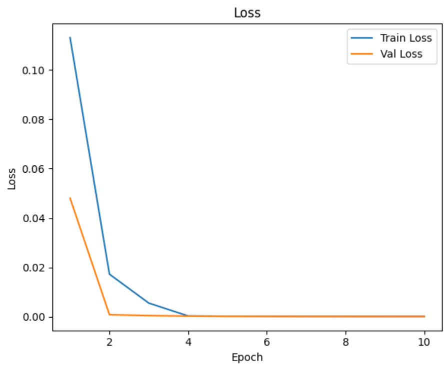

Figure 8: Training and validation loss over epochs in the Bi-component dataset.

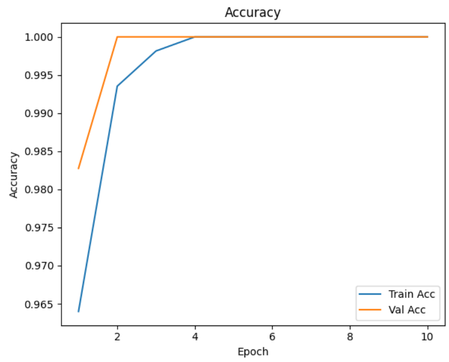

Figure 9: Training and validation accuracy over epochs in the Bi-component dataset.

4. Conclusion and future work

In the fastener manufacturing domain, ensuring product quality is a key factor for the safety and reliability of critical applications. In the present paper, we presented a deep learning classification approach, relying on a visual transformer-based model, for the identification of defective parts from fastener images, able to improve accuracy in defect detection. In particular, our approach demonstrates the effectiveness of a transfer learning approach, using a pre-trained ViT model for the automated classification of defects in fastener manufacturing. The fine-tuned model achieved high accuracy across three different datasets, showing its potential to improve

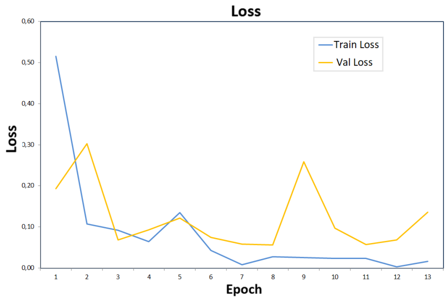

Figure 10: Training and validation loss over epochs in the Rounded Pieces dataset.

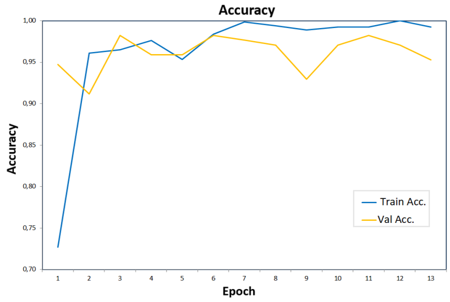

Figure 11: Training and validation accuracy over epochs in the Rounded Pieces dataset.

Epoch	Train Loss	Train Acc.	Val Loss	Val Acc.
1	0.0165	0.9957	0.0126	0.9978
2	0.0890	0.9713	0.0096	0.9978
3	0.0792	0.9770	0.0754	0.9732
4	0.0445	0.9871	0.0234	0.9911
5	0.0202	0.9938	0.0349	0.9911
6	0.0331	0.9909	0.0370	0.9888
7	0.0136	0.9971	0.0590	0.9777
8	0.0057	0.9986	0.0406	0.9911
9	0.0010	0.9990	0.0571	0.9911
10	0.0101	0.9976	0.0200	0.9978

Table 6
Training and validation metrics per epoch over the Screws dataset

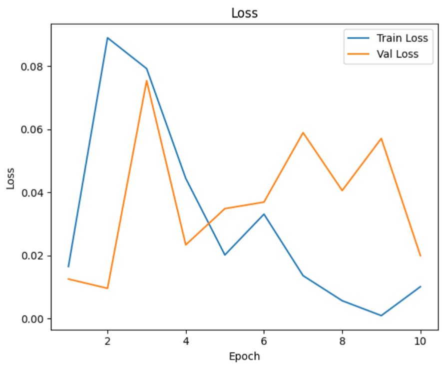

Figure 12: Training and validation loss over epochs in the Screws dataset.

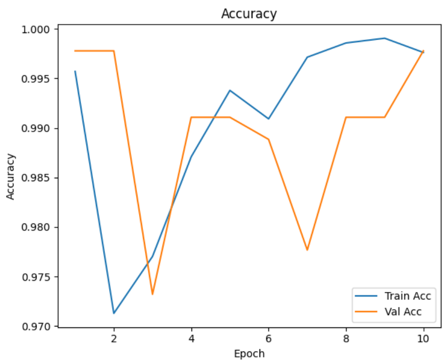

Figure 13: Training and validation accuracy over epochs in the Screws dataset.

quality control processes in the sector. Future work may focus on further optimizing the model's architecture and training process, extending the experimental part with larger industrial real-world datasets, covering a broader set of defect types and of modified imaging conditions such as, for example, different light, soft focus and camera positioning. We also plan to explore the application of these models to other types of industrial defect detection.

Acknowledgments

The present work has been performed within the project QAIS (Quality and Artificial Intelligence Services) funded by Regione Piemonte and European Union under the SWITCH program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Hudgins, B. James, Fatigue of threaded fasteners, AM&P Technical Articles (2014).
[2] C. H. Cundiff, R. W. Crites, Structural integrity of fasteners, in: NASA Conference Publication 3296, NASA Langley Research Center, 1995. URL: https://tinyurl.com/2d2pnkyr.
[3] ASTM International, Standard practice for fastener sampling for specified mechanical properties and performance inspection (astm f1470-24), 2024. URL: https://tinyurl.com/2zrmv74c.
[4] Int. Org. for Standardization, Fasteners — quality assurance system, 2002. ISO 16426.
[5] V. Verdhan, Computer Vision using Deep Learning, Apress, 2021.
[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.
[7] N. Cauli, M. Murgia, D. R. Recupero, G. Scarpi, Leveraging transformers architectures and augmentation for efficient classification of fasteners and natural language searches, Discov. Comput. 27 (2024) 10. doi:10.1007/S10791-024-09443-8.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.
[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[10] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, 2022. URL: https://arxiv.org/abs/2108.08810. arXiv:2108.08810.
[11] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, M. A. Azim, Transfer learning: a friendly introduction, J. Big Data 9 (2022) 102. doi:10.1186/S40537-022-00652-W.