Aprimorando o Controle de Qualidade na Fabricação por Fundição com Vision Transformers

Lucas Matos A. Dias[1], **Emanuelle S. Gil**[1], **Alternei S. Brito**[1], **Felipe G. Oliveira**[1]

[1]Instituto de Ciências Exatas e Tecnologia
Universidade Federal do Amazonas (UFAM) – Itacoatiara, AM – Brazil

emanuelle.gil@ufam.edu.br, lucas.dias@ufam.edu.br,

felipeoliveira@ufam.edu.br, alternei@ufam.edu.br

Abstract. *Quality control is vital in modern manufacturing to ensure product reliability and competitiveness. This paper addresses defect detection in casting discs for submersible pump impellers using automated visual inspection. We propose a method based on Vision Transformers (ViT), which leverage self-attention to learn visual patterns effectively. Real and simulated experiments showed high accuracy (99.22%) and strong robustness to image noise, maintaining 95.45% and 98.28% accuracy under Gaussian and Salt-and-Pepper noise. The results confirm the method's reliability and potential to optimize the casting process.*

Resumo. *O controle de qualidade é essencial na manufatura moderna para garantir a confiabilidade do produto e a competitividade. Este artigo aborda a detecção de defeitos em discos de fundição para impulsores de bombas submersíveis utilizando inspeção visual automatizada. Propomos um método baseado em Vision Transformers (ViT), que utilizam mecanismos de autoatenção para aprender padrões visuais de forma eficaz. Experimentos reais e simulados mostraram alta precisão (99,22%) e forte robustez a ruídos nas imagens, mantendo 95,45% e 98,28% de precisão com ruídos Gaussiano e Sal-e-Pimenta, respectivamente. Os resultados confirmam a confiabilidade do método e seu potencial para otimizar o processo de fundição.*

1. Introdução

Nas últimas décadas, os avanços em inteligência artificial proporcionaram melhorias significativas nos processos industriais, especialmente na garantia da qualidade. No entanto, setores como circuitos integrados [Rocha et al. 2016], semicondutores [Silva. et al. 2022] e fundição [Omar et al. 2022] ainda demandam maior automação. A fundição é um processo onde um material líquido é vertido em um molde e deixado para solidificar [Duan et al. 2021].

No entanto, esse processo pode apresentar defeitos como bolhas de ar, microporos, rebarbas, retrações e falhas metalúrgicas. A inspeção visual automatizada tem se mostrado eficaz na identificação desses problemas, contribuindo para o aumento da eficiência na produção [Rocha et al. 2016].

Neste artigo, propomos uma abordagem baseada em Vision Transformer para detecção de falhas em rotores de bombas submersíveis durante a fundição. A arquitetura inclui etapas de Divisão em Patches, Embedding e Codificador Transformer, para

aprender características visuais representativas. Os experimentos utilizam um conjunto de dados consolidado, com variações de posição e iluminação dos discos fundidos.

As principais contribuições do nosso trabalho são resumidas a seguir:

- Apresentamos um método que utiliza a arquitetura Vision Transformer (ViT), uma técnica de classificação de ponta, para identificar defeitos no processo de fabricação por fundição. Essa abordagem inovadora oferece uma solução robusta para a inspeção complexa de discos fundidos utilizados em rotores de bombas submersíveis.
- Conduzimos experimentos extensivos para validação da nossa metodologia, utilizando um conjunto de dados bem definido sob condições reais e simuladas. Esses experimentos visaram comparar quantitativamente modelos de aprendizado profundo com o nosso modelo proposto, especialmente em cenários desafiadores envolvendo ruído. Os resultados fornecem insights significativos que podem orientar futuras pesquisas nessa área.

2. Metodologia

Este artigo aborda o problema de inspeção visual automática para detecção de falhas em fundição, baseado em Vision Transformer [De Souza Gil et al. 2024]. A metodologia proposta aborda a detecção de falhas durante a produção de peças metálicas em processos de fundição, cujos detalhes serão apresentados nas próximas subseções.

2.1. Arquitetura Vision Transformer para Classificação em Fundição

Vision Transformer (ViT) é uma arquitetura de rede neural que aplica o modelo transformer, originalmente desenvolvido para NLP, à classificação de imagens. ViTs demonstram desempenho competitivo, frequentemente superando CNNs em benchmarks [Dosovitskiy et al. 2021]. A arquitetura é composta por três etapas principais: Divisão em Patches, Embedding dos Patches e Codificador Transformer, descritas nas subseções a seguir.

2.1.1. Divisão em Patches

Dada uma imagem de entrada () de tamanho $(H \times W \times C)$, onde H e W são a altura e largura, e C é o número de canais, a imagem é dividida em patches não sobrepostos de tamanho $P \times P$. O número de patches N é dado por:

$$N = \frac{H.W}{P^2}. \tag{1}$$

Cada patch X_i é então achatado em um vetor:

$$X_i \in \mathbb{R}^{P^2.C}, \quad i = 1, 2, ..., N. \tag{2}$$

2.1.2. Embedding dos Patches

Os patches achatados são projetados linearmente em um espaço de embedding de menor dimensão usando uma matriz de projeção aprendível E:

$$Z_0^i = EX_i + e_{pos}^i, \tag{3}$$

onde $E \in \mathbb{R}^{D\times(P^2.C)}$ é a matriz de embedding aprendível. $e_{pos}^i \in \mathbb{R}^D$ é o embedding posicional para o patch i.

Assim, a sequência de patches embutidos forma a entrada para o transformer:

$$Z_0 = [Z_0^1; Z_0^2; ...; Z_0^N] \in \mathbb{R}^{N\times D}. \tag{4}$$

2.1.3. Codificador Transformer

O codificador transformer consiste em L camadas, cada uma composta por blocos de Multi-head Self-Attention (MSA) e Perceptron Multicamadas (MLP) com Normalização de Camada (LN) e conexões residuais:

Multi-Head Self-Attention (MSA): Para cada camada l, o mecanismo de autoatenção calcula pontuações de atenção A e saídas, considerando os relacionamentos entre diferentes patches dados por:

$$A = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{5}$$

onde as matrizes de consulta (Q), chave (K) e valor (V) são calculadas como: $Q = Z_{l-1}W_Q$; $K = Z_{l-1}W_K$; $V = Z_{l-1}W_V$. As matrizes de projeção aprendíveis são $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$.

Depois, as saídas de múltiplas cabeças de atenção são combinadas da seguinte forma:

$$Z_{MSA} = Concat(head_1, head_2, ..., head_h)W_0 \tag{6}$$

onde h é o número de cabeças de atenção e $W_0 \in \mathbb{R}^{(h.D)\times D}$ é a matriz de projeção de saída.

Normalização de Camada e Conexões Residuais: A normalização de camada (LN) é aplicada para estabilizar e acelerar o treinamento, normalizando a entrada nas características de cada amostra. No codificador transformer, a normalização de camada é aplicada após a adição da conexão residual da saída do MSA à entrada do bloco MSA. A operação LN para uma entrada x é definida como:

$$Z_l^{'} = LN(Z_{l-1} + Z_{MSA}) \tag{7}$$

Perceptron Multicamadas (MLP): O bloco MLP consiste em duas camadas totalmente conectadas com uma função de ativação não linear GELU (Gaussian Error Linear Unit). Esse bloco é responsável por transformar ainda mais a representação dos patches de entrada. O bloco MLP é definido como:

$$Z_l = LN(Z_l^{'} + MLP(Z_l^{'})) \tag{8}$$

onde:

$$MLP(Z) = max(0, ZW_1 + b_1)W_2 + b_2 \tag{9}$$

com W_1, W_2 e b_1, b_2 sendo parâmetros aprendíveis.

Cabeça de Classificação: Após o processamento pelo codificador transformer, um token de classificação z_L^0 (previamente adicionado à sequência de entrada) é utilizado para a classificação final:

$$Output = softmax(W_{cls} z_L^0) \tag{10}$$

onde W_{cls} é uma matriz de projeção aprendível.

Nossa arquitetura ViT extrai patches de 16×16, projeta-os em uma dimensão 8, utiliza 4 cabeças de autoatenção, 8 blocos codificadores transformer e apresenta unidades MLP de 512 e 256. O treinamento do modelo ViT emprega o otimizador Adam (taxa de aprendizado 0,001), lote de 16, validação cruzada 5-fold, com 20 épocas por divisão. A acurácia média e o desvio padrão são usados como métricas. A Figura 1 ilustra a arquitetura.

3. Experimentos

3.1. Conjunto de Dados

Para avaliar a abordagem proposta, utilizamos o conjunto de dados Casting Product Image [Dabhi 2020], composto por 7348 imagens em tons de cinza (300×300 pixels) de discos de fundição vistos de cima. Para melhorar a robustez e generalização do modelo, o conjunto foi expandido artificialmente com técnicas de aumento de dados, como cutout, mixup e color jittering. A Figura 2 apresenta exemplos de discos de fundição que compõem o conjunto de dados.

3.2. Detalhes de Implementação

Utilizamos os frameworks OpenCV e TensorFlow em um computador Dell com processador Intel® Xeon^{TM} Silver 4114 2.20GHz, memória principal DDR4-2133 de 128 GB e uma GPU NVIDIA® GeForce® RTX A4000 de 16 GB GDDR6. A fase de treinamento do modelo ViT proposto envolveu o uso de Grid Search para otimizar os hiperparâmetros visando alta acurácia.

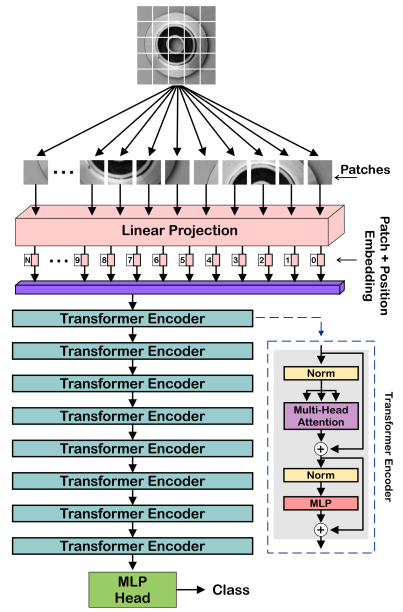


Figura 1. Arquitetura ViT proposta.

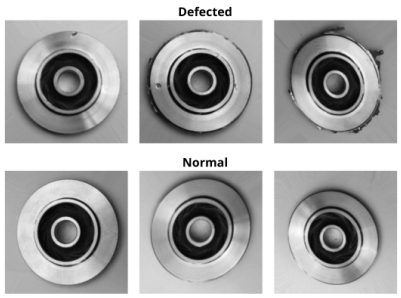


Figura 2. Exemplos de discos de fundição que compõem o conjunto de dados adotado, destacando peças defeituosas e normais.

3.3. Avaliação da Classificação de Falhas em Fundição

Este experimento avalia a acurácia do método proposto para inspeção de discos de fundição, comparando seis estratégias de classificação: i) Vision Transformer (ViT) proposto; ii) CNN; iii) CNN com Random Forest (RF); iv) CNN com SVM; v) VGG16; e vi) EfficientNet. Técnicas de aprendizado profundo são utilizadas por seu desempenho superior em inspeção visual automática [Omar et al. 2022, Dong et al. 2018, Kumaresan et al. 2021, Tan and Le 2019, Simonyan and Zisserman 2015].

Os resultados, apresentados na Tabela I, mostram que o ViT supera os demais métodos em acurácia e desvio padrão. Isso se deve à sua capacidade de capturar relações espaciais complexas por meio de autoatenção, permitindo uma representação hierárquica mais eficaz das características visuais [Dosovitskiy et al. 2021].

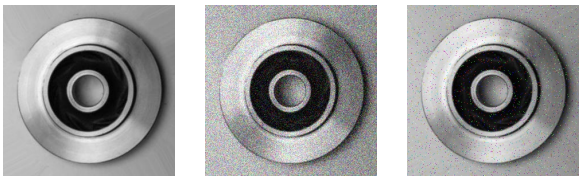
3.4. Avaliação da Classificação de Falhas em Fundição na Presença de Ruído

Neste experimento, avaliamos a robustez da abordagem proposta para inspeção de fundição sob condições de ruído, aplicando ruído Gaussiano e Sal e Pimenta às imagens. Avaliamos diferentes técnicas de aprendizado profundo, incluindo o modelo ViT

Tabela 1. Resultados para Detecção de Falhas em Fundição, como um problema de classificação. Este experimento apresenta a acurácia dos métodos ViT (nosso), CNN, CNN + RF, CNN + SVM, VGG 16 e Efficient Net.

Método	Acurácia
CNN [Omar et al. 2022]	92.00 ± 1.70
CNN + RF [Dong et al. 2018]	91.94 ± 1.14
CNN + SVM [Kumaresan et al. 2021]	92.97 ± 2.44
VGG 16 [Simonyan and Zisserman 2015]	98.21 ± 0.57
Efficient Net [Tan and Le 2019]	98.91 ± 0.28
ViT (Nosso)	**99.22 ± 0.17**

proposto. O treinamento foi realizado com imagens sem ruído, enquanto o teste utilizou imagens ruidosas, por meio de aprendizado por transferência.



((a)) ((b)) ((c))

Figura 3. Exemplos de imagens de discos de fundição. Na Figura 3(a), é apresentada uma imagem de disco de fundição sem ruído. A Figura 3(b) e 3(c) apresentam a mesma imagem com ruído Gaussiano e Sal e Pimenta, com densidade de ruído 0.02.

Os resultados deste experimento indicam que o modelo ViT proposto apresenta desempenho superior mesmo em condições com ruído, como mostrado nas Tabelas 2 e 3. A Tabela 2 apresenta os resultados de classificação para ruído Gaussiano, incluindo acurácia e desvio padrão. A Tabela 3 apresenta os resultados para ruído Sal e Pimenta, também com acurácia e desvio padrão. Esses resultados demonstram que a abordagem proposta supera outros métodos de classificação, destacando sua robustez.

Tabela 2. Resultados da avaliação da robustez da detecção de falhas de fundição, como um problema de classificação, considerando a presença de ruído. Nesse experimento, todos os métodos considerados são avaliados usando ruído Gaussiano.

	Gaussiano		
Densidade de ruído	**0.005**	**0.01**	**0.02**
CNN [Omar et al. 2022]	90.36 ± 4.42	88.21 ± 5.27	84.89 ± 7.26
CNN + RF [Dong et al. 2018]	88.34 ± 9.98	83.98 ± 10.02	81.27 ± 8.29
CNN + SVM [Kumaresan et al. 2021]	88.72 ± 7.62	83.93 ± 8.84	82.11 ± 10.23
VGG 16 [Simonyan and Zisserman 2015]	95.23 ± 8.01	94.04 ± 8.71	93.17 ± 9.85
Efficient Net [Tan and Le 2019]	95.34 ± 2.87	94.89 ± 3.69	92.73 ± 5.39
ViT (Nosso)	**96.84 ± 2.91**	**96.01 ± 3.79**	**95.45 ± 4.38**

Tabela 3. Resultados da avaliação da robustez da detecção de falhas de fundição, como um problema de classificação, considerando a presença de ruído. Nesse experimento, todos os métodos considerados são avaliados usando o ruído Sal e Pimenta.

	Sal e Pimenta		
Densidade de ruído	**0.005**	**0.01**	**0.02**
CNN [Omar et al. 2022]	91.58 ± 5.99	88.35 ± 10.68	85.12 ± 9.74
CNN + RF [Dong et al. 2018]	89.11 ± 10.11	82.96 ± 10.20	72.50 ± 8.24
CNN + SVM [Kumaresan et al. 2021]	89.45 ± 7.36	84.58 ± 4.38	75.83 ± 10.85
VGG 16 [Simonyan and Zisserman 2015]	95.71 ± 7.75	94.82 ± 7.82	93.99 ± 9.77
Efficient Net [Tan and Le 2019]	96.31 ± 1.37	95.37 ± 3.19	93.02 ± 10.36
ViT (Nosso)	**98.85 ± 2.84**	**98.51 ± 0.54**	**98.28 ± 1.63**

4. Conclusão

Este artigo aborda o problema da inspeção visual automática de discos utilizados em impulsores de bombas submersíveis. O método proposto, baseado em técnicas de visão computacional e aprendizado profundo, demonstrou alta acurácia na detecção de falhas, inclusive em cenários desafiadores. Além disso, mostrou-se promissor para integração em linhas de produção, contribuindo para o aumento da eficiência, da confiabilidade e do controle de qualidade industrial. Os resultados experimentais confirmam sua viabilidade técnica e robustez sob diferentes condições operacionais, tanto reais quanto simuladas, reforçando seu potencial para aplicações práticas na indústria.

Referências

Dabhi, R. (2020). Casting product image data for quality inspection. `https://www.kaggle.com/datasets/ravirajsinh45/real-life-industrial-dataset-of-casting-product`. Accessed: 2024-05-08.

De Souza Gil, E., De Abreu Dias, L. M., De Souza Brito, A., and Oliveira, F. G. (2024). Enhancing casting manufacturing quality control with vision transformers. In *2024 Brazilian Symposium on Robotics (SBR) and 2024 Workshop on Robotics in Education (WRE)*, pages 162–167.

Dong, X., Taylor, C. J., and Cootes, T. F. (2018). Small defect detection using convolutional neural network features and random forests. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Duan, L., Yang, K., and Ruan, L. (2021). Research on automatic recognition of casting defects based on deep learning. *IEEE Access*, 9:12209–12216.

Kumaresan, S., Aultrin, K. J., Kumar, S., and Anand, M. D. (2021). Transfer learning with cnn for classification of weld defect. *Ieee Access*, 9:95097–95108.

Omar, F., Sohrab, H., Saad, M., Hameed, A., and Bakhsh, F. I. (2022). Deep learning binary-classification model for casting products inspection. In *2022 2nd Int. Conf. on*

Power Electronics IoT Applications in Renewable Energy and its Control (PARC), pages 1–6.

Rocha, C. S., Menezes, M. A., and Oliveira, F. G. (2016). Detecção automática de microcomponentes smt ausentes em placas de circuito impresso. In *Workshop on Industry Applications (WIA) in the 29th Conference on Graphics, Patterns and Images (SIBGRAPI 2016), São José dos Campos, Sp, Brazil*, volume 1.

Silva., C. N., Ferreira., N. P., Meireles., S. S., Otani., M., J. da Silva., V., O. de Freitas., C. A., and Oliveira., F. G. (2022). The visual inspection of solder balls in semiconductor encapsulation. In *Proceedings of the 19th International Conference on Informatics in Control, Automation and Robotics - ICINCO*, pages 750–757. INSTICC, SciTePress.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. conference on machine learning*, pages 6105–6114. PMLR.