Detecção de Anomalias Visuais com Autoencoder Convolucional e Perda Híbrida no Dataset MVTec AD

Vinicius Pereira Tavares de Sousa
Universidade Tecnológica Federal do Paraná
devviniciustavares@gmail.com

***Abstract*—This paper presents an approach for visual anomaly detection using a Convolutional Autoencoder trained with a hybrid loss function that combines Mean Squared Error (MSE) and Structural Similarity Index (SSIM). Applied to the *bottle* category of the MVTec AD dataset, the model was evaluated using average SSIM between original and reconstructed images. Experimental results show that the hybrid loss (0.5 MSE, 0.5 SSIM) outperformed individual losses, achieving an SSIM of 0.9055 ± 0.0161 for normal images and 0.8553 ± 0.0251 for defective ones. These values indicate enhanced class separation and higher sensitivity to structural anomalies, confirming the effectiveness of the hybrid formulation for industrial visual inspection.**

***Index Terms*—Detecção de anomalias visuais, autoencoder convolucional, SSIM, visão computacional, inspeção industrial, MVTec AD.**

I. Introdução

A inspeção visual automática tem papel fundamental no controle de qualidade industrial, sendo responsável por detectar e classificar defeitos em produtos manufaturados. Este processo tradicionalmente é realizado manualmente por operadores treinados, o que pode gerar custos elevados e inconsistências devido à fadiga e subjetividade. Com o avanço da **Inteligência Artificial (IA)** [1] e do **Aprendizado Profundo (AP)** [2], também conhecido como **Deep Learning (DL)** [3], métodos automáticos têm ganhado destaque por sua capacidade de análise precisa e em alta velocidade.

O desafio central na detecção de anomalias está na raridade e diversidade dos defeitos, que tornam inviável a criação de conjuntos balanceados para treinamento supervisionado. Assim, métodos baseados em aprendizado não supervisionado, como autoencoders, têm sido amplamente estudados [4], [5], [6], [7]. Esses modelos aprendem uma representação compacta dos dados normais e detectam anomalias quando uma entrada apresenta alto erro de reconstrução.

Este trabalho propõe o uso de um autoencoder convolucional com função de perda híbrida combinando erro quadrático médio (MSE) e índice estrutural de similaridade (SSIM). Essa combinação visa preservar detalhes estruturais importantes durante a reconstrução, o que, como evidenciado nos resultados, melhora significativamente a sensibilidade na identificação e localização dos defeitos. A aplicação foi realizada na categoria *bottle* do dataset MVTec AD, um benchmark amplamente utilizado para avaliação de métodos de detecção de anomalias visuais.

II. Trabalhos Relacionados

A literatura sobre detecção de anomalias visuais é vasta e crescente. Modelos baseados em autoencoders simples têm sido usados com sucesso, explorando a capacidade de compressão e reconstrução dos dados normais [4], [5]. Porém, a função de perda padrão, geralmente MSE, pode falhar em capturar diferenças perceptuais importantes, especialmente em contextos industriais onde pequenas alterações estruturais são cruciais [8].

Para contornar essas limitações, trabalhos recentes têm incorporado métricas perceptuais na função de perda, como o SSIM, que avalia similaridade em termos de luminância, contraste e estrutura [9], alinhando-se melhor com a percepção humana. Além disso, abordagens adversariais combinam autoencoders com GANs para produzir reconstruções mais realistas, embora com maior complexidade computacional e dificuldade de treinamento [10].

Neste cenário, a proposta deste trabalho é uma solução intermediária, combinando MSE e SSIM de forma simples e eficaz, oferecendo uma boa qualidade de reconstrução sem elevar a complexidade do modelo, com resultados superiores aos obtidos por perdas individuais, como demonstrado experimentalmente.

III. Metodologia

A. Dataset

O dataset MVTec AD [11] é um benchmark público amplamente utilizado para detecção de anomalias visuais em ambientes industriais. Ele contém imagens reais de produtos com e sem defeitos, distribuídas em 15 categorias diferentes.

Nesta pesquisa, focamos na categoria *bottle*, composta por 209 imagens de produtos normais (sem defeito) e 57 imagens contendo defeitos variados, incluindo subcategorias como *broken_large*, *broken_small* e *contamination*.

As imagens possuem resolução original variando entre 1024×1024 até 1920×1080 pixels, com alta qualidade e detalhamento, em formato colorido (RGB). Para simplificar o processamento, todas as imagens foram convertidas para escala de cinza e redimensionadas para 256×256 pixels.

A divisão dos dados para treinamento e teste foi realizada utilizando as imagens normais para treinamento (209 imagens) e imagens defeituosas para teste (19 da subcategoria *broken_large*), além de um subconjunto de imagens normais no teste para comparação.

B. Pré-processamento dos Dados

Todas as imagens foram convertidas para escala de cinza e redimensionadas para 256×256 pixels, permitindo padronização para o modelo. A normalização foi feita para intervalo $[0, 1]$, melhorando estabilidade numérica no treinamento.

Durante o treinamento, apenas imagens sem defeitos foram usadas para que o modelo aprendesse o padrão normal. Para avaliação, imagens defeituosas foram processadas para medir a capacidade do modelo em identificar anomalias.

C. Arquitetura do Autoencoder

O modelo proposto é dividido em duas partes principais: encoder e decoder. O encoder é responsável por extrair uma representação latente compacta da imagem de entrada, enquanto o decoder reconstrói a imagem original a partir dessa representação.

A entrada do modelo são imagens monocromáticas redimensionadas para 256×256 pixels, com um canal único. O encoder é composto por quatro camadas convolucionais com filtros de tamanho 5×5, todas com *stride* 2×2 e preenchimento *same*, o que reduz progressivamente a dimensão espacial da imagem. Cada uma dessas camadas é seguida de uma ativação LeakyReLU, normalização por lotes (Batch Normalization) e Dropout com taxa de 20%, contribuindo para a estabilidade do treinamento e para a prevenção de overfitting.

Em seguida, o volume de ativação resultante é achatado (Flatten) e projetado em um vetor latente unidimensional de dimensão 128 por meio de uma camada densa. A dimensão latente de 128 foi escolhida empiricamente, balanceando compressão e fidelidade. Valores menores causaram perdas de informação, enquanto valores maiores não trouxeram ganhos significativos, apenas aumentando o custo computacional.

O decoder reconstrói a imagem original a partir do vetor latente. Ele inicia com uma camada densa que expande o vetor para uma forma tridimensional $(16, 16, 128)$, seguida de uma sequência de três camadas de convolução transposta (*Conv2DTranspose*) com filtros de tamanho 5×5, *stride* 2×2 e ativação LeakyReLU, cada uma acompanhada por normalização por lotes e Dropout. A camada final é uma convolução transposta com um único filtro 3×3 e ativação sigmoidal, responsável por gerar uma saída com dimensão $(256, 256, 1)$, normalizada entre 0 e 1, compatível com as imagens de entrada.

TABLE I: Arquitetura detalhada do Autoencoder proposto

Bloco	Camada	Filtros/Unidades	Saída	Observações
Encoder	Conv2D + BN + Dropout	16	(128, 128, 16)	Stride 2
Encoder	Conv2D + BN + Dropout	32	(64, 64, 32)	Stride 2
Encoder	Conv2D + BN + Dropout	64	(32, 32, 64)	Stride 2
Encoder	Conv2D + BN + Dropout	128	(16, 16, 128)	Stride 2
Encoder	Flatten	-	(32768,)	-
Encoder	Dense	128	(128,)	Vetor latente
Decoder	Dense + BN + Dropout	-	(16, 16, 128)	Reshape
Decoder	Conv2DTranspose + BN + Dropout	64	(32, 32, 64)	Stride 2
Decoder	Conv2DTranspose + BN + Dropout	32	(64, 64, 32)	Stride 2
Decoder	Conv2DTranspose + BN + Dropout	16	(128, 128, 16)	Stride 2
Decoder	Conv2DTranspose (sigmoid)	1	(256, 256, 1)	Stride 2

D. Função de Perda Híbrida

A função de perda usada combina dois termos importantes:

- **Erro Quadrático Médio (MSE):** Mede a diferença pontual média entre a imagem original e a reconstruída, incentivando fidelidade pixel a pixel.
- **Índice Estrutural de Similaridade (SSIM):** Avalia a similaridade considerando luminância, contraste e estrutura, promovendo preservação dos detalhes estruturais e texturas, essenciais para detectar pequenas anomalias.

A perda é dada por:

$$\mathcal{L}(x, \hat{x}) = 0.5 \times \text{MSE}(x, \hat{x}) + 0.5 \times (1 - \text{SSIM}(x, \hat{x}))$$

Este balanço foi escolhido empiricamente para maximizar a qualidade perceptual das reconstruções, fundamental para o sucesso na detecção de defeitos.

E. Configurações de Treinamento

O treinamento foi realizado utilizando o otimizador Adam com taxa de aprendizado 0.002, batch size de 8, durante 500 épocas. O uso de batch normalization em todas as camadas convolucionais auxilia na aceleração do treinamento e evita overfitting.

A função de perda híbrida foi monitorada para garantir convergência, e o modelo final foi salvo para uso em testes e análise.

IV. Experimentos

A. Avaliação Quantitativa

Para avaliar o desempenho, utilizamos o índice SSIM médio entre imagens originais e reconstruídas, comparando imagens normais e defeituosas. A Tabela II apresenta os resultados obtidos com diferentes combinações de pesos na função de perda: somente MSE, somente SSIM e uma combinação equilibrada dos dois.

Observa-se que as imagens com defeito apresentam SSIM significativamente menor em relação às imagens normais em todas as configurações, indicando que o modelo é sensível às anomalias. A melhor separação entre classes foi obtida com a combinação MSE + SSIM (0.5, 0.5), reforçando a eficácia da função de perda híbrida.

TABLE II: Média e desvio padrão do SSIM nas imagens de teste para diferentes pesos na função de perda

Categoria	MSE (1,0)	SSIM (0,1)	MSE + SSIM (0.5,0.5)
Imagens normais	0.8662 ± 0.0130	0.8461 ± 0.0153	**0.9055 ± 0.0161**
Imagens com defeito	0.8383 ± 0.0160	0.7938 ± 0.0270	**0.8553 ± 0.0251**

B. Visualização dos Resultados

A Figura 1 exemplifica imagens com e sem defeito, suas reconstruções e os mapas de erro, evidenciando a capacidade do modelo em localizar anomalias apenas nas imagens defeituosas.

V. Discussão

Os resultados demonstram que o uso combinado de MSE e SSIM na função de perda promove reconstruções que preservam detalhes estruturais e texturais relevantes para a identificação das anomalias. Como evidenciado na Tabela II, a configuração híbrida (0.5, 0.5) alcançou os maiores valores médios de SSIM tanto para imagens normais quanto para imagens com defeito, além da maior separação entre essas categorias. Isso indica que o modelo treinado com a função de perda híbrida foi mais eficaz em preservar padrões normais e destacar desvios anômalos, favorecendo a geração de mapas de erro mais informativos e localizados.

Em contraste, o uso isolado de MSE ou SSIM resultou em reconstruções com menor fidelidade perceptual e menor sensibilidade às anomalias, demonstrando que a sinergia entre as duas métricas é essencial para otimizar o desempenho do modelo em tarefas de inspeção visual.

Limitações do modelo incluem menor sensibilidade a defeitos muito sutis e a dependência da qualidade das imagens de entrada. Ruídos, variações de iluminação e diferentes condições de aquisição podem afetar o desempenho, o que exige cuidados no pré-processamento.

VI. Análise de Robustez e Limitações

Embora o modelo apresente desempenho robusto na detecção de defeitos visuais evidentes, sua capacidade de identificar anomalias muito sutis, como pequenas manchas ou fissuras pouco contrastadas, pode ser limitada. Isso ocorre porque tais defeitos podem gerar erro de reconstrução baixo, dificultando a distinção entre normalidade e anomalia.

Além disso, a abordagem atual depende fortemente da qualidade do pré-processamento e padronização das imagens. Condições adversas de iluminação, ruído e variações de pose podem comprometer a qualidade da reconstrução, levando a falsos positivos ou negativos.

Outro ponto crítico é o custo computacional. A arquitetura convolucional profunda e as operações de SSIM implicam em tempo considerável de treinamento e inferência, o que pode

ser um desafio para sistemas de inspeção em tempo real. Investigações futuras podem explorar otimizações e arquiteturas mais eficientes.

VII. Perspectivas Futuras

Para ampliar a aplicabilidade do método, propõe-se a integração com técnicas de atenção espacial, que destacam regiões relevantes da imagem, potencializando a detecção de defeitos localizados. Além disso, o uso de redes adversariais (GANs) para melhorar a qualidade da reconstrução pode aumentar a sensibilidade a anomalias sutis.

Explorar aprendizado multimodal, combinando dados visuais com informações de sensores industriais, também pode enriquecer o diagnóstico.

Repositório e Código

Todo o código e os experimentos descritos neste artigo estão disponíveis no GitHub:
https://github.com/ViniciusTavaresSousa/Deteccao-de-Anomalias-Visuais-com-Autoencoder-Convolucional-e-SSIM-no-Dataset-MVTec-AD

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," *Proceedings of the MLSDA*, 2014.

[5] Y. Xia, W. Liu, and W. Wang, "Learning discriminative reconstructions for unsupervised outlier removal," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3519–3531, 2015.

[6] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, 2015. [Online]. Available: https://arxiv.org/abs/1512.09300

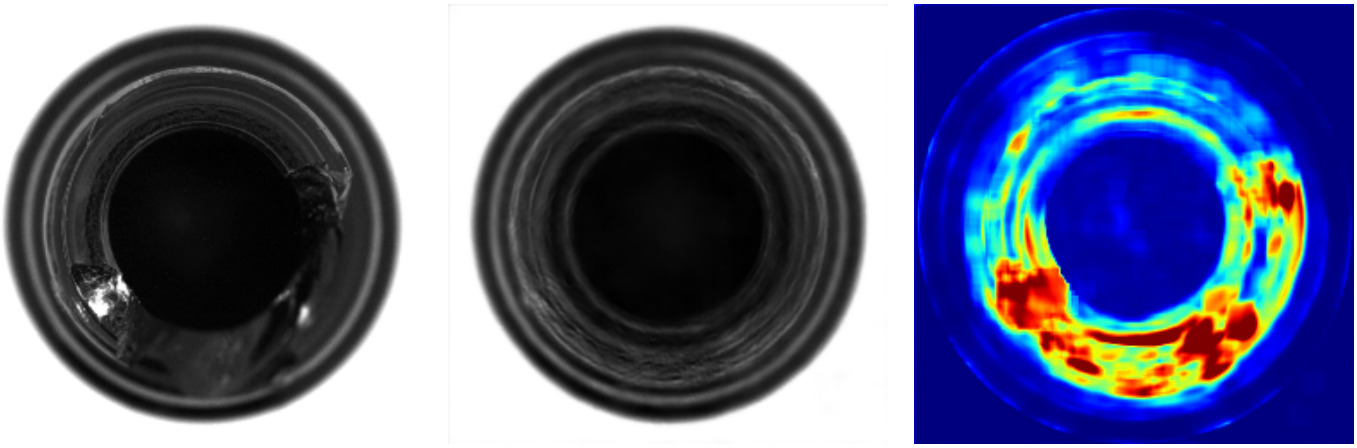
[7] L. Ruff, R. A. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, A. Binder, K.-R. Müller, and M. Kloft, "Deep one-class classification," *International Conference on Machine Learning (ICML)*, 2018. [Online]. Available: https://proceedings.mlr.press/v80/ruff18a.html

[8] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

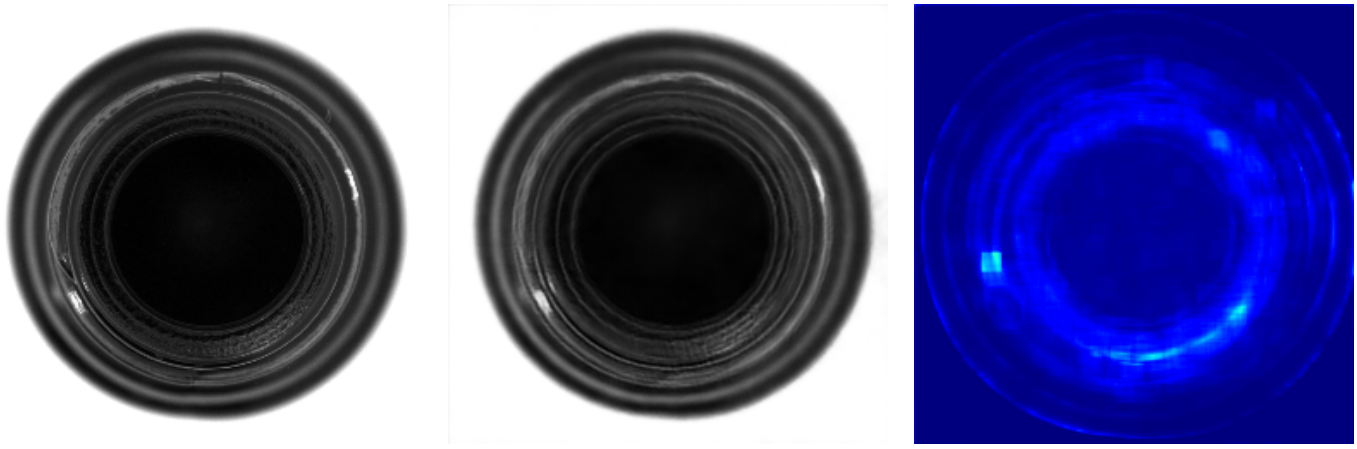
[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Information Processing in Medical Imaging (IPMI)*, 2017.

[11] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad - a comprehensive real-world dataset for unsupervised anomaly detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600, 2019.



(a) Imagem com defeito (b) Reconstrução da imagem com defeito (c) Mapa de erro (1-SSIM) com defeito



(d) Imagem normal (e) Reconstrução da imagem normal (f) Mapa de erro (1-SSIM) normal

Fig. 1: Exemplos de imagens originais, reconstruções e mapas de erro para imagens com e sem defeitos.