Abordagem de Super-Resolução Baseada em Autoencoder para Navegação de Robôs Aéreos

Jhonathan A. Oliveira[1] **e Felipe G. Oliveira**[2]

[1]Inst. de Computação (ICOMP) - Univ. Federal do Amazonas (UFAM)

[2]Inst. de Ciências Exatas e Tecnologia (ICET) - Univ. Federal do Amazonas (UFAM)

jholiveira@icomp.ufam.edu.br, felipeoliveira@ufam.edu.br

Resumo. *Imagens são amplamente utilizadas na navegação de veículos aéreos autônomos, mas seu processamento local pode exigir alta capacidade computacional e consumo energético. Como alternativa, este artigo propõe o método AESR (Autoencoder-based Super-Resolution) para viabilizar o processamento remoto, reconstruindo imagens de alta resolução a partir de versões de baixa resolução capturadas pelo veículo. A abordagem utiliza autoencoders com conexões de salto e módulos de autoatenção. Os resultados mostram que o AESR supera técnicas do estado da arte em métricas como PSNR, SSIM, LPIPS, DISTS, NIQE e BRISQUE, mesmo em cenários desafiadores.*

1. Introdução

Na robótica móvel, o processamento remoto de dados dos sensores é uma prática comum que reduz as exigências computacionais locais e o consumo de energia do robô, além de possibilitar, em alguns casos, a transmissão de um volume menor de dados, aliviando a carga na comunicação. As câmeras, amplamente utilizadas nesses sistemas, fornecem imagens fundamentais para tarefas como detecção de objetos, navegação e mapeamento. No entanto, a qualidade dessas imagens pode ser afetada por limitações do dispositivo de captura, transmissão ou compressão, comprometendo a operação autônoma em tempo real [Abdullah et al. 2021]. Nesse contexto, a super-resolução surge como uma abordagem promissora para reconstruir imagens de alta resolução a partir de versões de baixa resolução, aumentando a nitidez e o nível de detalhes. Essa técnica tem se mostrado eficaz em aplicações como navegação autônoma [Angarano et al. 2023] e sistemas de vigilância [Gonzalez et al. 2022], sendo especialmente útil em veículos aéreos não tripulados (VANTs) [Lin et al. 2023]. A Figura 1 ilustra a aplicação da abordagem de super-resolução proposta, destacando seu potencial para auxiliar na navegação autônoma de veículos aéreos por meio do processamento remoto de imagens.

Neste trabalho, propõe-se o AESR (Autoencoder-based Super-Resolution), uma abordagem de aprendizado profundo baseada em arquitetura de autoencoder para reconstrução de imagens de alta resolução a partir de dados de baixa resolução. O método foi desenvolvido para processar imagens aéreas, visando facilitar o processamento remoto na navegação de veículos aéreos autônomos. Para isso, o autoencoder foi integrado a conexões de salto (skip connections) e módulos de autoatenção (self-attention modules), aprimorando o processo de aprendizado e a capacidade de generalização do modelo.

As principais contribuições deste trabalho são apresentadas a seguir:

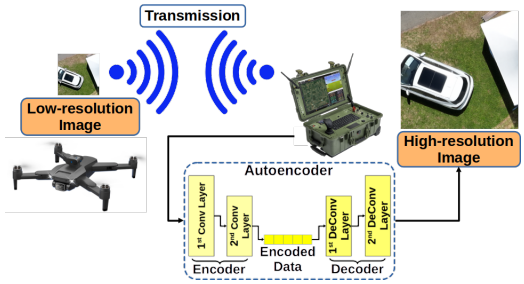


Figura 1: Exemplo de aplicação de uma técnica de super-resolução para apoiar a navegação de veículos aéreos.

- Uma abordagem inovadora de aprendizado para o problema de super-resolução, capaz de gerar imagens aéreas de alta qualidade a partir de dados de baixa resolução. A metodologia proposta extrai conhecimento a partir de um conjunto de dados desafiador, composto por diversas imagens aéreas, para estimar valores de pixels desconhecidos durante o processo de aumento de resolução;
- Combinação de conexões de salto e módulos de autoatenção em uma arquitetura de autoencoder, integrando a preservação de informações relevantes e a melhoria da capacidade do modelo de focar em regiões específicas da imagem.

2. Metodologia

Este artigo aborda o problema da super-resolução, reconstruindo imagens aéreas em alta resolução a partir de dados de baixa resolução [Oliveira et al. 2024]. A metodologia proposta será detalhada nas subseções seguintes.

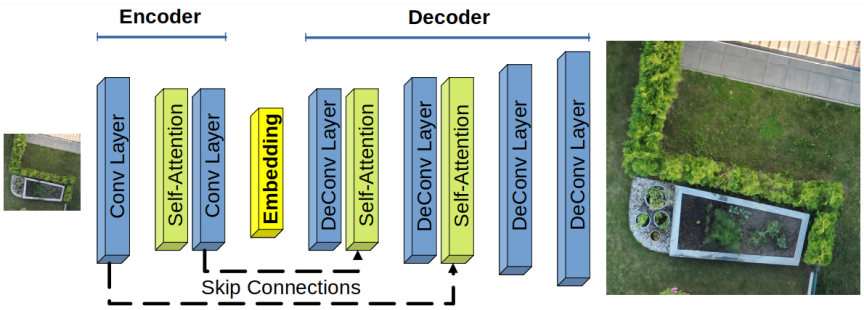


Figura 2: Visão geral da metodologia proposta para super-resolução de imagens aéreas.

2.1. Autoencoder para Super-Resolução

Autoencoders são um tipo de arquitetura de rede neural projetada para aprender representações eficientes dos dados. Esse processo permite que o modelo capture características essenciais enquanto elimina ruídos [Liu et al. 2021]. Um autoencoder é composto por dois principais componentes: uma função de codificação f e uma função de decodificação g. O codificador mapeia a entrada x para uma representação latente z em um espaço de menor dimensão, i.e., $z = f(x)$. O decodificador tenta reconstruir a entrada a partir dessa representação latente, produzindo $\hat{x} = g(z)$. O objetivo do treinamento é minimizar o erro de reconstrução, geralmente medido por uma função de perda. Neste trabalho, utilizou-se a função de perda baseada no Índice de Similaridade Estrutural (SSIM), que é mais adequada para preservar informações perceptuais e estruturais.

A função de perda SSIM mede a similaridade entre a entrada original x e a saída reconstruída $\hat{x}$ em termos de luminância, contraste e estrutura, aspectos cruciais para a percepção visual humana. A função de perda pode ser representada como:

$$L(x, \hat{x}) = 1 - \text{SSIM}(x, \hat{x}) \tag{1}$$

onde:

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{2}$$

Nesta equação, μ_x e $\mu_{\hat{x}}$ representam as médias de x e $\hat{x}$; σ_x^2 e $\sigma_{\hat{x}}^2$ são as variâncias; $\sigma_{x\hat{x}}$ é a covariância entre x e $\hat{x}$; e C_1 e C_2 são constantes utilizadas para estabilizar a divisão [Wang et al. 2004]. Ao utilizar a função de perda SSIM, o autoencoder é orientado a preservar a integridade estrutural e perceptual das imagens durante o processo de reconstrução.

O processo de codificação é realizado por meio de uma sequência de camadas convolucionais, funções de ativação e pooling, retendo apenas as informações mais relevantes. As camadas convolucionais aplicam filtros aprendidos à imagem de entrada, capturando progressivamente características de nível mais elevado enquanto reduzem as dimensões espaciais. A arquitetura do codificador proposta é composta por duas camadas convolucionais, com 64 e 128 filtros na primeira e segunda camada, respectivamente. Essas camadas são seguidas pela função de ativação ReLU e pelo Max Pooling de tamanho (2×2), resultando na representação de embedding.

O processo de decodificação, por sua vez, tem como objetivo reconstruir a imagem de entrada a partir da representação de embedding z. Esse processo é realizado por meio de camadas de deconvolução, também conhecidas como camadas convolucionais transpostas, que executam o upsampling progressivo dos mapas de características até o espaço original da entrada. A arquitetura do decodificador proposta é composta por quatro camadas de deconvolução, com 128, 64, 32 e 16 filtros, respectivamente, da primeira à quarta camada. Essas camadas são seguidas pela função de ativação ReLU e pelo Max Pooling de tamanho (2×2), resultando na imagem aérea predita em alta resolução.

2.1.1. Conexões de Salto

As conexões de salto conectam diretamente a saída de camadas iniciais a camadas posteriores, mitigando o problema do desaparecimento do gradiente e preservando detalhes finos ao possibilitar que a rede contorne certas transformações. Isso permite que informações espaciais cruciais não são perdidas durante o processo de codificação e podem ser diretamente utilizadas durante a reconstrução. Na arquitetura de autoencoder proposta, a primeira conexão de salto conecta a primeira camada do codificador à primeira camada do decodificador, após a aplicação do módulo de autoatenção. A segunda conexão de salto conecta a segunda camada do codificador à segunda camada do decodificador, também após a aplicação do módulo de autoatenção, conforme ilustrado na Figura 2.

2.1.2. Módulos de Autoatenção

Os módulos de autoatenção permitem que a rede concentre-se em regiões importantes da imagem ao ponderar dinamicamente diferentes partes da entrada, facilitando a captura de dependências locais e globais dentro da imagem. Na arquitetura de autoencoder proposta, o módulo de autoatenção é composto por duas camadas convolucionais com 64 e 128 filtros, na primeira e segunda camada, respectivamente. Os filtros possuem tamanho $1x1$ e são seguidos por normalização em batch (batch normalization) e função de ativação ReLU. Os módulos de autoatenção são aplicados após a primeira camada do codificador, após a primeira camada do decodificador e após a segunda camada do decodificador (conforme ilustrado na Figura 2).

3. Experimentos

3.1. Ambiente Experimental

A abordagem proposta utiliza os frameworks OpenCV e TensorFlow em um computador Dell, equipado com processador Intel® Xeon^{TM} Silver 4114 de 2,20 GHz, 128 GB de memória DDR4-2133 e uma placa NVIDIA® GeForce® RTX A4000 com 16 GB de memória GDDR6. Na etapa de treinamento, foi utilizado o framework Grid Search para ajuste dos hiperparâmetros, incluindo o número de filtros, o tamanho dos filtros, o tamanho do lote (batch size), o número de épocas (epochs), o algoritmo de otimização e a taxa de aprendizado, com o objetivo de alcançar alta precisão. Como resultado desse ajuste, o tamanho do lote foi definido como 16, o número de épocas como 50 e a taxa de aprendizado como 0,0001.

3.2. Conjunto de Dados

Nos experimentos, utilizou-se o desafiador conjunto de dados DSR (Drone Super-Resolution) [Lin et al. 2023] para validar a metodologia proposta de super-resolução no domínio aéreo. Esse conjunto de dados é composto por 2132 imagens aéreas, sendo 1066 imagens de baixa resolução e 1066 imagens correspondentes de alta resolução. As imagens de baixa resolução possuem dimensão de 180×180, enquanto as imagens de alta resolução têm dimensão de 720×720, ou seja, são $4\times$ maiores do que as imagens de baixa resolução.

3.3. Métricas de Qualidade da Imagem

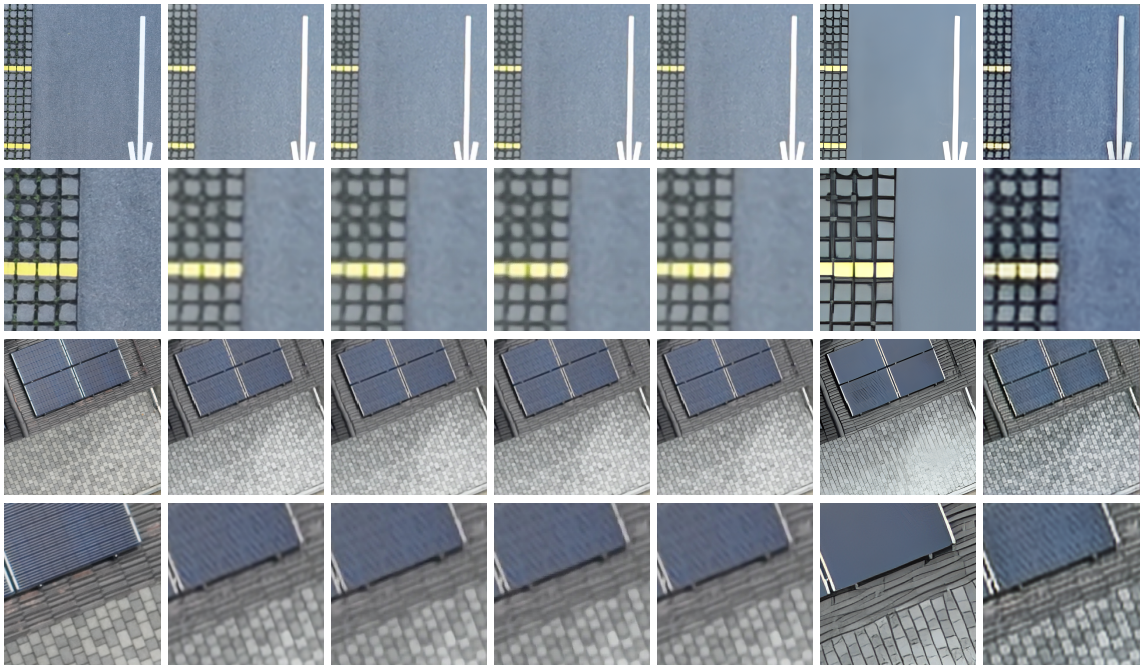
Na análise quantitativa, foram avaliados os resultados obtidos pelo método proposto (AESR) e pelas técnicas da literatura utilizadas como comparação. Na análise com referência, em que a imagem predita é comparada com a imagem de referência, foram utilizados os indicadores PSNR [Pratt 1978], SSIM [Wang et al. 2004], LPIPS [Zhang et al. 2018] e DISTS [Ding et al. 2022]. Já na análise sem referência, em que a imagem predita não é comparada com imagens de referência, foram utilizadas as métricas NIQE [Mittal et al. 2013] e BRISQUE [Bosse et al. 2018].

3.4. Avaliação de Desempenho

3.4.1. Análise Qualitativa

A análise qualitativa envolve a comparação do método proposto de Super-Resolução (AESR) com técnicas consolidadas da literatura. Os experimentos consideram o aumento

da dimensão das imagens aéreas em $4\times$ em relação às imagens de baixa resolução. A qualidade visual das diferentes técnicas no conjunto de dados DSR é analisada, com ênfase nas diferenças em cor, contraste, nitidez e qualidade visual geral. Essa abordagem comparativa oferece insights sobre as vantagens e limitações da técnica proposta, constituindo um passo fundamental para validar sua capacidade de gerar imagens de alta resolução a partir de dados de baixa resolução, especialmente em aplicações aéreas.



(a) Raw (b) EDSR (c) ESPCN (d) FSRCNN (e) LapSRN (f) Real-ESRGAN (g) AESR (Proposta)

Figura 3: Comparação qualitativa das imagens de super-resolução no conjunto de dados DSR. Da esquerda para a direita, são apresentadas as imagens aéreas originais (raw) e os resultados dos métodos EDSR, ESPCN, FSRCNN, LapSRN, Real-ESRGAN e da abordagem AESR proposta. A primeira e a terceira linhas representam imagens aéreas do conjunto de dados, enquanto a segunda e a quarta linhas mostram regiões extraídas das imagens correspondentes.

A Figura 3 apresenta as imagens super-resolvidas após a aplicação dos algoritmos de reconstrução de alta resolução, considerando o conjunto de dados DSR. As diferentes linhas mostram cenários distintos, com características desafiadoras para reconstrução, destacando detalhes em cenas variadas. As colunas exibem, da esquerda para a direita, a imagem original, os métodos comparativos da literatura e o método proposto (AESR). Diferentemente dos métodos existentes, o método proposto destaca-se no aumento da resolução de imagens aéreas de baixa qualidade, evitando borramento, contraste excessivo, distorção de cor e recuperando detalhes originais. EDSR, ESPCN, FSRCNN e LapSRN apresentaram reconstruções borradas, com baixa fidelidade nos detalhes. Por outro lado, o Real-ESRGAN mostrou boa qualidade na nitidez e redução do borramento. Contudo, vale destacar que, no Real-ESRGAN, vários detalhes da cena real são perdidos devido ao uso de um filtro passa-baixa para mitigar o efeito de borramento após a reconstrução. O método AESR preservou detalhes nas imagens geradas, apresentando qualidade em cor e contraste e reduzindo os efeitos de borramento nas imagens aéreas reconstruídas.

3.4.2. Análise Quantitativa

A análise quantitativa avalia os resultados gerados pelo método AESR proposto e os compara com técnicas de referência na área, incluindo EDSR[Lim et al. 2017], ESPCN[Shi et al. 2016], FSRCNN[Dong et al. 2016], LapSRN [Lai et al. 2017] e Real-ESRGAN [Wang et al. 2021]. As medições quantitativas são utilizadas para determinar a precisão do processo de super-resolução. Para garantir uma avaliação rigorosa, foram empregadas métricas de qualidade com e sem referência. PSNR, SSIM, LPIPS (AlexNet, VGG e SqueezeNet) e DISTS compõem as métricas de análise com referência completa, enquanto NIQE e BRISQUE avaliam a qualidade sem referência.

Tabela 1: Avaliação da qualidade das imagens, com e sem referência, utilizando as métricas de qualidade PSNR, SSIM, LPIPS (AlexNet, VGG e SqueezeNet), DISTS, NIQE e BRISQUE no conjunto de dados DSR.

Métricas	EDSR (CVPR 2017)	ESPCN (CVPR 2016)	FSRCNN (ECCV 2016)	LapSRN (CVPR 2017)	Real-ESRGAN (CVPR 2021)	AESR (Proposta)
PSNR ↑	20.152	20.125	20.101	20.126	18.556	**21.005**
SSIM ↑	0.724	0.721	0.718	0.722	0.646	**0.758**
LPIPS$_{Alex}$ ↓	0.608	0.561	0.544	0.579	0.421	**0.402**
LPIPS$_{VGG}$ ↓	0.520	0.528	0.543	0.536	0.491	**0.443**
LPIPS$_{Squeeze}$ ↓	0.497	0.448	0.405	0.459	0.263	**0.228**
DISTS ↓	0.233	0.235	0.233	0.237	0.207	**0.189**
NIQE ↓	0.951	0.954	0.952	0.952	0.901	**0.826**
BRISQUE ↓	18.947	18.949	18.968	18.980	19.022	**18.845**

A Tabela 1 apresenta a comparação entre o método proposto e vários algoritmos reconhecidos na literatura, avaliados por métricas de qualidade de imagem com e sem referência. Os resultados mostram que o método proposto supera os demais métodos, mesmo quando testado em um conjunto de dados desafiador de imagens aéreas e considerando diferentes métricas de qualidade. Entre os métodos concorrentes, Real-ESRGAN e EDSR apresentaram o melhor desempenho. O Real-ESRGAN destacou-se em precisão nas métricas LPIPS (AlexNet, VGG e SqueezeNet), DISTS, NIQE e BRISQUE. Em contrapartida, o EDSR alcançou alta precisão nas métricas tradicionais PSNR e SSIM. Embora o AESR tenha apresentado alta precisão em todas as métricas de qualidade, demonstra sua robustez e eficiência para aplicações aéreas.

4. Conclusões

Este artigo aborda o problema da super-resolução, reconstruindo imagens aéreas de alta resolução a partir de dados de baixa resolução. Dessa forma, a abordagem proposta viabiliza a transmissão de imagens de baixa resolução capturadas por veículos aéreos para processamento remoto, o que contribui para a melhoria da navegação e da tomada de decisão. O método proposto alcançou alta precisão no DSR, um conjunto de dados desafiador de imagens aéreas para super-resolução. Os resultados indicam que o processo de super-resolução proposto apresenta robustez mesmo em cenas complexas e sob diversas condições, oferecendo alta precisão em diferentes cenários.

Referências

Abdullah, Q., Shah, N. S. M., Mohamad, M., Ali, M. H. K., Farah, N., Salh, A., Aboali, M., Mohamad, M. A. H., and Saif, A. (2021). Real-time autonomous robot for object tracking using vision system. *CoRR*.

Angarano, S., Salvetti, F., Martini, M., and Chiaberge, M. (2023). Generative adversarial super-resolution at the edge with knowledge distillation. *Engineering App. of Artificial Intelligence*, 123:106407.

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2018). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. on Im. Proc.*, 27(1):206–219.

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2022). Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. on Pat. Analysis and Machine Intel.*, 44(5):2567–2581.

Dong, C., Loy, C. C., and Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 391–407, Cham. Springer International Publishing.

Gonzalez, D., Patricio, M. A., Berlanga, A., and Molina, J. M. (2022). A super-resolution enhancement of uav images based on a convolutional neural network for mobile devices. *Personal and Ubiquitous Computing*, 26:1193–1204.

Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140.

Lin, X., Ozaydin, B., Vidit, V., El Helou, M., and Süsstrunk, S. (2023). Dsr: Towards drone image super-resolution. In Karlinsky, L., Michaeli, T., and Nishino, K., editors, *Computer Vision – ECCV 2022 Workshops*, pages 361–377, Cham. Springer Nature Switzerland.

Liu, Z.-S., Siu, W.-C., and Chan, Y.-L. (2021). Photo-realistic image super-resolution via variational autoencoders. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1351–1365.

Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

Oliveira, J. A., Drews, P. L. J., and Oliveira, F. G. (2024). Autoencoder-based super-resolution approach for aerial robot navigation. In *2024 Brazilian Symposium on Robotics (SBR) and 2024 Workshop on Robotics in Education (WRE)*, pages 85–90.

Pratt, W. K. (1978). *Digital Image Processing*. John Wiley & Sons, Nashville, TN.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE CVPR*, pages 1874–1883.

Wang, X., Xie, L., Dong, C., and Shan, Y. (2021). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.