Preenchimento de Lacunas (Gap-Filling) em DTMs HiRISE usando Inferência Monocular com Vision Transformers (ViT)

Bruno R. B. F. Holanda[1]

[1]Instituto do Hardware BR – (HBR)
Av. Alan Turing, n°: 776 – 13.083-898 – Campinas – SP – Brasil

brunorodriguesholanda@gmail.com

Abstract. *numero de palavras 150 - 250*

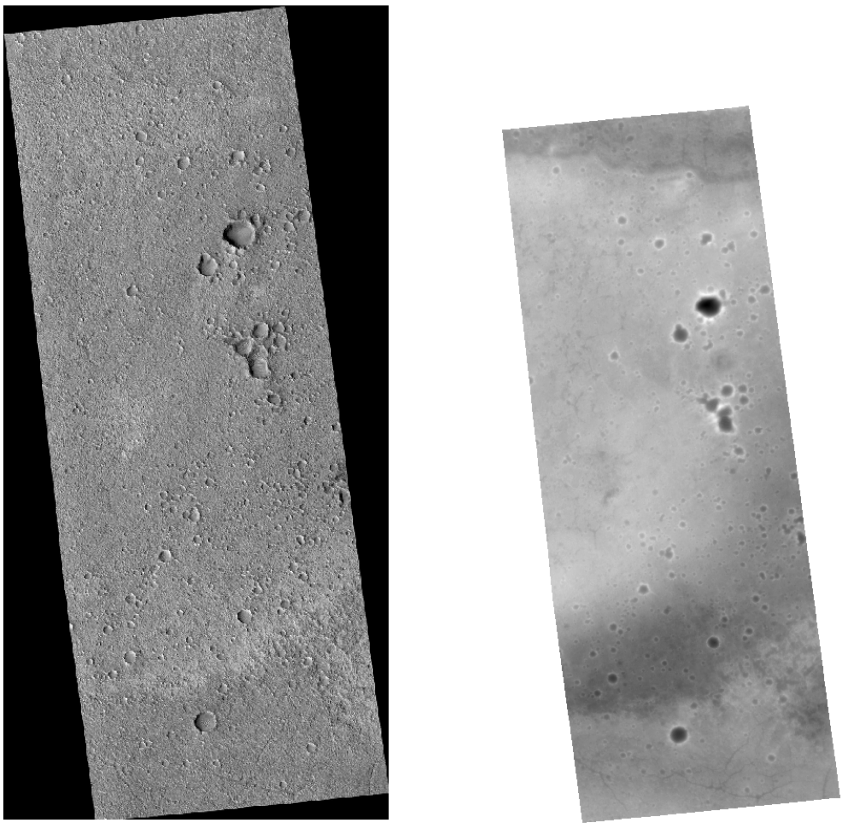
Resumo. *Numero de palavras: 150 - 250 - igual ao abstract porém em pt-BR Espaço Visual (Páginas): 0.5 Foco Principal: Venda o peixe: Problema + Solução (ViT) + Melhor Resultado.*

1. Introdução

A exploração da superfície de Marte atingiu um novo patamar de detalhamento com a câmera *High Resolution Imaging Science Experiment* (HiRISE), a bordo da sonda *Mars Reconnaissance Orbiter* (MRO). Os produtos gerados por este instrumento, especificamente os Modelos Digitais de Terreno (DTMs), oferecem uma resolução espacial na ordem de metros por pixel, constituindo um recurso inestimável para a comunidade científica. Estes dados são fundamentais para estudos geomorfológicos de precisão, modelagem de processos de superfície e, crucialmente, para o planejamento estratégico e segurança de missões robóticas, como o rover Perseverance (MCEWEN et al., 2007).

A metodologia predominante para a geração destes modelos topográficos é a fotogrametria estéreo, que reconstrói a estrutura 3D da superfície através da identificação de pontos homólogos em pares de imagens orbitais (KIRK et al., 2011). Apesar de sua robustez geométrica, esta técnica possui limitações intrínsecas severas. O processo de correspondência de pixels (*stereo matching*) falha sistematicamente em condições adversas, notadamente em regiões de textura homogênea — como vastos campos de dunas — ou em áreas de iluminação extrema, caracterizadas por sombras profundas ou reflexos especulares em gelo. O resultado direto dessas falhas é a presença massiva de lacunas de dados (pixels com valor *NoData*) nos produtos finais.

A descontinuidade causada por estas lacunas compromete a integridade dos dados, impedindo análises espaciais contínuas essenciais, como simulações de fluxo hidrológico e cálculos de declividade. Tradicionalmente, a mitigação deste problema recorre a métodos de interpolação matemática, como a bilinear ou *splines*. No entanto, tais abordagens tendem a gerar superfícies artificialmente suaves, ignorando a morfologia local e resultando em artefatos geologicamente implausíveis. A Figura 1 ilustra este cenário, contrastando a riqueza de textura visual disponível na ortoimagem com a ausência de informação altimétrica no DTM correspondente.



(a) Ortoimagem HiRISE (Textura) **(b) DTM com Lacunas (Preto = NoData)**

Figura 1. Exemplo do problema abordado: (a) A imagem visual contém informações completas de textura e sombreamento, enquanto (b) o DTM derivado via estéreo apresenta falhas significativas (áreas pretas) onde a correlação falhou.

Para superar essas limitações, este trabalho propõe uma abordagem baseada em Aprendizado Profundo (*Deep Learning*), especificamente no campo da Estimativa de Profundidade Monocular (MDE). A premissa central é que a ortoimagem monocular (Figura 1a), que permanece íntegra mesmo onde a fotogrametria falha, contém pistas visuais suficientes — através de sombreamento (*Shape-from-Shading*) e textura — para inferir a topografia subjacente.

Embora trabalhos seminais como o MADNet (TAO et al., 2021) tenham demonstrado a viabilidade do uso de Redes Adversariais Generativas (GANs) baseadas em convoluções (U-Net) para esta tarefa, propõe-se o uso de uma arquitetura mais recente e promissora: os *Vision Transformers* (ViT). Modelos como o DPT (*Dense Prediction Transformer*) (RANFTL et al., 2021) utilizam mecanismos de auto-atenção global, permitindo capturar relações de longo alcance na imagem que são frequentemente perdidas por janelas de convolução limitadas. A hipótese deste estudo é que um modelo ViT-DPT pode aprender a complexa função de transferência entre o albedo marciano e sua topografia de forma mais coesa que as abordagens anteriores.

Portanto, os objetivos deste trabalho são: (1) consolidar um conjunto de dados curado de pares Ortoimagem-DTM contendo apenas dados válidos para treinamento su-

pervisionado; (2) treinar uma arquitetura ViT-DPT para prever a topografia a partir da informação visual; e (3) aplicar o modelo treinado para preencher seletivamente as lacunas *NoData* em produtos oficiais, gerando DTMs híbridos ”prontos para análise”. Esta pesquisa visa entregar uma solução automatizada que aumente significativamente a usabilidade científica do vasto arquivo de dados da missão MRO.

2. Metodologia

A metodologia estruturou-se em três etapas sequenciais: a curadoria de um conjunto de dados livre de falhas para treinamento supervisionado, a adaptação e treinamento da arquitetura *Vision Transformer* (ViT), e o desenvolvimento de um pipeline de inferência com pós-processamento para a fusão topográfica.

2.1. Aquisição e Preparação de Dados

Para viabilizar o aprendizado supervisionado, consolidou-se um conjunto de dados a partir do *Planetary Data System* (PDS). Pares de Modelos Digitais de Terreno (DTMs) e Ortoimagens das missões PSP e ESP foram indexados e processados.

Inicialmente, cada DTM foi reprojetado e alinhado pixel a pixel com sua respectiva ortoimagem utilizando a biblioteca GDAL, aplicando reamostragem cúbica para minimizar artefatos. Em seguida, implementou-se uma estratégia de recorte (*tiling*) com janela deslizante de 512×512 pixels e passo (*stride*) de 256 pixels (Figura 2).

Um filtro de qualidade rigoroso foi aplicado: apenas recortes contendo 100% de pixels válidos de elevação foram mantidos, descartando-se qualquer amostra com valores *NoData*. Isso garantiu que o modelo aprendesse exclusivamente com a ”verdade terrestre”fotogramétrica. Os dados resultantes foram normalizados para o intervalo $[0, 1]$ e serializados em formato *Parquet* para otimização de I/O, divididos em conjuntos de treino (80%) e validação (20%).

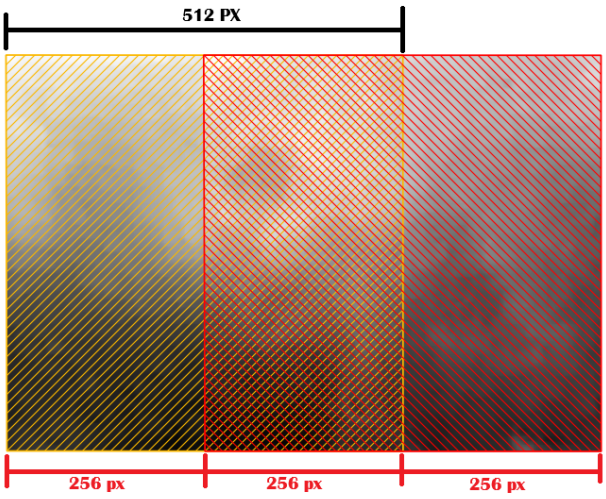


Figura 2. Estratégia de extração: Janela deslizante de 512px com stride de 256px. Apenas blocos íntegros (amarelo) foram utilizados, descartando-se falhas (vermelho).

2.2. Arquitetura de Rede e Treinamento

Adotou-se a arquitetura *Dense Prediction Transformer* (DPT) (Figura 3), utilizando o modelo `Intel/dpt-large` pré-treinado no dataset MiDaS. Esta abordagem substitui redes convolucionais puras por mecanismos de atenção global, permitindo correlacionar texturas locais com o contexto geomorfológico amplo.

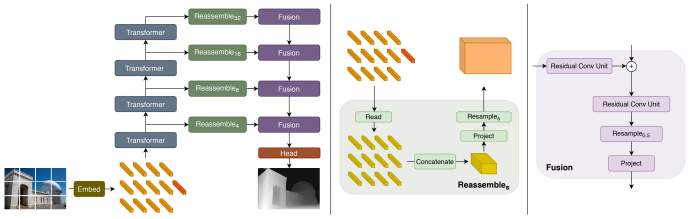


Figura 3. Arquitetura DPT: O codificador Transformer extrai representações em múltiplas escalas, refinadas por um decodificador convolucional para predição densa. Adaptado de Ranftl et al. (2021).

O treinamento foi orquestrado via *PyTorch* com *Distributed Data Parallel* (DDP). Utilizou-se o otimizador **AdamW** (decaimento de peso de 0.01) e precisão mista (AMP). Para evitar *overfitting*, aplicou-se *Early Stopping* com paciência de 5 épocas monitorando a perda de validação.

Para capturar a complexidade topográfica, definiu-se uma Função de Perda Combinada (L_{total}), conforme a Equação 1:

$$L_{total} = 0.6 \cdot \mathcal{L}_{L1} + 0.3 \cdot \mathcal{L}_{grad} + 0.1 \cdot \mathcal{L}_{SSIM} \tag{1}$$

Esta função pondera o erro absoluto ($\mathcal{L}_{L1}$), a consistência de bordas e inclinações através de operadores de Sobel ($\mathcal{L}_{grad}$), e a similaridade estrutural perceptual ($\mathcal{L}_{SSIM}$).

2.3. Pipeline de Inferência e Fusão

O preenchimento das lacunas não se resumiu à inferência direta. Desenvolveu-se um pipeline robusto para garantir consistência altimétrica e visual.

2.3.1. Inferência Contextual e Denormalização

Para mitigar efeitos de borda, adotou-se uma inferência com contexto expandido. Embora o bloco de predição final fosse de 512×512 pixels, a entrada da rede recebia uma área de 768×768 pixels (com margem de 128px).

Como o modelo estima profundidade relativa normalizada (Z_{pred}), foi necessário convertê-la para a altitude absoluta marciana (Z_{final}). Aplicou-se um alinhamento estatístico linear (Equação 2) baseado na média (μ) e desvio padrão (σ) dos pixels válidos na interseção entre a predição e o DTM original:

$$Z_{final} = Z_{pred} \cdot \left(\frac{\sigma_{real}}{\sigma_{pred} + \epsilon} \right) + \left(\mu_{real} - \mu_{pred} \cdot \frac{\sigma_{real}}{\sigma_{pred} + \epsilon} \right) \tag{2}$$

2.3.2. Fusão de Bordas (Blending)

Para eliminar descontinuidades na junção entre o pixel real e o sintético, aplicou-se um algoritmo de fusão. Uma máscara das lacunas foi dilatada e suavizada por filtro Gaussiano ($\sigma = 2.0$), criando uma zona de transição gradual (*feathering*) que interpolou os valores, garantindo continuidade topográfica.

2.4. Protocolo de Validação

Devido à inexistência de dados reais nas áreas de lacuna, a validação utilizou o método de ”Lacunas Sintéticas”. Máscaras de falha artificiais foram introduzidas em DTMs de teste originalmente perfeitos. As métricas RMSE (Erro Quadrático Médio), MAE (Erro Absoluto Médio) e SSIM foram calculadas comparando-se a reconstrução da IA com o dado original oculto.

3. Resultados

Numero de palavras: 400 - 600 Espaço Visual (Páginas): 1.5 a 2.0 Foco Principal: Crucial: Menos texto, mais tabelas de métricas e figuras comparativas.

4. Análise e Discução

Numero de palavras: 700 - 900 Espaço Visual (Páginas): 1.5 a 2.0 Foco Principal: Crucial: O ”porquê”dos resultados. Comparação com estado da arte.

5. Conclusão

Numero de palavras: 200 - 300 Espaço Visual (Páginas): 0.5 Foco Principal: Crucial: Retomada dos objetivos, limitações e trabalhos futuros.