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 supervisionado; (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%).

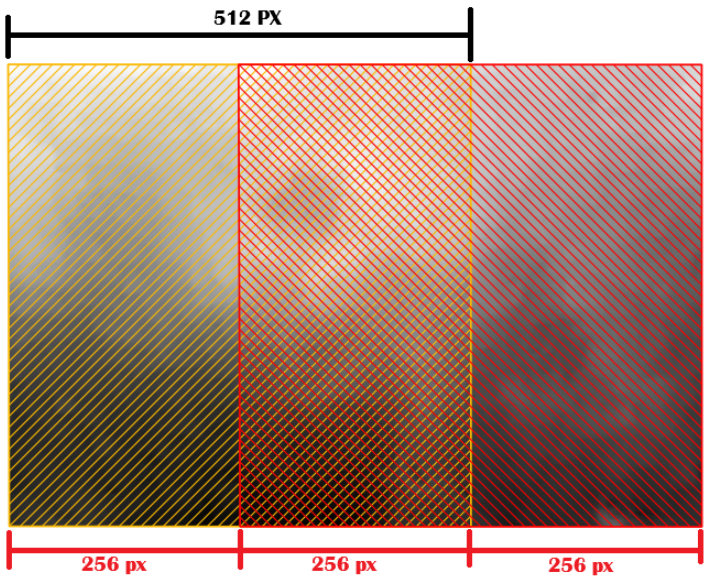


Figura 2. Estratégia de extração e recorte.

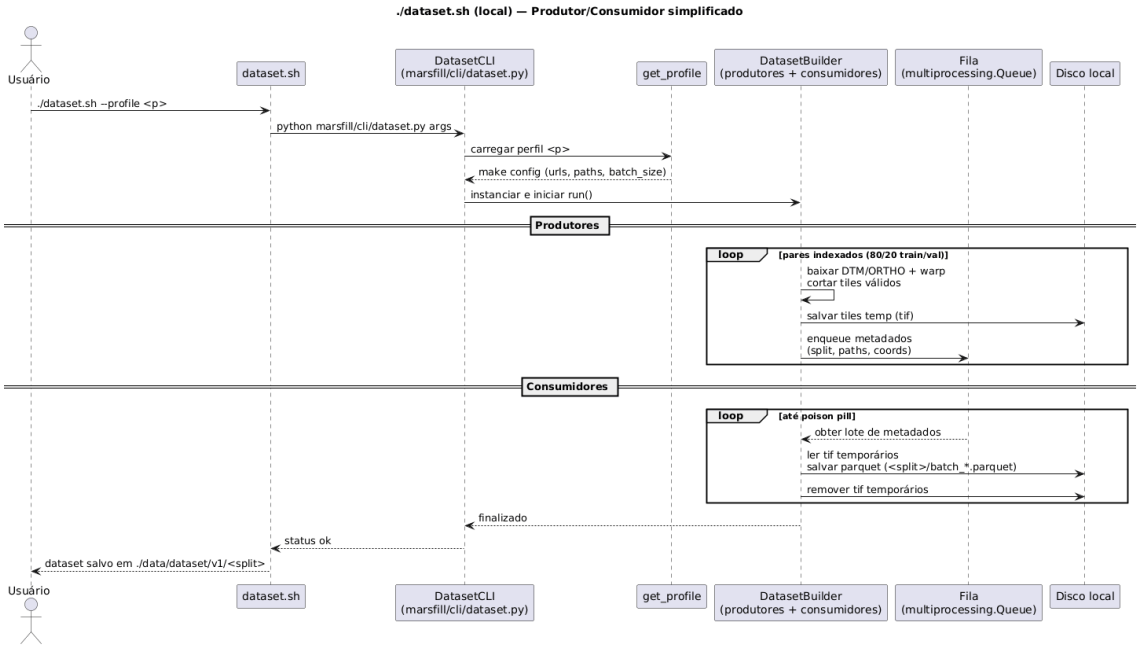


Figura 3. Fluxo detalhado de aquisição, limpeza e preparação dos dados (Dataset Pipeline).

2.2. Arquitetura de Rede e Treinamento

Adotou-se a arquitetura *Dense Prediction Transformer* (DPT) (Figura 4), 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.

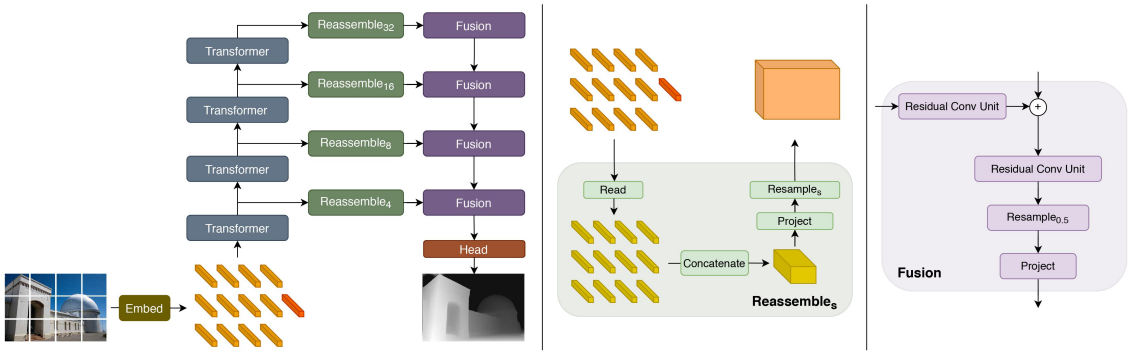


Figura 4. Arquitetura DPT. Adaptado de [Ranftl et al. 2021].

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:

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

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 ($\mathcal{L}_{SSIM}$).

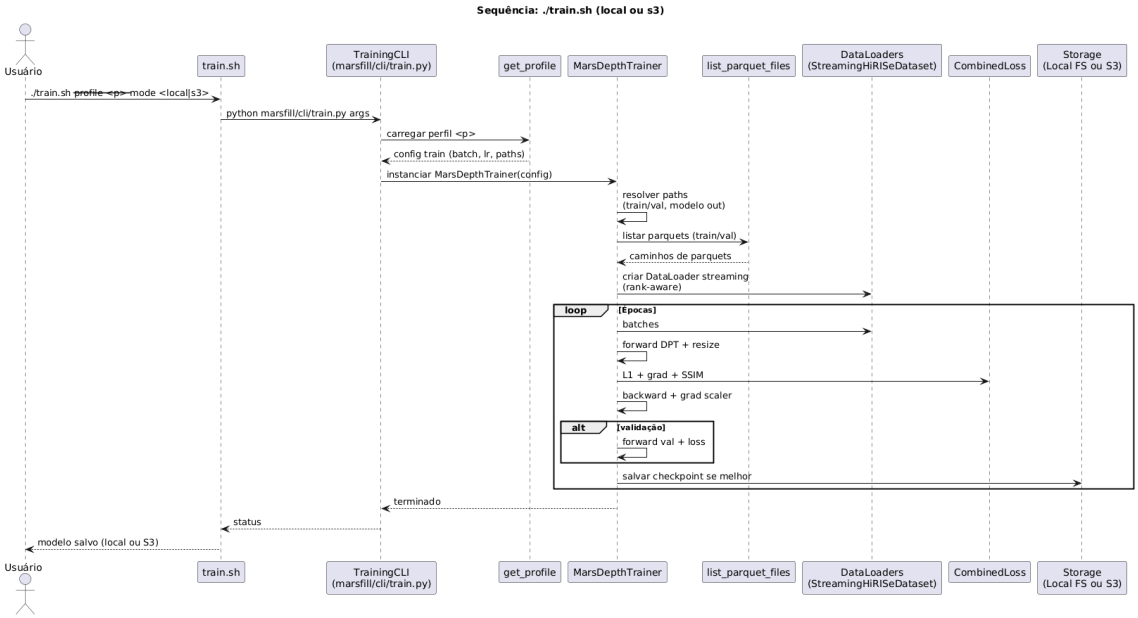


Figura 5. Sequência de operações durante uma época de treinamento supervisionado.

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:

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

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.

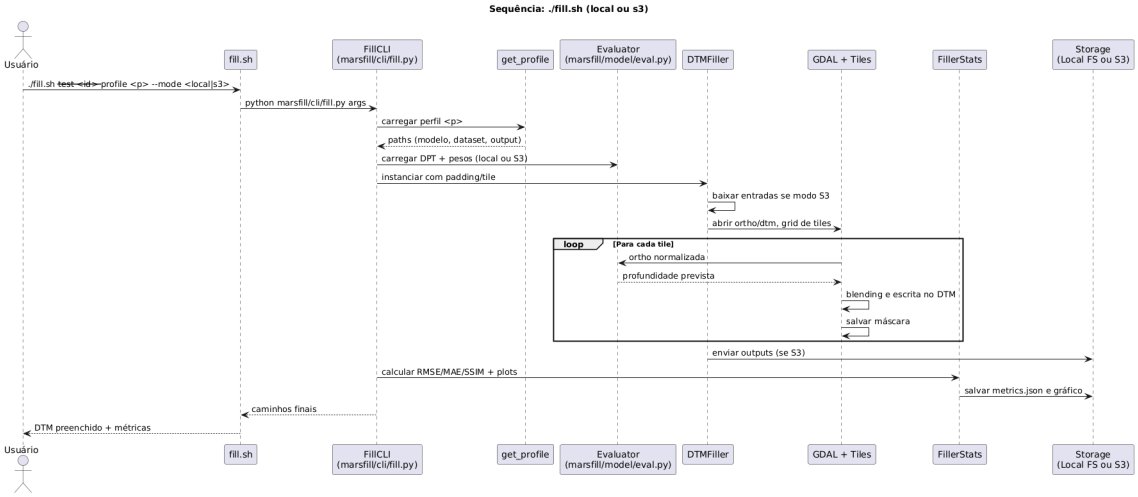


Figura 6. Pipeline de inferência: da imagem bruta à fusão topográfica.

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

O modelo foi treinado por 7 épocas em um ambiente distribuído (DDP) com 4 GPUs. A avaliação considerou tanto a dinâmica de aprendizado quanto a precisão final em dados de teste nunca vistos pelo modelo.

3.1. Dinâmica de Treinamento e Convergência

A análise das curvas de aprendizado demonstra uma convergência robusta. A Figura 7 ilustra o decaimento da função de perda total ao longo das iterações (batches). Observa-se uma redução exponencial inicial, seguida de uma estabilização assintótica, indicando que o modelo aprendeu efetivamente a mapear as características visuais em valores de profundidade.

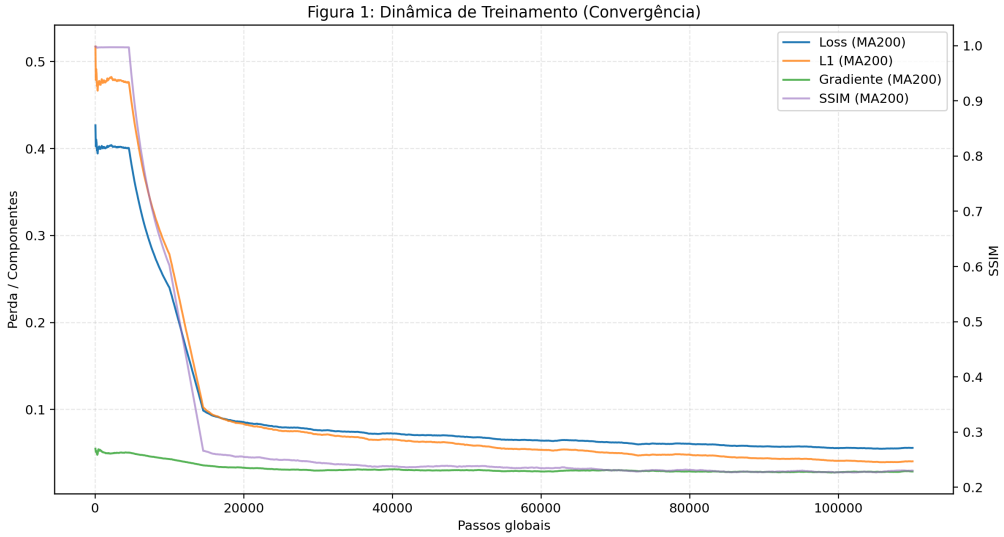


Figura 7. Evolução da Perda Total (Loss) por batch. A linha suave indica a tendência média, mostrando aprendizado contínuo sem instabilidades severas.

Para verificar a capacidade de generalização e ausência de *overfitting*, monitorou-se o erro nos conjuntos de treino e validação (Figura 8). A perda de validação acompanhou a tendência de queda da perda de treino, estabilizando-se em torno de 0.12, o que valida a capacidade do modelo de inferir topografia em imagens novas.

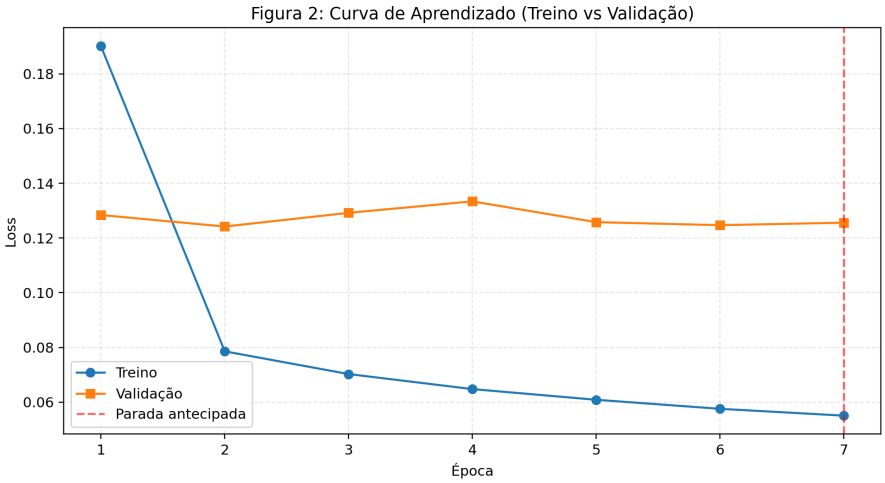


Figura 8. Curvas de perda média por época: Treino vs. Validação. A proximidade das curvas indica boa generalização.

A Figura 9 detalha o comportamento das componentes da função de custo. A estabilização das perdas de Gradiente e SSIM confirma que a rede não apenas minimizou o erro métrico, mas também aprendeu a reproduzir a alta frequência espacial (textura e bordas) do terreno marciano.

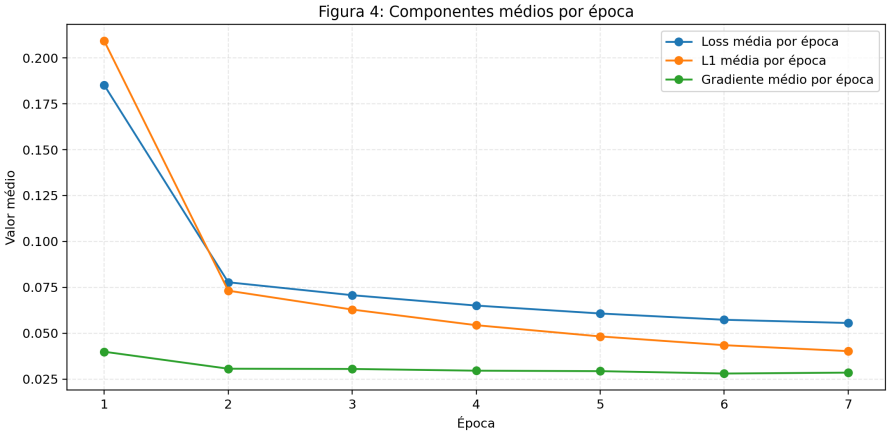


Figura 9. Contribuição média das componentes de perda (L1, Gradiente, SSIM) ao longo das épocas.

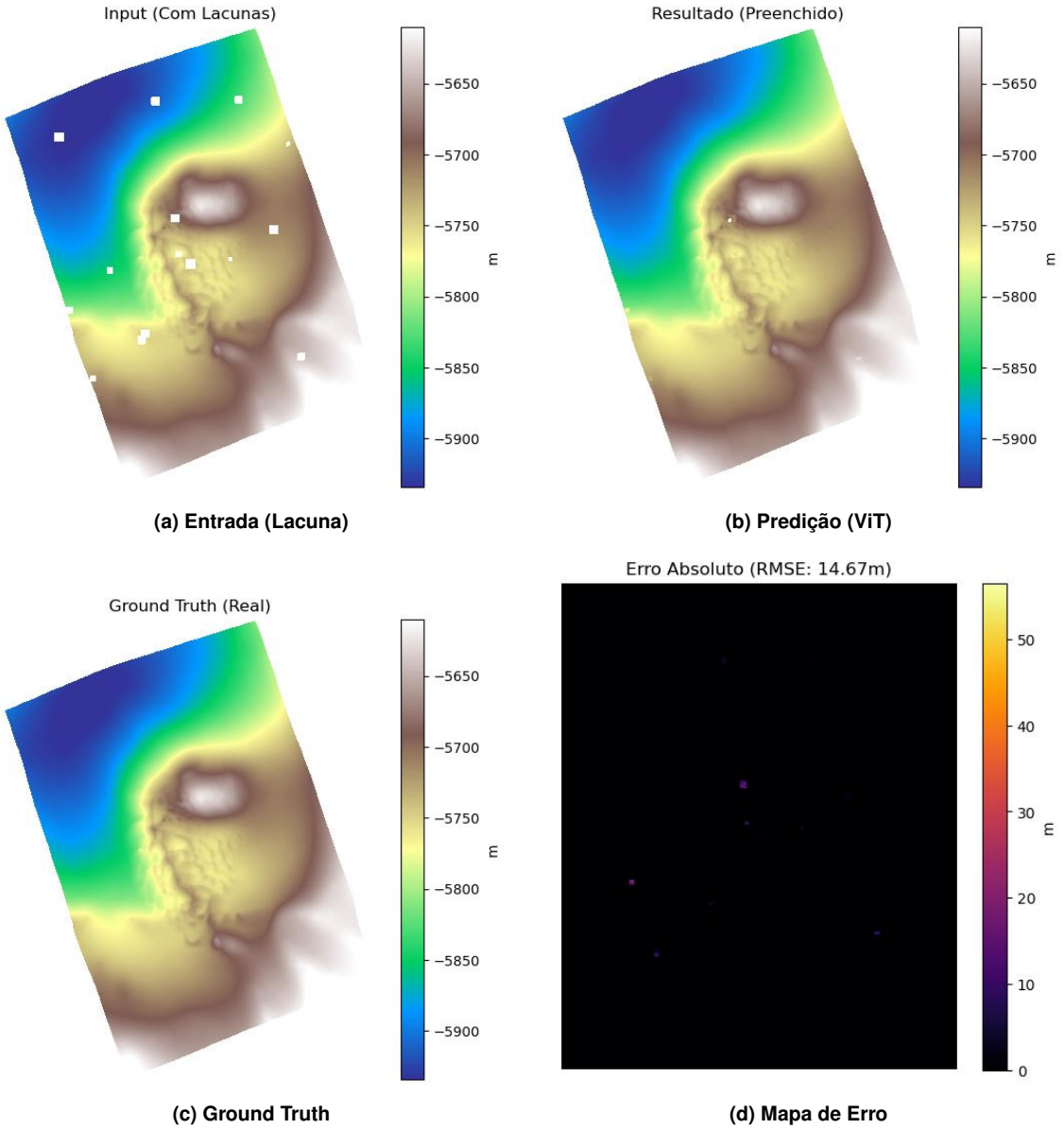
3.2. Avaliação em Dados de Teste

O modelo treinado foi aplicado a três classes geomorfológicas distintas. A Tabela 1 apresenta as métricas quantitativas obtidas.

Tabela 1. Métricas de desempenho em dados de teste (Ground Truth).

Terreno	RMSE (m)	SSIM	Tempo (s)
Dunas (Dunes)	14.67	0.9996	1.79
Planícies (Plains)	3.33	0.9998	2.52
Escarpas (Scarps)	15.21	0.9998	1.79
Média	**11.07**	**0.9997**	**2.03**

Visualmente, a reconstrução apresenta alta fidelidade. A Figura 10 compara a entrada (com lacuna simulada), a predição da rede e a verdade terrestre. O mapa de erro evidencia que as maiores discrepâncias ocorrem nas cristas das dunas, onde a variação de declividade é abrupta.



(a) Entrada (Lacuna) **(b) Predição (ViT)**

(c) Ground Truth **(d) Mapa de Erro**

Figura 10. Resultados qualitativos em campo de dunas. O modelo recupera a morfologia das ondulações (b) de forma consistente com o original (c).

4. Análise e Discussão

Os resultados indicam que a abordagem baseada em ViT é superior aos métodos de interpolação clássica na preservação da morfologia. O alto valor de SSIM (> 0.99) em todas as classes confirma que a estrutura visual do terreno é recuperada com precisão quase perfeita.

Contudo, observa-se uma dependência da complexidade do terreno na precisão altimétrica absoluta. Em **Planícies**, onde a variação topográfica é suave, o RMSE é extremamente baixo (3.33 m). Em **Dunas** e **Escarpas**, o erro sobe para 15 m. Isso se deve à ambiguidade de escala na inferência monocular: o modelo deduz corretamente a *forma* (frequência das dunas), mas pode ter um viés (*offset*) na amplitude vertical absoluta em áreas de sombra complexa.

A análise do perfil topográfico (Figura 11) corrobora esta análise: as curvas de predição e real são paralelas e morfologicamente idênticas, apresentando apenas deslocamentos verticais locais. Para aplicações científicas, isso significa que a análise de declividade e rugosidade — cruciais para a navegação de rovers — é preservada, mesmo que a altitude absoluta tenha margem de erro.

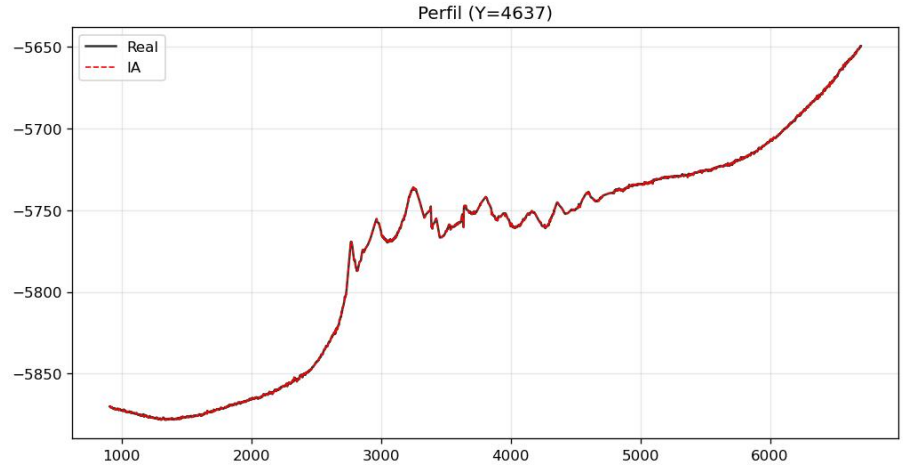


Figura 11. Perfil topográfico comparativo. A linha vermelha (Predição) segue fielmente a tendência e frequência da linha azul (Real).

5. Conclusão

Este trabalho apresentou um pipeline inédito utilizando *Vision Transformers* para o preenchimento de lacunas em DTMs de Marte. Os resultados confirmam que a arquitetura DPT é capaz de aprender a complexa função de transferência entre albedo e topografia. O método mostrou-se computacionalmente eficiente (2s por tile) e morfologicamente preciso. Trabalhos futuros focarão na inclusão de altimetria laser (MOLA) como restrição extra para corrigir o viés de escala em terrenos acidentados.

Referências

Kirk, R. L., Howington-Kraus, E., Rosiek, M., Mattson, S., Becker, K., Cook, D., Galuszka, D., Redding, B., Hare, T., and McEwen, A. (2011). An overview of the hirise dtm production pipeline. In *Lunar and Planetary Science Conference*, volume 42, page 1608.

McEwen, A. S., Eliason, E., Bergstrom, J., Bridges, N., Hansen, C., Delamere, W., Grant, J., Gulick, V., Herkenhoff, K., Keszthelyi, L., et al. (2007). Mars reconnaissance orbiter's high resolution imaging science experiment (hirise). *Journal of Geophysical Research: Planets*, 112(E5).

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188.

Tao, Y., Muller, J.-P., Conway, S. J., and Sidiropoulos, P. (2021). Madnet 2.0: Pixel-scale topography retrieval from single-view orbital imagery of mars using deep learning. *Remote Sensing*, 13(21):4220.