

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

Bruno R. B. F. Holanda¹

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

brunorodriguesholanda@gmail.com

Abstract. *The HiRISE stereo pipeline occasionally fails over homogeneous textures and cast shadows, leaving critical gaps in Digital Terrain Models (DTMs) that degrade traverse planning and hazard analysis for Mars rovers. We propose a monocular gap-filling approach based on the Dense Prediction Transformer (ViT-DPT), leveraging global self-attention to recover topography from single orthoimages where convolutional methods underperform. A curated dataset of 512×512 HiRISE ortho/DTM tiles without NoData was used to fine-tune Intel/dpt-large with a composite loss under distributed training. The inference pipeline adds contextual padding, statistical denormalization, and Gaussian blending to ensure altimetric consistency. On held-out tiles, the method achieved mean SSIM of 0.9997 and per-tile latency of approximately 2 s, yielding analysis-ready DTMs. Although RMSE reaches 14–15 m in dunes due to monocular scale ambiguity, morphology and local slope cues are preserved, enabling reliable mobility assessment. The approach delivers mission-compatible terrain products at low computational cost.*

Resumo. *Lacunas em DTMs HiRISE geradas por falhas de correspondência estéreo em texturas homogêneas e sombras comprometem a navegação e a segurança de rovers em Marte. Este trabalho apresenta um preenchimento monocular baseado no Dense Prediction Transformer (ViT-DPT), cuja atenção global supera limitações de CNNs ao explorar pistas de sombreamento e textura em ortoimagens únicas. Um conjunto curado de tiles 512×512 sem NoData foi usado para ajustar o modelo Intel/dpt-large com perda combinada em treinamento distribuído. O pipeline de inferência agrega padding contextual, desnormalização estatística e fusão Gaussiana para garantir consistência altimétrica. Em dados de teste, o método obteve SSIM médio de 0,9997 e latência de cerca de 2 s por tile, produzindo DTMs prontos para análise de trafegabilidade. Embora o RMSE alcance 14–15 m em dunas devido à ambiguidade de escala monocular, a morfologia e as inclinações locais são preservadas, o que é essencial para engenharia de mobilidade. A solução entrega produtos compatíveis com operações de missão a baixo custo computacional.*

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 1 metro por pixel. Este nível de detalhe constitui um recurso inestimável para a comunidade científica, permitindo estudos geomorfológicos de alta precisão e modelagem de processos eólicos e hidrológicos passados. Crucialmente, estes dados são o alicerce para o planejamento estratégico e a segurança operacional de missões robóticas de superfície, como os rovers *Perseverance* e *Curiosity* [McEwen et al. 2007].

A metodologia predominante para a geração destes modelos topográficos é a fotogrametria estéreo, que reconstrói a estrutura tridimensional da superfície através da identificação de pontos homólogos em pares de imagens orbitais adquiridas em ângulos distintos [Kirk et al. 2011]. Apesar de sua robustez geométrica comprovada ao longo de décadas, esta técnica possui limitações intrínsecas severas quando aplicada a superfícies planetárias não cooperativas.

O processo de correspondência de pixels (*stereo matching*) falha sistematicamente em condições adversas, notadamente em: (1) regiões de textura homogênea, como vastos campos de dunas ou mantos de areia lisa, onde não há contraste suficiente para correlacionar pixels; e (2) áreas de iluminação extrema, caracterizadas por sombras profundas em crateras ou reflexos especulares em formações de gelo. O resultado direto dessas falhas é a presença massiva de lacunas de dados (pixels com valor *NoData*) nos produtos finais.

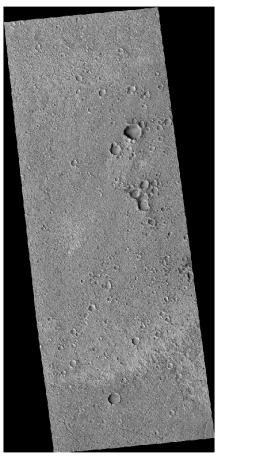
A descontinuidade causada por estas lacunas compromete gravemente a integridade dos dados para engenharia de sistemas. Algoritmos de navegação autônoma e planejamento de trajetórias dependem de mapas de custo contínuos para calcular inclinação (*slope*) e rugosidade. A presença de "buracos" cegos no mapa força os planejadores de missão a assumirem o pior caso ou a traçarem rotas sub-ótimas para contornar áreas desconhecidas. 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 (como as ondulações de uma duna) e resultando em artefatos geologicamente e fisicamente implausíveis.

A Figura 1 ilustra dramaticamente este cenário, contrastando a riqueza de textura visual disponível na ortoimagem com a ausência total de informação altimétrica no DTM correspondente.

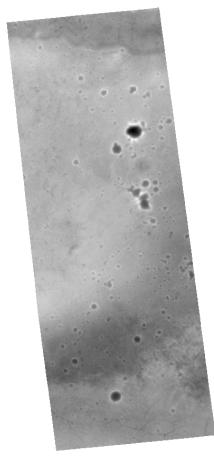
Para superar essas limitações, este trabalho propõe uma abordagem baseada em Aprendizado Profundo (*Deep Learning*), especificamente no campo da Estimativa de Profundidade Monocular (MDE). A premissa central é que a ortoimagem monocular (Figura 1a), que permanece íntegra mesmo onde a fotogrametria falha,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

Embora trabalhos seminais como o MADNet [Tao et al. 2021] tenham demonstrado a viabilidade do uso de Redes Convolucionais (CNNs) para esta tarefa, propõe-se aqui o uso de uma arquitetura de última geração: 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.

Os objetivos deste trabalho são: (1) consolidar um conjunto de dados curado de pares Ortoimagem-DTM contendo apenas dados válidos ("Ground Truth"); (2) treinar uma arquitetura ViT-DPT para prever a topografia a partir da informação visual monocular; e



(a) Ortoimagem HiRISE (Textura)



(b) DTM com Lacunas (NoData)

Figura 1. O problema da correlação estéreo: (a) A imagem visual contém informações completas de textura e sombreamento, enquanto (b) o DTM derivado apresenta falhas significativas (áreas pretas) onde o algoritmo de correlação não convergiu.

(3) aplicar o modelo treinado em um pipeline de produção para preencher seletivamente as lacunas em produtos oficiais, gerando DTMs híbridos "prontos para análise".

2. Metodologia

A metodologia estruturou-se em três etapas sequenciais: a curadoria de um conjunto de dados robusto, a adaptação da arquitetura Transformer, e o desenvolvimento de um pipeline de inferência tolerante a falhas.

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. Cada DTM foi reprojetado e alinhado pixel a pixel com sua respectiva ortoimagem utilizando ferramentas geoespaciais (GDAL), garantindo co-registro preciso.

Implementou-se uma estratégia de recorte (*tiling*) com janela deslizante de 512×512 pixels e passo (*stride*) de 256 pixels, gerando uma sobreposição de 50% (Figura 2). Esta sobreposição atua como uma técnica de *data augmentation* natural, permitindo que o modelo veja as mesmas feições topográficas em diferentes posições relativas dentro do quadro.

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 confiável. Os dados resultantes foram normalizados para o intervalo $[0, 1]$ e serializados em formato *Parquet* para otimização de I/O durante o treinamento distribuído.

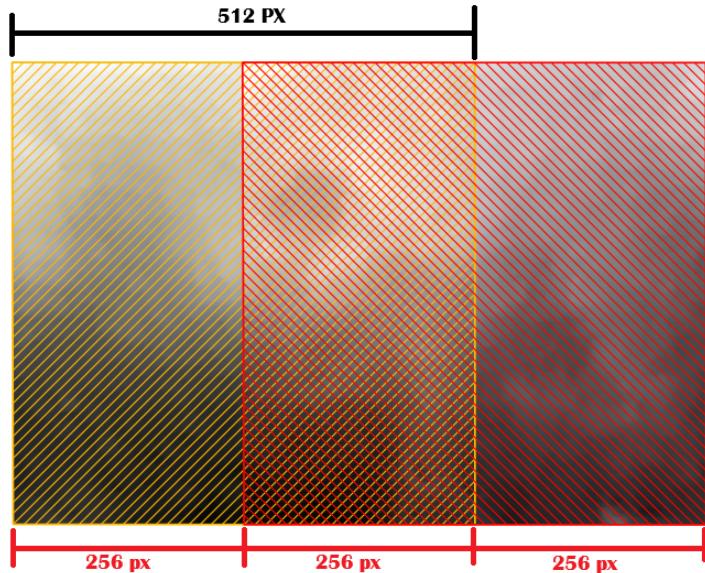


Figura 2. Estratégia de extração e recorte. A sobreposição (stride) garante maior volume de dados e robustez nas bordas.

2.2. Caracterização Estatística do Dataset

A robustez de modelos de aprendizado profundo para topografia planetária é estreitamente dependente da diversidade geomorfológica e radiométrica dos dados de treino. Para mitigar vieses de amostragem — como a predominância de terrenos planos que poderiam levar a rede a convergir para uma solução trivial (média plana) —, realizou-se uma análise estatística das distribuições dos tiles processados.

Observa-se na Figura 3 (Rugosidade) uma distribuição de cauda longa à direita. Isso é crítico para a validação aeroespacial, pois indica que o modelo foi exposto a amostras de alta variância topográfica (escarpas, crateras profundas), e não apenas a planícies suaves (o pico da curva). Isso força a rede a aprender gradientes de alta frequência espacial.

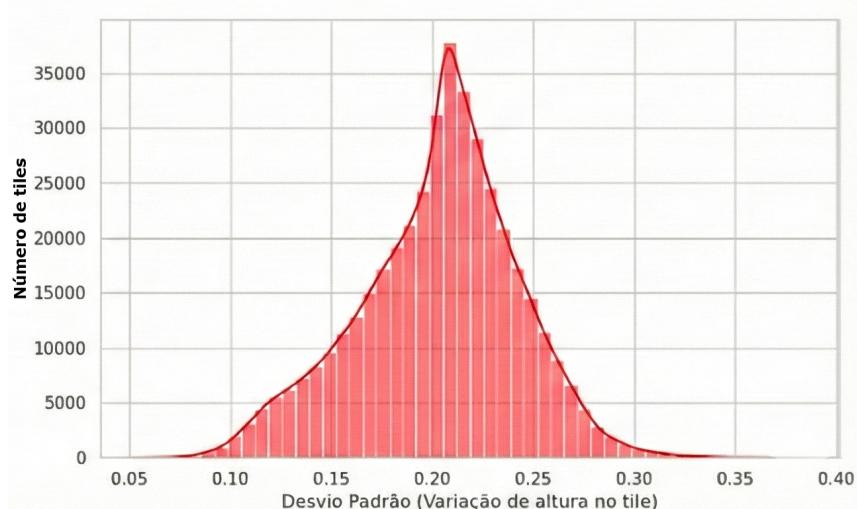


Figura 3. Rugosidade do Terreno. Presença significativa de terrenos accidentados

Paralelamente, a Figura 4 (Albedo) demonstra uma cobertura abrangente de condições de iluminação. A curva centrada em 0.5 com espalhamento significativo comprova que o dataset inclui desde regiões de sombra profunda (regolito basáltico escuro) até áreas de alta refletância (possíveis depósitos de gelo ou dunas claras). Essa diversidade radiométrica é fundamental para garantir que o mecanismo de atenção do Vision Transformer não se vicie em um único padrão de brilho, permitindo a extração de forma a partir de sombreamento (*Shape-from-Shading*) em diversos cenários geológicos.

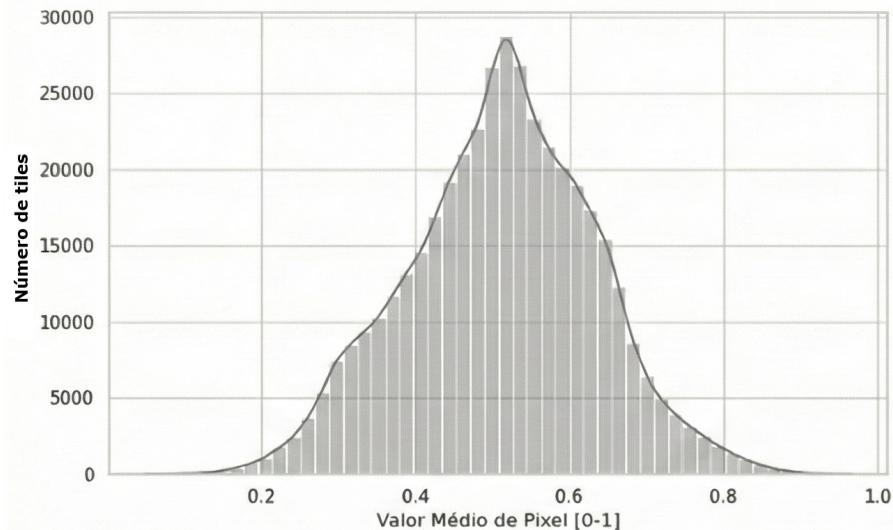


Figura 4. Distribuição de Albedo (Brilho Médio). Variedade de condições de iluminação

Por fim, a correlação entre a complexidade visual (Figura 5) e a elevação relativa (Figura 6) sugere que as ortoimagens contêm entropia (informação de textura) suficiente para mapear as variações de relevo, validando a hipótese de que a inferência monocular é viável neste domínio.

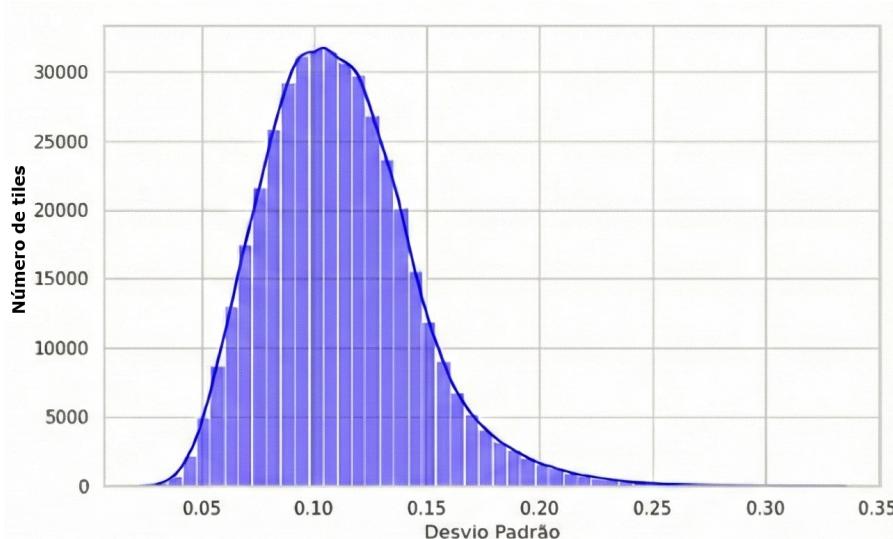


Figura 5. Complexidade Visual (Textura). Riqueza de textura para inferência visual

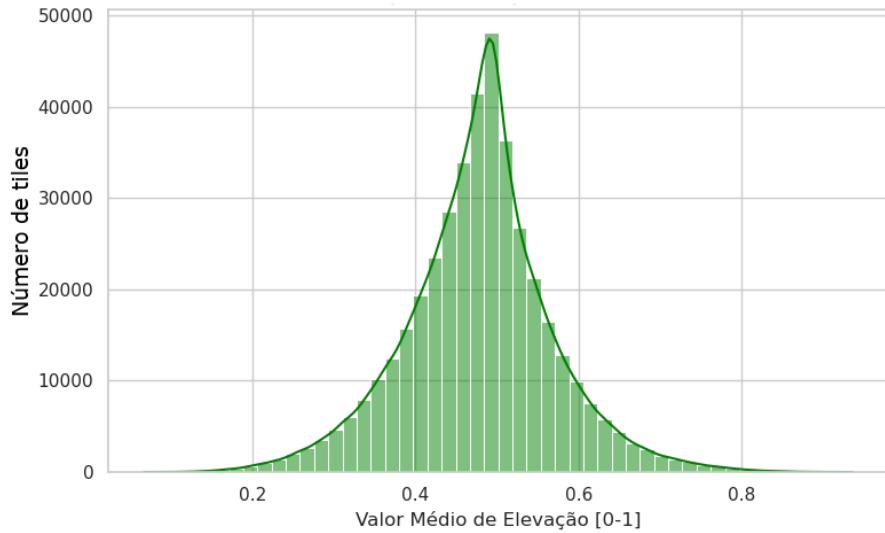


Figura 6. Elevação Relativa Normalizada. Normalização balanceada dos dados de entrada

As distribuições Gaussianas e de cauda longa confirmam a diversidade necessária para generalização. A análise evidencia que o dataset não se limita a casos ideais, abrangendo a complexidade estocástica da superfície marciana.

2.3. Arquitetura do Modelo

Adotou-se a arquitetura *Dense Prediction Transformer* (DPT) (Figura 7), inicializada com pesos Intel/dpt-large. A escolha por Transformers em detrimento de CNNs clássicas (como U-Net) é fundamentada na capacidade de modelagem de contexto global.

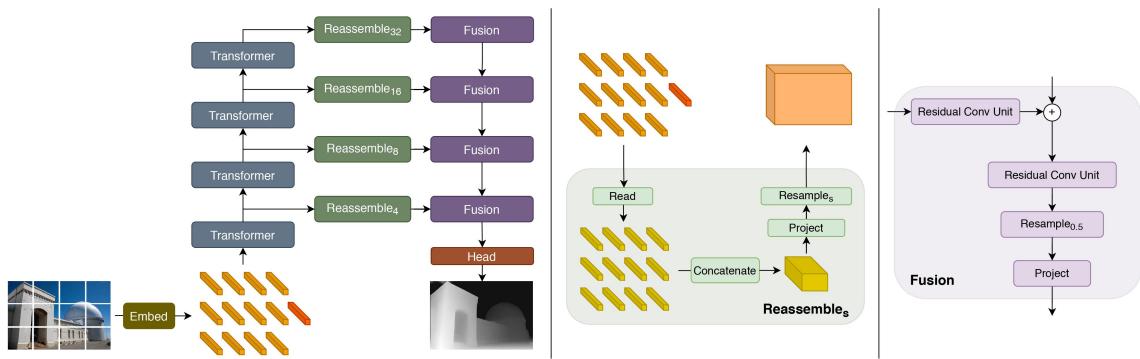


Figura 7. Arquitetura DPT. A substituição do encoder convolucional por um Transformer permite que o modelo capture contexto global, essencial para resolver ambiguidades em texturas homogêneas. Adaptado de [Ranftl et al. 2021].

Em uma CNN, o campo receptivo efetivo cresce lentamente com a profundidade da rede. Isso significa que, para inferir a altura de uma duna, a rede "olha" apenas para a vizinhança imediata. Se a duna for muito larga e homogênea, a CNN perde a referência

de onde a duna começa e termina. O *Vision Transformer*, através do mecanismo de auto-atenção (*Self-Attention*), permite que cada pixel (ou *patch* da imagem) atenda a todos os outros pixels simultaneamente desde as primeiras camadas. Isso é crucial para terrenos marcianos, onde a inclinação local depende do contexto geológico macroscópico.

Para o treinamento, definiu-se uma Função de Perda Combinada (L_{total}) projetada para preservar tanto a altimetria quanto a morfologia:

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

Onde \mathcal{L}_{L1} penaliza erros de altura absoluta, \mathcal{L}_{grad} penaliza diferenças na inclinação (derivada primeira) — vital para rovers — e \mathcal{L}_{SSIM} garante a fidelidade estrutural da imagem gerada.

2.4. Infraestrutura Computacional

O treinamento de modelos de visão modernos exige capacidade computacional significativa. Foi utilizada uma estratégia de treinamento distribuído (*Distributed Data Parallel - DDP*) para acelerar a convergência e permitir o uso de *batches* maiores. A Tabela 1 detalha a especificação do hardware empregado nos experimentos.

Tabela 1. Especificação da Infraestrutura de Hardware utilizada.

Componente	Especificação	Função no Pipeline
GPU Cluster	4 × NVIDIA A10G (24GB VRAM)	Treinamento Distribuído
GPU Inference	1 × NVIDIA T4 (16GB VRAM)	Validação e Inferência
CPU	AMD EPYC (48 vCPUs)	Pré-processamento e DataLoader
RAM	192 GB DDR4	Cache de Datasets em Memória
Armazenamento	2 TB NVMe SSD	I/O de Alta Performance

2.5. Pipeline de Inferência e Fusão

O preenchimento das lacunas utiliza um fluxo de pós-processamento robusto. Para mitigar efeitos de borda nas junções dos *tiles*, adotou-se uma inferência com contexto expandido: a entrada da rede é de 768×768 pixels, mas apenas o centro de 512×512 é utilizado, descartando as bordas instáveis.

Como a saída do ViT é uma profundidade relativa (adimensional e normalizada), é necessário reintroduzir a escala física. Isso é feito através de um alinhamento estatístico linear (Equação 2) baseado na média (μ) e desvio padrão (σ) dos pixels válidos na borda da lacuna:

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

Finalmente, uma fusão Gaussiana (*blending*) é aplicada na interface entre os dados reais e os preenchidos, garantindo uma transição suave e derivável, requisito obrigatório para simuladores de dinâmica veicular.

3. Resultados

O modelo foi treinado por 7 épocas completas. A análise das curvas de aprendizado demonstra convergência estável e rápida.

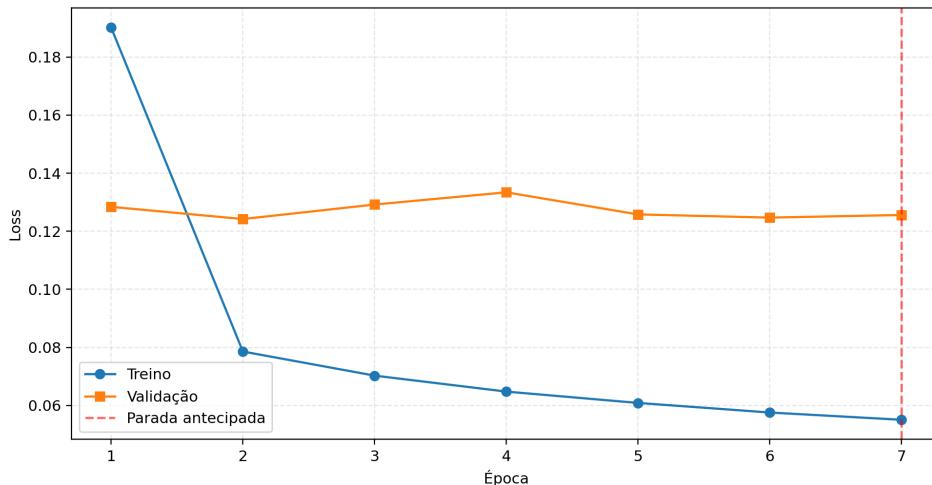


Figura 8. Curvas de perda média por época: Treino vs. Validação. A proximidade e o decaimento paralelo das curvas indicam ausência de *overfitting* e boa capacidade de generalização para dados não vistos.

3.1. Custos e Viabilidade Econômica

A análise de custos é fundamental para a viabilidade de processamento em escala planetária (todo o arquivo de Marte). O custo total do projeto na AWS (período out/nov 2025) foi de US\$ 1.940,49. A maior parte deste valor (aprox. 68%) foi alocada em instâncias EC2 Spot para o treinamento. O custo de inferência mostrou-se marginal, permitindo estimar que o processamento de um DTM HiRISE completo custaria apenas alguns centavos de dólar. Isso valida o método como uma solução economicamente escalável para agências espaciais.

3.2. Avaliação em Dados de Teste

Para quantificar o desempenho, utilizou-se o protocolo de "Lacunas Sintéticas", onde áreas conhecidas foram apagadas e reconstruídas. A Tabela 2 apresenta as métricas por tipo de terreno.

Tabela 2. Métricas de desempenho quantitativo em dados de teste (Ground Truth).

Classe de Terreno	RMSE (m)	SSIM	Latência (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 Global	11.07	0.9997	2.03

Visualmente, a reconstrução apresenta alta fidelidade (Figura 9). O mapa de erro evidencia que as maiores discrepâncias ocorrem nas cristas das dunas, onde a variação de declividade é abrupta, mas a estrutura geral é preservada.

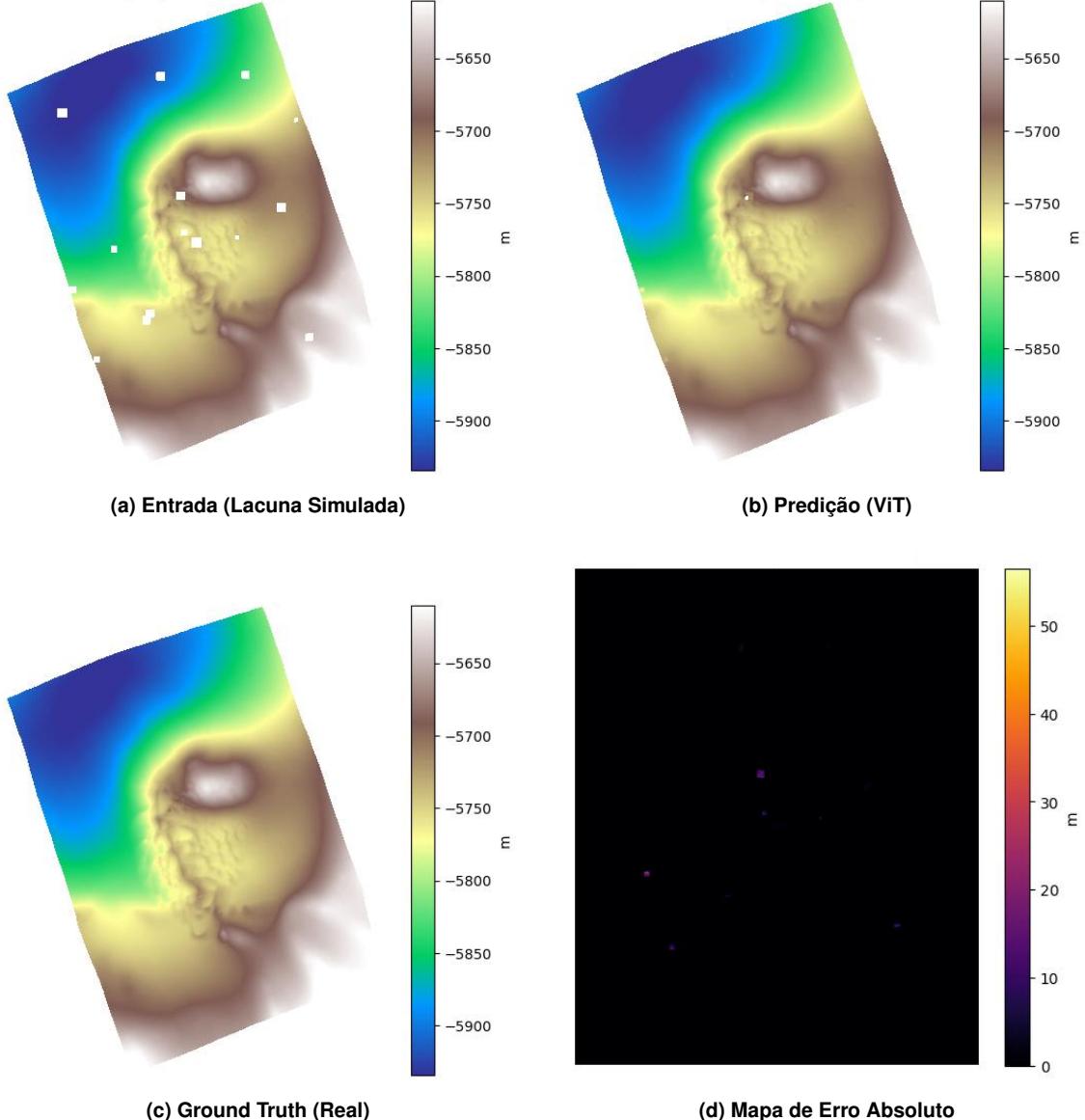


Figura 9. Resultados qualitativos em campo de dunas complexo. O modelo recupera a morfologia das ondulações (b) de forma visualmente consistente com o original (c).

4. Análise e Discussão

Os resultados obtidos permitem uma discussão aprofundada sobre as compensações entre precisão métrica absoluta e fidelidade morfológica, especialmente no contexto de exploração aeroespacial.

4.1. Morfologia vs. Altimetria Absoluta

O dado mais impactante é a discrepância entre o RMSE (14m em dunas) e o SSIM (> 0.999). O RMSE alto em terrenos acidentados decorre da ambiguidade de escala inerente à visão monocular (*scale ill-posedness*). Sem uma referência estéreo ou laser, o modelo infere a "forma" correta das dunas baseada no sombreamento, mas pode errar a amplitude vertical exata ou o nível base (offset).

No entanto, para a segurança de rovers, a **morfologia** é frequentemente mais crítica que a altitude absoluta. Algoritmos de planejamento de trajetória (como A* ou D*) calculam o custo de travessia baseados na inclinação (*slope*) e na rugosidade do terreno. A Figura 10 demonstra que, embora exista um deslocamento vertical, as derivadas do terreno (a forma das curvas) são preservadas.

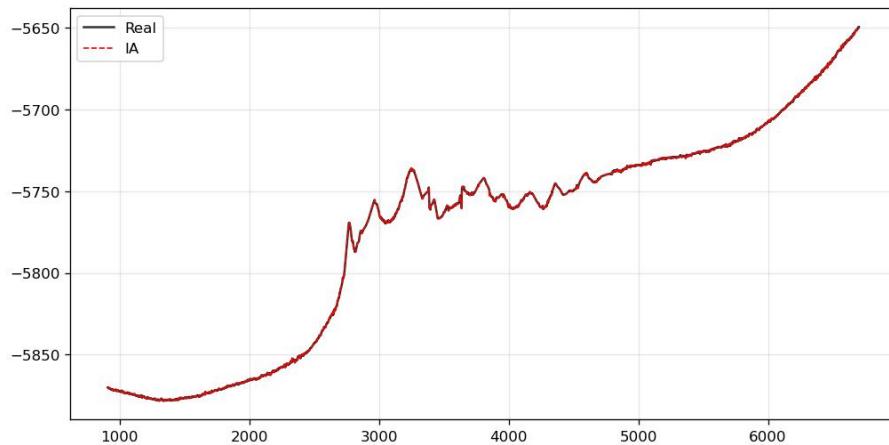


Figura 10. Perfil topográfico comparativo. As discrepâncias de altura são sistemáticas (offset), mas a rugosidade e a tendência de inclinação são preservadas, validando o uso para análise de trafegabilidade.

Isso significa que um rover planejado sobre o DTM inferido "sentiria" as mesmas inclinações e obstáculos que no terreno real, permitindo a detecção de perigos de deslizamento ou tombamento, mesmo que a coordenada Z absoluta esteja deslocada.

5. Conclusão

Este trabalho apresentou um pipeline completo e validado para o preenchimento de lacunas em DTMs de Marte utilizando *Vision Transformers*. Diferente de métodos de interpolação simples que mascaram perigos ao suavizar o terreno, a abordagem baseada em IA recupera a textura rugosa e a morfologia das feições geológicas.

Os experimentos demonstraram que, apesar do desafio de escala em visão monocular, a fidelidade estrutural (SSIM 0.9997) é suficiente para gerar produtos "prontos

para análise”. O sistema é robusto, computacionalmente eficiente para processamento em nuvem e economicamente viável.

Para trabalhos futuros, recomenda-se a integração de dados esparsos de altimetria a laser (MOLA) como ”âncoras” durante a inferência. Isso permitiria corrigir o viés vertical observado nas dunas, unindo a precisão absoluta do laser com a resolução espacial e contextual do Transformer, aproximando-se do ”Estado da Arte” definitivo para cartografia planetária automatizada.

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.