THE ROBUSTNESS OF NATURAL IMAGE PRIORS IN REMOTE SENSING: A ZERO-SHOT VAE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the robustness of variational autoencoders (VAEs) pre-trained on natural images when applied zero-shot to remote sensing data. Despite domain mismatch, these VAEs demonstrate superior reconstruction performance, suggesting natural image priors transfer effectively to satellite imagery.

1 INTRODUCTION

While specialized visual generative models for remote sensing (Liu et al., 2025; Khanna et al., 2024; Yellapragada et al., 2025; Yu et al., 2025; Pang et al., 2026; Sastry et al., 2024; Pan et al., 2025; Sebaq & ElHelw, 2024) have emerged, standard VAEs pre-trained on large-scale natural image datasets (e.g., LAION (Schuhmann et al., 2022) and ImageNet (Russakovsky et al., 2015)) remain commonly used without domain adaptation. We evaluate whether these zero-shot VAEs can effectively compress and reconstruct satellite imagery, despite distinct viewing geometries and spatial resolutions (Rolf et al., 2024).

2 PRELIMINARY: VARIATIONAL AUTOENCODERS

Variational Autoencoders (VAEs) (Kingma & Welling, 2014) learn to map input x to latent representation z via encoder $q_\phi(z|x)$ and reconstruct via decoder $p_\theta(x|z)$. The objective maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\lambda(z)) \tag{1}$$

where the first term is reconstruction likelihood and the second regularizes the latent space against prior $p_\lambda(z)$ (typically $\mathcal{N}(0, I)$). In large-scale generative models, VAEs compress high-dimensional pixel data into manageable latent spaces for efficient modeling.

3 DATASETS AND METHODS

We evaluate zero-shot VAE performance on two benchmarks: **RESISC45** (Cheng et al., 2017) (31,500 images, 45 classes, 20cm–30m/px GSD) and **AID** (Xia et al., 2017) (10,000 images, 30 classes, 600×600px). We test VAEs from Stable Diffusion (Rombach et al., 2022; Podell et al., 2024), FLUX (Black Forest Labs, 2025b;a), SANA (Xie et al., 2025), and Qwen (Wu et al., 2025) families. All models and latents are evaluated using *bfloat16* precision for computational efficiency. Evaluation metrics: PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017). Results are summarized in Table 1.

4 RESULTS

Overall, VAEs demonstrate mostly good reconstruction performance across the evaluated metrics (Table 1). We observe that results on RESISC45 are slightly worse than AID, which we attribute to RESISC45 containing a number of low-resolution images that lack sufficient high-frequency features, leading to VAE reconstruction failures. Qualitative examples shown in Figure 1 demonstrate visually near-identical reconstructions across all models.

Model	GFLOPs	Latent Shape	PSNR↑		SSIM↑		LPIPS↓		FID↓	
			RESISC45	AID	RESISC45	AID	RESISC45	AID	RESISC45	AID
SD21-VAE	894.91	(4,32,32)	25.71	26.66	0.672	0.709	0.095	0.094	4.13	3.08
SDXL-VAE	894.91	(4,32,32)	25.83	26.80	0.692	0.726	0.098	0.098	4.98	3.11
SD35-VAE	895.25	(16,32,32)	29.71	30.72	0.862	0.876	0.035	0.037	1.11	0.69
FLUX1-VAE	895.25	(16,32,32)	33.30	33.63	0.923	0.918	0.022	0.025	0.38	0.26
FLUX2-VAE	895.71	(32,32,32)	**33.42**	**34.46**	**0.925**	**0.926**	**0.021**	**0.022**	**0.46**	**0.37**
SANA-VAE	846.76	(32,8,8)	23.36	N/A*	0.558	N/A*	0.124	N/A*	8.69	N/A*
Qwen-VAE	1143.88	(16,32,32)	30.38	31.46	0.874	0.889	0.080	0.077	9.51	0.42

Table 1: VAE model statistics and zero-shot performance on RESISC45 (31.5K images, 45 classes, 20cm–30m/px GSD) and AID (10K images, 30 classes, 600×600px). Input shape: (3, 256, 256). *SANA-VAE cannot process AID images because 600 is not divisible by 32 (SANA-VAE's spatial compression factor).

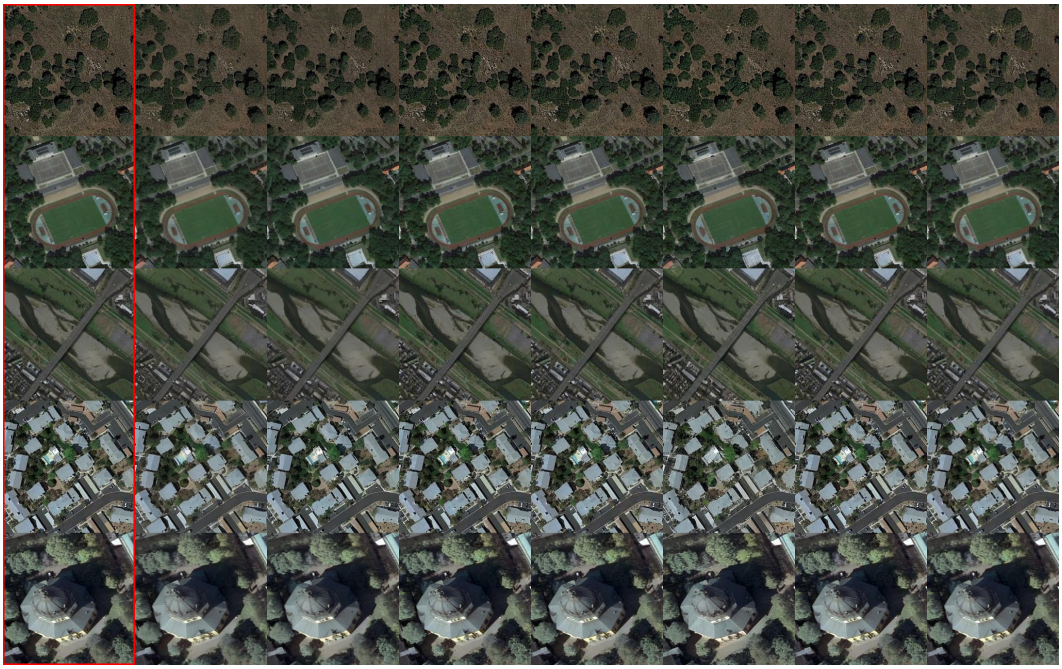

Figure 1: Qualitative reconstructions from 5 random RESISC45 samples. Each column shows (left to right): Original, SD21-VAE, SDXL-VAE, SD35-VAE, FLUX1-VAE, SANA-VAE, FLUX2-VAE, and Qwen-VAE. No significant visual difference appears.

5 Insights and Conclusion

Insights 1: We find that VAEs reconstruct remote sensing images remarkably well, with reconstructions appearing visually nearly identical to the input. We argue that VAEs may have the potential to implicitly deblur and denoise input images, where the reconstructed image serves as a better data source for model training (e.g., representation learning) with possibly improved statistics.

Insights 2: As the compression appears effectively lossless, we argue for directly storing latent representations instead of original images as datasets to reduce storage requirements.

In this work, we explored the robustness of natural image priors in VAEs for remote sensing. Our findings indicate that these models, when used zero-shot, can provide significant utility in data compression across various categories. We will release the reconstructed images along with their corresponding latents for community exploration and further research.

REFERENCES

Black Forest Labs. FLUX.2: Frontier Visual Intelligence, November 2025a.

Black Forest Labs. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, June 2025b.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi: 10.1109/JPROC.2017.2675998.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *International Conference on Learning Representations*, 2024.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

Chenyang Liu, Keyan Chen, Rui Zhao, Zhengxia Zou, and Zhenwei Shi. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–23, 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3560455.

Jiancheng Pan, Shiye Lei, Yuqian Fu, Jiahao Li, Yanxing Liu, Yuze Sun, Xiao He, Long Peng, Xiaomeng Huang, and Bo Zhao. EarthSynth: Generating informative earth observation with diffusion models, August 2025.

Li Pang, Xiangyong Cao, Datao Tang, Shuang Xu, Xueru Bai, Feng Zhou, and Deyu Meng. HSI-Gene: A Foundation Model for Hyperspectral Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):730–746, January 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3610927.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission Critical – Satellite Data is a Distinct Modality in Machine Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 42691–42706. PMLR, July 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. GeoSynth: Contextually-Aware High-Resolution Satellite Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 460–470, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Ahmad Sebaq and Mohamed ElHelw. RSDiff: Remote sensing image generation from text using diffusion model. *Neural Computing and Applications*, 36(36):23103–23111, December 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10363-3.

Zhou 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*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report, August 2025.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Text-to-Image Synthesis with Linear Diffusion Transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Srikar Yellapragada, Alexandros Graikos, Kostas Triaridis, Prateek Prasanna, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras. ZoomLDM: Latent Diffusion Model for Multi-scale Image Generation. In *CVPR*, pp. 23453–23463, 2025.

Zhiping Yu, Chenyang Liu, Liqin Liu, Zhenwei Shi, and Zhengxia Zou. MetaEarth: A Generative Foundation Model for Global-Scale Remote Sensing Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1764–1781, March 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3507010.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.