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

Zhenyuan Chen
Zhejiang University
School of Earth Sciences
bili_sakura@zju.edu.cn

Feng Zhang
Zhejiang University
School of Earth Sciences
zfcarnation@zju.edu.cn

ABSTRACT

This paper explores the robustness of variational autoencoders (VAEs) pre-trained on natural image data, such as ImageNet, when applied to the remote sensing domain in a zero-shot manner. We investigate whether these natural image priors embedded in standard VAEs can serve as effective compressors and reconstructors for satellite images, even when applied in a different manner across various settings compared to natural cases. Our study evaluates several state-of-the-art VAE architectures across multiple remote sensing categories and reconstruction metrics to demonstrate their potential.

1 INTRODUCTION

The rapid development of visual foundation models has transformed the landscape of generative AI, with milestone architectures like GANs (Goodfellow et al., 2014; Karras et al., 2018; Brock et al., 2019; Karras et al., 2020; Sauer et al., 2022) and diffusion models (Ho et al., 2020; Song et al., 2020; 2021; Dhariwal & Nichol, 2021; Karras et al., 2022; Lu et al., 2022; Rombach et al., 2022) setting new standards for high-fidelity image synthesis in the general domain. This momentum has recently extended to the remote sensing (RS) domain, where specialized models such as Text2Earth (Liu et al., 2025), DiffusionSat (Khanna et al., 2024), and other Earth observation foundation models (Lu et al., 2025; Tuia et al., 2025) have been developed to capture complex geospatial distributions, alongside other RS generative models (Yellapragada et al., 2025; Yu et al., 2025; Pang et al., 2026; Sastry et al., 2024; Pan et al., 2025; Sebaq & ElHelw, 2024). Despite these advancements, RS imagery presents unique challenges compared to natural images, including distinct viewing geometries, multi-spectral bands, and varying spatial resolutions, as highlighted in several position papers (Rolf et al., 2024). A common practice remains the use of standard VAEs pre-trained on natural image priors (e.g., ImageNet) without domain-specific adaptation. In this work, we investigate the robustness of these zero-shot VAEs in the RS context, focusing on their effectiveness as compressors and reconstructors for satellite data.

2 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 represents the reconstruction likelihood and the second term is the Kullback-Leibler (KL) divergence regularizing the latent space against a prior distribution $p_\lambda(z)$, typically a standard Gaussian $\mathcal{N}(0, I)$. Modern VAEs often employ advanced architectures such as VQ-GAN (Esser et al., 2021) or flow-matching based decoders to improve reconstruction fidelity. In the context of large-scale generative models, these VAEs serve as essential components by compressing high-dimensional pixel data into a manageable latent space for downstream diffusion or transformer-based modeling.

3 EXPERIMENTS

In this study, we evaluate several state-of-the-art VAE architectures in a zero-shot manner on remote sensing data. We include models from the Stable Diffusion family (SD21-VAE, SDXL-VAE, SD35-VAE) (Rombach et al., 2022; Podell et al., 2024), the FLUX family (FLUX.1-VAE, FLUX.2-VAE) (Black Forest Labs, 2025b;a), and other efficient architectures such as SANA-VAE (Xie et al., 2025) and Qwen-VAE (Wu et al., 2025). These models were primarily pre-trained on natural image datasets like ImageNet and LAION, and we test their direct applicability to RS benchmarks without any fine-tuning.

Model	GFLOPs	Spatial Comp. Ratio	Latent Ch.	PSNR↑		SSIM↑		LPIPS↓		FID↓	
				RESISC45	AID	RESISC45	AID	RESISC45	AID	RESISC45	AID
SANA-VAE	846.76	32	32	23.36	24.72	0.558	0.606	0.124	0.123	8.69	5.01
SD21-VAE	894.91	8	4	25.71	26.66	0.672	0.709	0.095	0.094	4.13	3.08
SDXL-VAE	894.91			25.83	26.80	0.692	0.726	0.098	0.098	4.98	3.11
SD35-VAE	895.25		16	29.71	30.72	0.862	0.876	0.035	0.037	1.11	0.69
FLUX1-VAE	895.25			33.30	33.63	0.923	0.918	0.022	0.025	**0.38**	**0.26**
Qwen-VAE	1143.88			30.38	31.46	0.874	0.889	0.080	0.077	9.51	0.42
FLUX2-VAE	895.71		32	**33.42**	**34.46**	**0.925**	**0.926**	**0.021**	**0.022**	0.46	0.37

Table 1: VAE model statistics and zero-shot performance on the full RESISC45 (31.5K images, 45 classes, 20cm–30m/px GSD) and AID (10K images, 30 classes, 600×600px) datasets, evaluated at their original image sizes (RESISC45: 256×256; AID: 600×600). Spatial comp. ratio denotes the per-dimension spatial downsampling factor (input:latent), and latent ch. denotes the number of latent channels.

Model	GFLOPs	Spatial Comp. Ratio	Latent Ch.	PSNR↑	SSIM↑	LPIPS↓	FID↓	CMMD↓
SANA-VAE	846.76	32	32	22.33	0.564	0.112	28.64	0.0002
SD21-VAE	894.91	8	4	25.81	0.688	0.082	16.43	0.0172
SDXL-VAE	894.91			25.92	0.705	0.084	15.97	0.0203
SD35-VAE	895.25		16	30.06	0.858	0.030	6.85	0.0001
FLUX1-VAE	895.25			31.73	0.899	0.020	**5.19**	0.0010
Qwen-VAE	1143.88			30.76	0.873	0.064	15.83	0.0106
FLUX2-VAE	895.71		32	**32.16**	**0.901**	**0.019**	4.23	**0.0001**

Table 2: Zero-shot performance on the UCMerced dataset (2.1K images, 21 classes, 256×256px), evaluated at original image size.

4 EXPERIMENTAL RESULTS

We evaluate the performance of various VAE architectures on multiple benchmark remote sensing datasets: NWPU-RESISC45 (Cheng et al., 2017), AID (Xia et al., 2017), and UCMerced (Yang & Newsam, 2010). Our evaluation focuses on zero-shot reconstruction quality across diverse aerial scene categories, using the full datasets at their original image sizes (RESISC45: 256×256; AID: 600×600; UCMerced: 256×256).

4.1 METRICS AND MAIN RESULTS

Reconstruction quality is assessed using standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al., 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and reconstruction Fréchet Inception Distance (FID) (Heusel et al., 2017). Table 1 summarizes the quantitative performance across the RESISC45 and AID datasets, while Table 2 presents results on the UCMerced dataset. We observe that while all pre-trained VAEs demonstrate remarkable zero-shot transfer capability, the FLUX and SDXL families consistently outperform older architectures in preserving fine-grained geospatial textures.

We also run a quick reconstruction sanity check on 10 samples per modality using the single-channel expansion strategy in `scripts/quick_vae_reconstruction.py`, with averages summarized in Table 3.

Data	VAE	Res	MAE↓	PSNR↑	SSIM↑
IR	SD21-VAE	1024	0.0213	28.49	0.7594
	SD35-VAE	1024	0.0164	31.83	0.8858
	FLUX2-VAE	1024	**0.0091**	**36.80**	**0.9548**
EO	SD21-VAE	256	0.0102	34.69	0.9331
	SD35-VAE	256	0.0051	41.44	0.9810
	FLUX2-VAE	256	**0.0042**	**43.83**	**0.9903**
RGB	SD21-VAE	1024	0.0299	25.64	0.6732
	SD35-VAE	1024	0.0228	28.50	0.8221
	FLUX2-VAE	1024	**0.0145**	**32.76**	**0.9173**
SAR	SD21-VAE	1024	**0.0057**	42.67	0.9789
	SD35-VAE	1024	0.0096	39.94	0.9823
	FLUX2-VAE	1024	**0.0057**	**43.83**	**0.9938**

Table 3: Quick reconstruction sanity check (10 images per modality) from `scripts/run_quick_vae_reconstruction.sh`, averaged over per-image MAE/PSNR/SSIM.

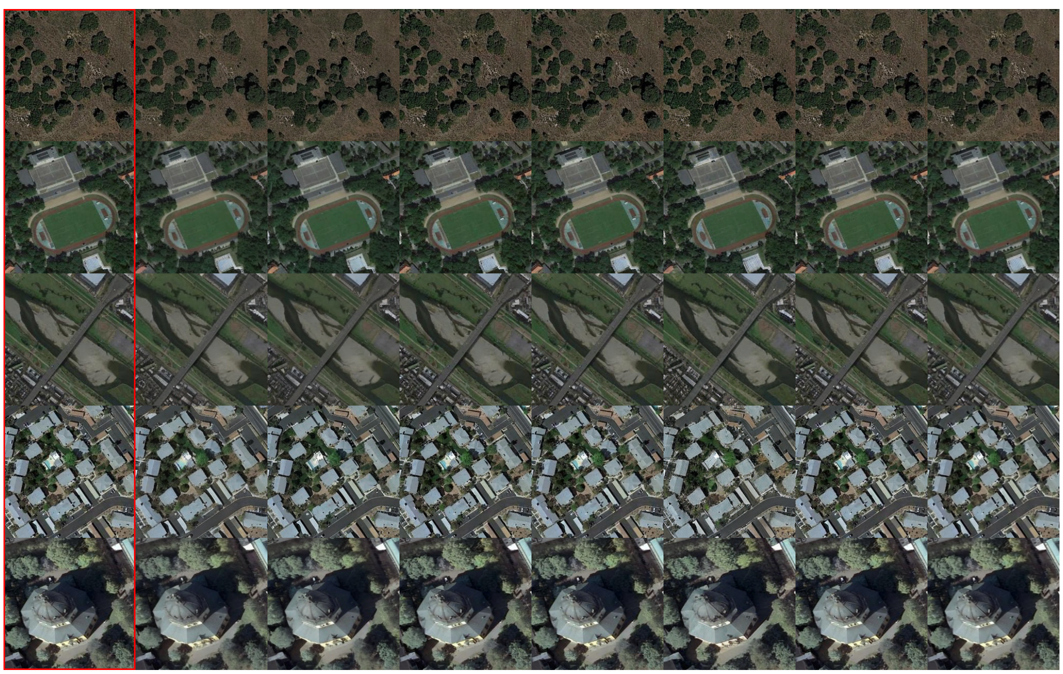

Figure 1: Qualitative reconstructions from 5 random RESISC45 samples. Each column shows (left to right): Original, SD21-VAE, SDXL-VAE, SD35-VAE, FLUX1-VAE, SANA-VAE, FLUX2-VAE, and Qwen-VAE. No significant visual difference appears.

5 Insights

Based on our extensive experiments, we highlight several key insights regarding the application of natural image VAEs to the remote sensing domain:

Key Insights

- **Generalization of Natural Priors:** Standard VAEs pre-trained on ImageNet/LAION exhibit surprising robustness to RS data, suggesting that low-level visual features (edges, textures) are highly transferable across domains.
- **Model Architecture Matters:** Flow-matching based decoders (e.g., FLUX) provide significantly higher reconstruction fidelity for high-resolution satellite imagery compared to earlier KL-regularized architectures.
- **Zero-Shot Utility:** These models can serve as effective ”cheap” pre-processors for denoising and initial data compression in RS pipelines without the need for expensive domain-specific re-training.

6 Conclusion

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 and pre-processing tasks across various sensing categories. Future work could further explore the integration of domain-specific priors to enhance these capabilities.

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.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

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.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. 35:5775–5787, December 2022.

Siqi Lu, Junlin Guo, James R. Zimmer-Dauphinee, Jordan M. Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A. Wernke, and Yuankai Huo. Vision Foundation Models in Remote Sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):190–215, September 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3541952.

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. HSIGene: 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.

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.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. October 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Devis Tuia, Konrad Schindler, Begüm Demir, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N. van Rijn, Holger H. Hoos, Fabio Del Frate, Mihai Datcu, Volker Markl, Bertrand Le Saux, Rochelle Schneider, and Gustau Camps-Valls. Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geoscience and Remote Sensing Magazine*, 13(4):119–141, December 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2024.3425961.

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.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279. ACM, 2010.

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.