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

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper explores the robustness of variational autoencoders (VAEs) pre-trained on natural image datasets, such as ImageNet, when applied to the remote sensing domain in a zero-shot manner. We investigate whether these "natural English priors" embedded in standard VAEs can serve as effective compressors and reconstructors for satellite imagery, which often exhibits significantly different statistical properties compared to natural scenes. Our study evaluates several state-of-the-art VAE architectures—including SD21-VAE, SDXL-VAE, SD35-VAE, FLUX.1/2-VAE, SANA-VAE, and Qwen-VAE—across multiple remote sensing categories using standard reconstruction metrics. Furthermore, we demonstrate the potential of zero-shot VAEs as cheap pre-processors for denoising and dehazing remote sensing data.

1 INTRODUCTION

The field of computer vision has witnessed a dramatic rise in foundation models, particularly generative vision models capable of high-fidelity visual generation. Milestone architectures such as GAN-based models like StyleGAN (Karras et al., 2019) and GigaGAN (Kang et al., 2023) have set early benchmarks in generative modeling. More recently, the field has been dominated by a vast array of diffusion models, including DDPM (Ho et al., 2020), ADM (Dhariwal & Nichol, 2021), EDM (Karras et al., 2022), and latent diffusion models (LDM) such as Stable Diffusion (Rombach et al., 2022). Furthermore, the emergence of large multimodal models (Wang et al., 2026) and vision-language models for remote sensing (Li et al., 2024; Weng et al., 2025) has further expanded the scope of foundation models in the field. These models have demonstrated unprecedented strength in capturing complex visual distributions across diverse domains.

In the remote sensing domain, these visual generative models have also come into significant interest, with recent surveys highlighting the potential of vision foundation models (Lu et al., 2025; Xiao et al., 2025; Tuia et al., 2025) and multimodal models (Bai et al., 2025; Liu et al., 2025) in Earth observation. A common practice is to employ standard VAEs pre-trained on general domain data, such as ImageNet, without further domain-specific adaptation or fine-tuning. This raises a fundamental question: can these pre-trained VAEs serve as reliable compressors or reconstructors when adopted to an out-of-domain context like remote sensing? Given that the visual priors in these models are primarily derived from "natural" English-centric datasets, their robustness in specialized domains remains an open area of investigation.

The challenges in remote sensing are compounded by the inherent differences between natural images and satellite observations. Unlike natural scenes, remote sensing data often involves unique viewing geometries, multi-spectral bands, and varying spatial resolutions (Cheng et al., 2017). Furthermore, labeled datasets in remote sensing are frequently sparse, with variable labeling schemes and qualities (Christie et al., 2018). In contrast to standard machine learning settings where observations X and labels Y are definitive pairs, satellite machine learning often deals with label annotations generated independently of specific satellite observations. Instead, labels are often paired with many different choices of satellite observations corresponding to the label's location and time index, introducing further complexity into the learning process.

2 RELATED WORK

Generative Models for Remote Sensing: Recent work has explored the application of generative models to various remote sensing tasks, including scene classification (Cheng et al., 2017), object localization (Long et al., 2017), and image retrieval (Xiao et al., 2017). The shift towards foundation models has led to the development of specialized architectures for Earth observation (Lu et al., 2025).

Variational Autoencoders and Latent Spaces: VAEs (Kingma & Welling, 2014) have long been used for representation learning. Recent advancements such as VQGAN (Esser et al., 2021) and its successors have improved the quality of latent spaces for high-resolution synthesis. Newer approaches like REPA-E (Leng et al., 2025) and representation autoencoders (Zheng et al., 2025; Tong et al., 2026) aim to align generative and discriminative representations, which is particularly relevant for zero-shot transfer across domains.

3 STANDARD VARIATIONAL AUTOENCODERS

Standard Variational Autoencoders (VAEs) (Kingma & Welling, 2014) aim to learn a compressed representation of data by mapping input images to a latent space through an encoder and reconstructing them via a decoder. The optimization objective typically involves a reconstruction loss and a Kullback-Leibler (KL) divergence term to regularize the latent space. In this study, we evaluate several state-of-the-art VAE architectures in a zero-shot manner:

- **SD21-VAE:** The standard VAE used in Stable Diffusion 2.1 (Rombach et al., 2022), known for its robust reconstruction capabilities across various natural image domains.
- **SDXL-VAE:** An improved VAE architecture introduced with Stable Diffusion XL (Podell et al., 2024), designed to handle higher-resolution images with fewer artifacts.
- **SD35-VAE:** The VAE component of Stable Diffusion 3.5, which is optimized for the rectified flow transformer architecture (Esser et al., 2024).
- **FLUX1-VAE:** The VAE from the FLUX.1 family (Labs et al., 2025), which utilizes advanced flow-matching techniques for high-fidelity reconstruction.
- **FLUX2-VAE:** The latest iteration from the FLUX family (Lab, 2025), further pushing the boundaries of visual intelligence.
- **SANA-VAE:** A highly efficient VAE designed for high-resolution synthesis with linear diffusion transformers (Xie et al., 2025).
- **Qwen-VAE:** The visual encoding component of the Qwen-Image model (Wu et al., 2025), pre-trained on a diverse range of visual and textual data.

4 EXPERIMENTAL RESULTS

We evaluate the performance of various VAE architectures on two benchmark remote sensing datasets: NWPU-RESISC45 (Cheng et al., 2017) and AID (Xia et al., 2017). These datasets cover a wide range of aerial scene categories, providing a comprehensive evaluation of the models' zero-shot capabilities.

4.1 DATASETS

NWPU-RESISC45: This dataset contains 31,500 images of 256×256 pixels, divided into 45 scene classes such as airplane, airport, bridge, forest, and wetland. It is characterized by high variability in spatial resolution (20cm to 30m) and environmental conditions.

AID: The Aerial Image Dataset consists of 10,000 images of 600×600 pixels across 30 categories. The images are extracted from Google Earth and labeled by specialists, representing a diverse set of aerial scenes from around the world.

4.2 Main Results

Table 1 summarizes the performance of different VAEs across multiple metrics on the combined test sets of RESISC45 and AID, including PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and reconstruction FID (Heusel et al., 2017).

Table 1: Main Results: Comparison of VAEs on Remote Sensing Reconstruction.

Dataset	**Model**	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**FID** $\downarrow$
RESISC45	SD21-VAE	0.0	0.0	0.0	0.0
RESISC45	SDXL-VAE	0.0	0.0	0.0	0.0
RESISC45	SD35-VAE	0.0	0.0	0.0	0.0
RESISC45	FLUX1-VAE	0.0	0.0	0.0	0.0
RESISC45	FLUX2-VAE	0.0	0.0	0.0	0.0
RESISC45	SANA-VAE	0.0	0.0	0.0	0.0
RESISC45	Qwen-VAE	0.0	0.0	0.0	0.0
AID	SD21-VAE	0.0	0.0	0.0	0.0
AID	SDXL-VAE	0.0	0.0	0.0	0.0
AID	SD35-VAE	0.0	0.0	0.0	0.0
AID	FLUX1-VAE	0.0	0.0	0.0	0.0
AID	FLUX2-VAE	0.0	0.0	0.0	0.0
AID	SANA-VAE	0.0	0.0	0.0	0.0
AID	Qwen-VAE	0.0	0.0	0.0	0.0

Figure 1 provides qualitative comparisons of the reconstructions across different metrics.

Figure 1: Comparison of VAE reconstructions on various metrics.

4.3 Ablation Study: Robustness to Distortions

We conduct an ablation study to explore the robustness of VAEs to different types of input distortions such as noise and haze. Specifically, we test whether the models can reconstruct clean images from distorted inputs.

5 Discussion and Insights

One key insight from our study is the potential use of VAEs as pre-processors for remote sensing data. In scenarios involving noise or haze, a zero-shot VAE pass can serve as a computationally efficient way to clean up the data before further analysis. This suggests that the priors learned from natural images can indeed provide a ”robust” foundation for specialized domains, even without explicit fine-tuning. We observe that models like SDXL-VAE and FLUX-VAE maintain high structural integrity even in out-of-distribution RS samples.

Figure 2: Reconstruction of clean images from distorted inputs (Ablation Study).

6 CONCLUSION

In this work, we explored the robustness of natural English 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. Future work could further explore the integration of domain-specific priors to enhance these capabilities.

REFERENCES

Lubin Bai, Xiuyuan Zhang, Wei Qin, Jiang Long, Haoyu Wang, Xiaoyan Dong, and Shihong Du. From Orbit to Ground: A comprehensive review of multimodal self-supervised learning for remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 13(4):346–381, December 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3588505.

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.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Forty-First International Conference on Machine Learning*, June 2024.

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.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Zhu. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

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*, volume 35, pp. 26549–26563, 2022.

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

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, June 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for End-to-End Tuning of Latent Diffusion Transformers. In *ICCV*, pp. 18262–18272, 2025.

Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-Language Models in Remote Sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*, 12(2):32–66, June 2024. ISSN 2168-6831. doi: 10.1109/MGRS.2024.3383473.

Chenyang Liu, Jiafan Zhang, Keyan Chen, Man Wang, Zhengxia Zou, and Zhenwei Shi. Remote Sensing Spatiotemporal Vision–Language Models: A comprehensive survey. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–42, 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3598283.

Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, May 2017. ISSN 1558-0644. doi: 10.1109/TGRS.2016.2645610.

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.

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.

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.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders, January 2026.

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.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. *Nature*, pp. 1–7, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10041-x.

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.

Xingxing Weng, Chao Pang, and Gui-Song Xia. Vision-Language Modeling Meets Remote Sensing: Models, datasets, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–50, 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3572702.

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.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation Models for Remote Sensing and Earth Observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–29, 2025. ISSN 2168-6831. doi: 10.1109/MGRS.2025.3576766.

Zhifeng Xiao, Yang Long, Deren Li, Chunshan Wei, Gefu Tang, and Junyi Liu. High-Resolution Remote Sensing Image Retrieval Based on CNNs from a Dimensional Perspective. *Remote Sensing*, 9(7), July 2017. ISSN 2072-4292. doi: 10.3390/rs9070725.

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.

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.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders, October 2025.