EO-VAE: TOWARDS A MULTI-SENSOR TOKENIZER FOR EARTH OBSERVATION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art generative image and video models rely heavily on tokenizers that compress high-dimensional inputs into more efficient latent representations. While this paradigm has revolutionized RGB generation, Earth observation (EO) data presents unique challenges due to diverse sensor specifications and variable spectral channels. We propose EO-VAE, a multi-sensor variational autoencoder designed to serve as a foundational tokenizer for the EO domain. Unlike prior approaches that train separate tokenizers for each modality, EO-VAE utilizes a single model to encode and reconstruct flexible channel combinations via dynamic hypernetworks. Our experiments on the TerraMesh dataset demonstrate that EO-VAE achieves superior reconstruction fidelity compared to the TerraMind tokenizers, establishing a robust baseline for latent generative modeling in remote sensing.

1 INTRODUCTION

The Stable Diffusion generative model introduced by Rombach et al. (2022) was a central breakthrough in high resolution image generation. Both training and inference efficiency, as well as overall generation performance rely on a pretrained Variational Autoencoder (Kingma & Welling, 2014) that has since become a fundamental building block and idea in numerous domains like video generation (Brooks et al., 2024), weather forecasting (Nguyen et al., 2025), and newer text-to-image models (BlackForestLabs, 2025).

In contrast to RGB data, processing earth observation data poses several challenges such as non-fixed pixel value ranges, multispectral channels, sensor diversity and large data volumes. With the growing volume of available earth observation data, reaching the petabyte scale, latent modeling approaches have similarly appealing properties of reducing memory requirements and improving efficiency. Previous works of earth observation generative models like Khanna et al. (2024) and Jakubik et al. (2025) also make use of pretrained Autoencoders but have limitations. Khanna et al. (2024) use the pretrained SD-VAE model which works extremely well in the RGB domain but cannot operate on channel varying satellite imagery. In contrast, Jakubik et al. (2025) train a separate tokenizer based on the ViT-VQGAN (Yu et al., 2022) framework for each of the separate modalities contained in the Terramesh dataset (Blumenstiel et al., 2025). We instead propose EO-VAE, a single autoencoder model that can encode and reconstruct a variable number of channels conditioned on the channel wavelengths. Our experiments demonstrate superior reconstruction performance compared to the TerraMind tokenizers.

2 METHODOLOGY

2.1 DATASET

We choose the TerraMesh dataset (Blumenstiel et al., 2025) to train and evaluate our EO-VAE because this allows for a fair comparison against the TerraMind tokenizers which have been trained on the same data. We follow the same z-score normalization scheme of TerraMind but also find that there is a misalignment in the Sentinel 2 data based on a new processing mode introduced in January 2022, as well as missing data for some modalities. More information is provided in the appendix. We do experiments with both a native and a corrected data corpus for which results are provided in

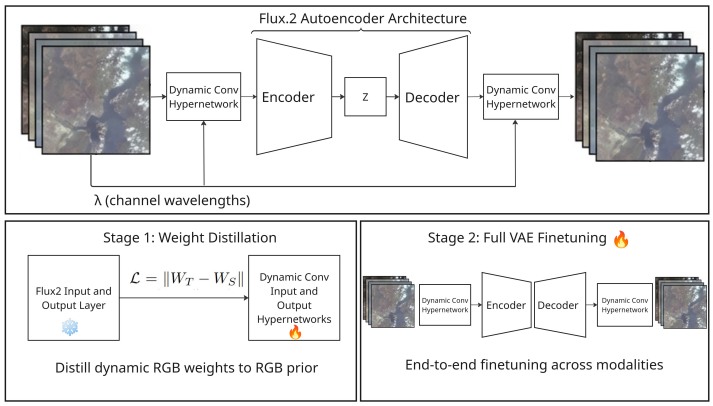


Figure 1: EO-VAE Architecture and Training Regime. The first and last convolutional layer of the Flux.2 Autoencoder architecture are replaced with dynamic convolution hypernetworks (Xiong et al., 2024). After weight distillation of the frozen Flux.2 convolutional weights, we finetune end-to-end on the multimodal TerraMesh dataset.

the appendix. We train and evaluate our model on the native Sentinel-2 L2A, and Sentinel-1 RTC modalities for which data is complete and pretrained tokenizers available. Due to storage demands we are only able to train on a subset of the TerraMesh data (first 25 shards). We generate a separate test split from the complete TerraMesh validation splits, where shards 0-6 are validation and shards 6-8 the test split and use an image size of 256x256px.

2.2 Model

We use the recently introduced Flux.2 Autoencoder (BlackForestLabs, 2025) as a base architecture and pretrained checkpoint. To accommodate a flexible number of input channels from various satellite modalities, we replace the first and last convolutional layer with dynamic hypernetworks convolutional layers that generate the convolutional weights conditioned on the channel wavelengths as proposed in the DOFA model (Xiong et al., 2024). This base model is able to reconstruct flexible number of channel combinations. Overall our training regime consists of two stages, which is depicted in Figure 1:

1. First, we use weight distillation (Lin et al., 2021) of the Flux.2 autoencoder's first and last convolutional layer (teacher) into the dynamic weight layers (student), by minimizing $\mathcal{L} = \|W_T - W_S\|$ where W_T are the teacher weights and W_S are the student weights through gradient descent optimization. We find this distillation to be crucial for faster convergence. The RGB channel provides a strong prior before exposing them to multispectral data.
2. Second, we conduct full finetuning across all three modalities via pixel-wise reconstruction loss.

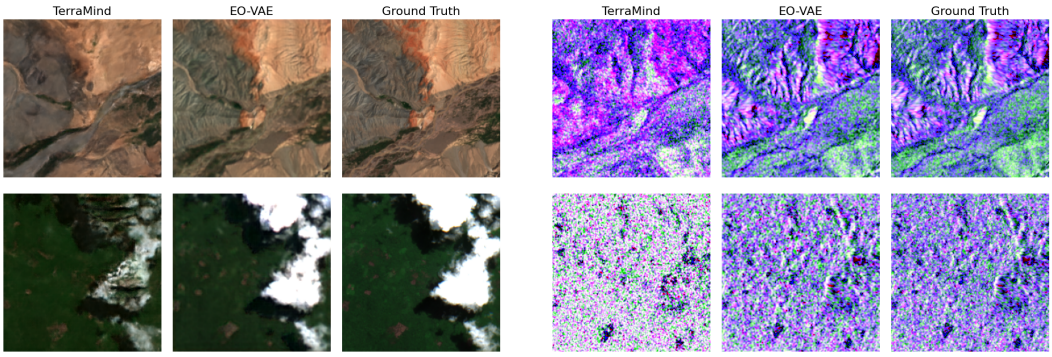
For the reconstruction objective, let $x \in \mathbb{R}^{C\times H\times W}$ denote an input satellite image with C spectral channels and corresponding channel wavelengths $\lambda = (\lambda_1, \ldots, \lambda_C)$. Let $E_{\theta_E}(\cdot;\lambda)$ and $D_{\theta_D}(\cdot;\lambda)$ denote the encoder and decoder of the Flux.2 autoencoder, where the first and last convolutional layers are replaced by dynamic hypernetwork layers that generate convolutional weights conditioned on λ. The reconstructed output is given by

$$\hat{x} = D_{\theta_D}(E_{\theta_E}(x;\lambda);\lambda) \tag{1}$$

The objective is to minimize a reconstruction loss over the training dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{rec}}(\theta_E, \theta_D) = \mathbb{E}_{(x,\lambda)\sim\mathcal{D}}\left[\ell(x, \hat{x})\right], \tag{2}$$

where $\ell(\cdot,\cdot)$ is a per-pixel reconstruction loss. We use an equally weighted Charbonier (Charbonnier et al., 1994) and multiscale structure similarity index (Wang et al., 2003) loss. All experiments were run on a 48GB NVIDIA RTX A6000.



(a) S2L2A modality reconstruction.

(b) S1RTC modality reconstruction.

Figure 2: Qualitative samples of reconstructed modalities. EO-VAE reconstructs details in both modalities better than the TerraMind tokenizers.

3 RESULTS

We evaluate EO-VAE on two capabilities: (1) high-fidelity reconstruction of multi-modal satellite imagery, and (2) its utility as a frozen latent tokenizer for downstream generative tasks. To capture complementary components of image prediction quality, we use RMSE, PSNR, SSIM, and SAM. Additionally, for the S2 modality, we also evaluate the reconstructed Normalized Difference Vegetation Index (NDVI) in terms of Mean Absolute Error (MAE) to assess physical consistency.

3.1 RECONSTRUCTION

Table 1 demonstrates that EO-VAE substantially outperforms the TerraMind Tokenizers across all metrics for both the S2L2A and S1RTC modality. On the S2L2A modality, EO-VAE achieves a PSNR of 42.80 dB, nearly 20 dB higher than TerraMind (22.95 dB). This quantitative gap aligns with the qualitative results in Figure 2, where EO-VAE preserves high-frequency details significantly better for both modalities. Additionally, our model achieves a substantial 3.5× reduction in NDVI MAE of reconstructed S2L2A images, and therefore better captures this crucial inter-band ratio.

Model	**S1RTC**				**S2L2A**				
	RMSE↓	PSNR↑	SSIM↑	SAM↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	NDVI-MAE↓
EO-VAE	**0.1401**	**37.23**	**0.9372**	**0.1601**	**0.0686**	**42.80**	**0.9720**	**0.0842**	**0.0410**
TerraMind	0.6711	23.65	0.2803	0.7285	0.7004	22.95	0.7543	0.3568	0.1403

Table 1: Reconstruction performance across modalities. EO-VAE outperforms the TerraMind tokenizers across both modalities and all metrics.

3.2 DOWNSTREAM TASK: LATENT SUPER RESOLUTION

To demonstrate the utility of our compressed latent space, we evaluate EO-VAE as a fixed tokenizer for a Latent Diffusion Model (LDM) super-resolution task on the Cross-Sensor Sen2NAIP dataset (Aybar et al., 2024). This dataset consists of spatially aligned Sentinel-2 and NAIP imagery with RGBN bands and a resolution factor of 4 from 128 to 512 pixels. We adopt a checkerboard-style geospatial split inspired by MOSAIK (Rolf et al., 2021) to ensure spatial separation between training, validation, and test regions, resulting in 2417/288/146 data points per split.

We formulate the task as a Latent Diffusion Model (LDM). Both low- and high-resolution images are encoded into the latent space using a frozen autoencoder. We train a standard UNet backbone (Ronneberger et al., 2015) to predict the high-resolution latents, conditioned on the upsampled low-resolution latents via concatenation. We use the EDM (Karras et al., 2022) diffusion model implemented in the `azula` library.[1] The EDM loss is defined as $\frac{\alpha_t^2+\sigma_t^2}{\sigma_t^2}||\mu_\phi(x_t) - x||^2$, where α and σ

[1] See https://azula.readthedocs.io/stable/.

Figure 3: Qualitative Results between EO-VAE and Flux-VAE for reconstructed super-resolution predictions.

denote the noise parameters from the variance preserving schedule used by Song et al. (2021). For inference we use DDIM sampler with 50 steps (Song et al., 2020).

This experiment highlights a critical limitation of the TerraMind tokenizers as they cannot support this task as it lacks a pretrained model for the specific RGBN modality, and training one from scratch is computationally prohibitive. In contrast, EO-VAE flexibly adapts to the 4-channel input without architectural changes. Consequently, we compare EO-VAE against two available baselines: First, a Flux.2 VAE baseline restricted to RGB-only channels (dropping NIR), as the original model cannot process multispectral data. And second, a native pixel space diffusion model trained directly on the low-and high-resolution images.

Quantitative results for the downstream super-resolution task are presented in Table 2. We observe that the EO-VAE achieves performance on par with the frozen RGB Flux.2 VAE, indicating that our multisensor adaptation effectively preserves the generative fidelity of the original architecture. Crucially, both latent-based approaches surpass the pixel-space baseline in reconstruction quality, a qualitative improvement visually confirmed in Figure 3. Beyond fidelity, Table 2 highlights the significant efficiency benefits of our approach, demonstrating that operating in the latent space yields substantial computational gains compared to pixel-space diffusion.

Model	Bands	PSNR↑	SSIM↑	RMSE↓	SAM↓	Time (ms)↓	Throughput (img/s)↑	Peak Memory (GB)↓	Params (M) Total (Diffusion)
EO-VAE	RGB+NIR	21.60	0.6234	0.0836	0.0672	389.7	2.57	1.53	106.5 (11.0)
Flux.2 VAE	RGB	**21.94**	**0.6434**	**0.0810**	**0.0609**	**374.7**	**2.67**	**1.41**	95.0 (11.0)
PIXELDiff	RGB+NIR	21.76	0.3437	0.7616	0.6905	7097.9	0.14	1.69	10.8 (10.8)

Table 2: Test set metrics on Super-Resolution experiments. EO-VAE performs on par with RGB Flux.2 model. Computational inference metrics are averaged over 50 iterations. Latent diffusion approaches are 18x more efficient measured by Time (ms) than pixel space approach.

4 Conclusion and Future Work

We have presented EO-VAE, a modality-agnostic tokenizer that bridges the gap between high-fidelity reconstruction and the spectral complexity of EO data. Our experiments demonstrate that EO-VAE significantly outperforms the existing foundation model baseline, achieving improvements across all metrics while better preserving NDVI error, despite only being trained on a much smaller subset of the overall available data. Beyond reconstruction, we validated the model's utility in a downstream generative task, where our latent space formulation enabled an $18\times$ inference speedup compared to pixel-space diffusion without sacrificing generative quality. We hope our work outlines a pathway for a collaborative effort for earth observation tokenizers that are flexible and efficient enough to handle the data requirements of EO data and provide an accessible and powerful building block for latent generative models and data compression in this domain. Future work can evolve in numerous ways: data scaling in terms of sensor variety, channel composition and resolution, further perceptual refinement, and extending to 3D architectures that handle spatio-temporal timeseries.

REFERENCES

Cesar Aybar, David Montero, Julio Contreras, Simon Donike, Freddie Kalaitzis, and Luis Gómez-Chova. Sen2naip: A large-scale dataset for sentinel-2 image super-resolution. *Scientific Data*, 11 (1):1389, 2024.

BlackForestLabs. Flux.2: Analyzing and enhancing the latent space of flux – representation comparison. `https://bfl.ai/research/representation-comparison`, 2025. [Accessed 19-12-2025].

Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabe Moreno, et al. Terramesh: A planetary mosaic of multimodal earth observation data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2394–2402, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, pp. 168–172. IEEE, 1994.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 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 *ICLR*, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, 2014. URL `https://arxiv.org/abs/1312.6114`.

Ye Lin, Yanyang Li, Ziyang Wang, Bei Li, Quan Du, Tong Xiao, and Jingbo Zhu. Weight distillation: Transferring the knowledge in neural network parameters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2076–2088, 2021.

Tung Nguyen, Tuan Pham, Troy Arcomano, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Omnicast: A masked latent diffusion model for weather forecasting across time scales. *arXiv preprint arXiv:2510.18707*, 2025.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 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.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations*, 2022.

A Appendix

A.1 TerraMesh S1-GRD Modality

At the point of submission, there was no matching validation data for the S1-GRD modality available HuggingFace repo.

A.2 TerraMesh Sentinel 2 S2L2A modality

When inspecting the Sentinel-2 L2A surface reflectance imagery in the TerraMesh dataset, we found a clear inconsistency that we found was not clearly documented. On January 25, 2022, ESA introduced a new processing baseline update (Baseline 04.00), which changes the way Sentinel-2 pixel values are represented that shifted the data range by adding a constant offset to allow negative reflectance values to be encoded. In practice this means that for imagery processed under this new baseline, the digital numbers (DNs) include values down to about –1000, whereas earlier imagery sticks to values near zero for very low reflectance and NoData. Blumenstiel et al. (2025) state that "the +1000 offset is removed from post-2022 data" and reports a value range of [0, 10000] for S2L2A. However, based on data analysis, it doesn't appear to have been applied consistently: histograms and time series of minimum pixel values show two distinct populations of Sentinel-2 values, with post-baseline images exhibiting significant density at negative values that line up with the offset.

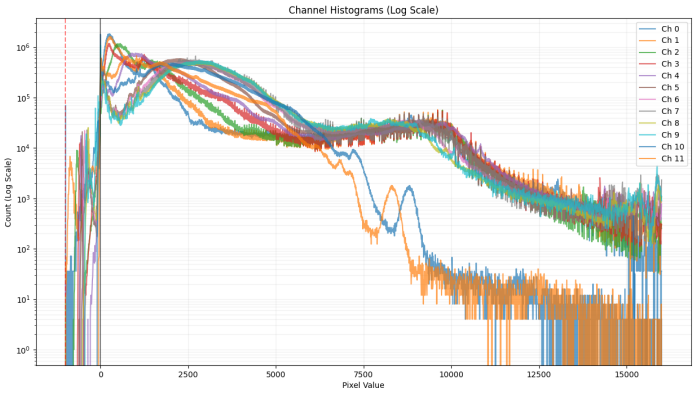


Figure 4: channelwise histogram of raw unnormalized data for the S2L2A modality, showing the range of >10000.

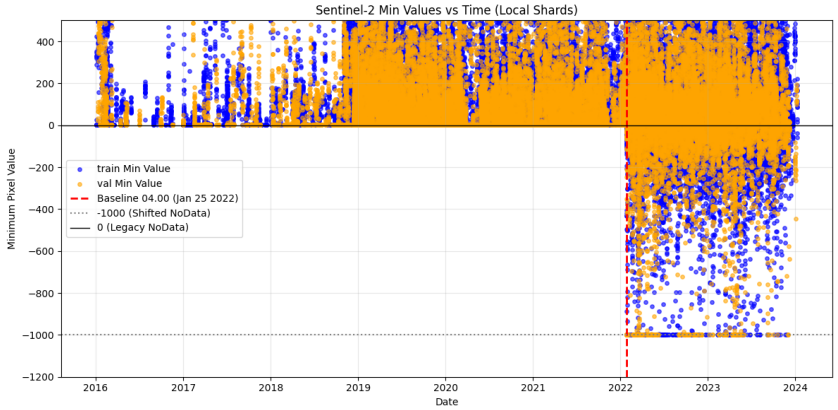


Figure 5: Minimum sample values plotted across time. The processing baseline change on January 22, 2022 becomes clearly visible.

Figure 6: Qualitative Results between EO-VAE and Flux-VAE for reconstructed super-resolution predictions

This leads to systematic differences in the basic statistics of the image data across the pre- and post-baseline change periods. The discontinuity is visible in the plots of minimum values over time as a sharp transition around the baseline change date, and the channel histograms highlight that negative values (around –1000) are concentrated in the later imagery. Including both conventions in the same dataset without adjustment therefore introduces a data inconsistency in the Sentinel-2 pixel value distributions. However, we find that correcting this bias does not change the results in a statistically significant manner.

A.3 Results with corrected S2L2A data

The following section lists results with the corrected data, where we have harmonized the data and computed new normalization statistics. Interestingly enough, the reconstruction results do not seem to improve with a EO-VAE trained on this corrected data.

Model	**S1RTC**				**S2L2A**			
	RMSE↓	PSNR↑	SSIM↑	SAM↓	RMSE↓	PSNR↑	SSIM↑	SAM↓
EO-VAE	**0.1401**	**37.23**	**0.9372**	**0.1601**	**0.0686**	**42.80**	**0.9720**	**0.0842**
EO-VAE*	0.1779	35.22	0.9010	0.1940	0.0904	37.58	0.9383	0.1135
TerraMind	0.6711	23.65	0.2803	0.7285	0.7004	22.95	0.7543	0.3568

Table 3: Reconstruction Performance Across Modalities. EO-VAE* denotes the version trained with the corrected aligned S2L2A data corpus.

In contrast, on the downstream task for super-resolution we observe, that the performance is slightly higher. In Table 4 shows

Model	Bands	PSNR↑	SSIM↑	RMSE↓	SAM↓	Time (ms)↓	Throughput (img/s)↑	Memory (GB)↓	Params (M)
EO-VAE	RGB+NIR	21.60	0.6234	0.0836	0.0672	389.7	2.57	1.53	106.5 (11.0)
EO-VAE*	RGB+NIR	**22.17**	**0.6556**	**0.0785**	**0.0584**	389.7	2.57	1.53	106.5 (11.0)
Flux.2 VAE	RGB	21.94	0.6434	0.0810	0.0609	374.7	2.67	1.41	95.0 (11.0)
PIXELDiff	RGB+NIR	21.76	0.3437	0.7616	0.6905	7097.9	0.14	1.69	10.8 (10.8)

Table 4: Test set metrics on Super-Resolution experiments. EO-VAE* denotes version trained on corrected S2L2A data corpus. Computational metrics are averaged over 50 iterations.

Fig 6 shows a qualitative example with very similar visual results.