


APPLICATION PAPER

Discrete variational autoencoders for synthetic nighttime visible satellite imagery

Mickell D. Als¹ , David Tomarov^{1,2} and Steve Easterbrook¹

¹Department of Computer Science, University of Toronto, Toronto, ON, Canada

²School of Electrical Engineering, Tel-Aviv University, Tel-Aviv, Israel

Corresponding author: Mickell D. Als; Email: mickellals@cs.toronto.edu

Received: 27 June 2025; **Revised:** 27 June 2025; **Accepted:** 30 June 2025

Keywords: deep learning; nighttime visible band; satellite remote sensing; variational autoencoders

Abstract

Visible satellite imagery (VIS) is essential for monitoring weather patterns and tracking ground surface changes associated with climate change. However, its availability is limited during nighttime. To address this limitation, we present a discrete variational autoencoder (VQVAE) method for translating infrared satellite imagery to VIS. This method departs from previous efforts that utilize a U-Net architecture. By removing the connections between corresponding layers of the encoder and decoder, the model learns a discrete and rich codebook of latent priors for the translation task. We train and test our model on mesoscale data from the Geostationary Operational Environmental Satellite (GOES) West Advanced Baseline Imager (ABI) sensor, spanning 4 years (2019 to 2022) using the Conditional Generative Adversarial Nets (CGAN) framework. This work demonstrates the practical use of a VQVAE for meteorological satellite image translation. Our approach provides a modular framework for data compression and reconstruction, with a latent representation space specifically designed for handling meteorological satellite imagery.

Impact Statement

This article introduces a discrete variational autoencoder (VQVAE) method for generating nighttime visible satellite imagery, demonstrating that high-quality generations can be achieved without relying on UNet architectures. By learning a versatile latent space, this approach provides a foundation for broader applications beyond single-task models, improving the detection of low-level atmospheric features during nocturnal hours and enabling more accurate and continuous Earth monitoring in reflectance bands beyond daylight.

1. Introduction

Meteorological satellites observe emitted, reflected, and scattered electromagnetic waves from within the Earth's atmosphere. Over the past decades, these satellites—particularly geostationary ones—have revolutionized weather forecasting by enabling near-continuous monitoring of atmospheric conditions. With their high spatiotemporal resolution, they deliver essential data for accurate weather analysis and serve as invaluable long-term records of climatic trends and land surface changes.

Since the mid-2010s, the Geostationary Operational Environmental Satellite (GOES) series has employed the Advanced Baseline Imager (ABI) sensor for the monitoring of weather and clouds in the Pacific and Atlantic regions across 16 spectral bands (Schmit et al., 2017). The ABI has since become the benchmark for many contemporary meteorological radiometer imagers (Park et al., 2021). Despite

its advancements, a significant limitation remains: six of its spectral bands depend on solar reflectance, rendering them ineffective during nighttime. This creates a gap in nighttime observation, restricting the use of visible satellite imagery (VIS), essential for accurate low-level cloud tracking and weather prediction (Conway, 1997), to daylight hours. Recent advancements in generative modeling present promising avenues to tackle this challenge. However, existing approaches (Kim et al., 2019; Harder et al., 2020; Kim et al., 2020; Park et al., 2021; Cheng et al., 2022; Han et al., 2022; Chirokova et al., 2023; Yan et al., 2023; Pasillas et al., 2024; Yao et al., 2024) have primarily focused on a limited subset of model architectures and training strategies, leaving significant room for further exploration and innovation.

Most previous efforts (Kim et al., 2019; Harder et al., 2020; Kim et al., 2020; Park et al., 2021; Cheng et al., 2022; Han et al., 2022; Yao et al., 2024) have predominantly focused on training Pix2Pix models (Isola et al., 2018) for this task. Pix2Pix utilizes a U-Net architecture (Ronneberger et al., 2015) trained adversarially with a PatchGAN discriminator (Isola et al., 2018). This approach is efficient for training, as the skip connections between corresponding encoder and decoder layers enable effective backpropagation with minimal loss and mitigate vanishing gradient issues. These skip connections also introduce an explicit spatial correspondence bias between input and target pixels, which is particularly beneficial for tasks like image translation and segmentation. However, to develop a more generalizable model for this task, it is essential to forego these skip connections and design a more robust latent representation. We conduct an initial investigation into constructing a robust latent representation without explicitly enforcing spatial correspondence as in Pix2Pix; instead, the correspondence is learned implicitly by the model.

In this work, we evaluate the potential of discrete variational autoencoders trained adversarially for generating synthetic visible imagery from GOES West Mesoscale data, particularly under nighttime conditions where traditional optical sensors have limitations. We also introduce a straightforward processing pipeline for converting ABI sensor L2 data from NetCDF files into usable NumPy arrays. This pipeline includes synthetic green band construction and tailored normalization to address the unique spectral characteristics of the sensor. Furthermore, our experiments explore the construction of a context-rich latent space that encapsulates atmospheric structure and provide valuable insights into how VQGAN-based synthesis performs relative to alternative methods such as Pix2Pix (Isola et al., 2018) and MLPs. These findings establish a foundation for future work in satellite image processing, including potential applications in super-resolution and latent diffusion to enhance the overall utility of meteorological data.

2. Background and related work

2.1. The advance baseline imager (ABI)

The ABI, developed by NASA in collaboration with NOAA, delivers continuous Earth observations from the geostationary orbit of the GOES family of satellites (Schmit et al., 2017). It captures imagery across 16 spectral wavelengths, as detailed in Table A1, providing a high spatiotemporal resolution that surpasses earlier meteorological Earth observation satellites (Menzel and Purdom, 1994). Vandal et al. (2022) emphasize that the transferability between different satellite sensors remains a limitation for current deep learning approaches in this domain. The ABI's broad spectral range and its role as a blueprint for subsequent imagers (Park et al., 2021) may help address this challenge. NOAA's public release of the extensive GOES dataset can facilitate the training of more generalizable deep learning models using imagery captured by the ABI and driving advancements in the field.

Expanding on the ABI's capabilities, the imager can operate in multiple scan modes to address various observational requirements. In scan mode 6, the current default, the ABI captures a full-disk image of Earth every 10 minutes and images of the contiguous United States (CONUS) every 5 minutes. Additionally, it acquires two mesoscale images every minute, each covering approximately 1,000 km² (Schmit and Gunshor, 2024). These mesoscale images target regions of meteorological significance, with locations determined by NOAA staff and adjusted hourly or daily based on evolving conditions.

This ability to focus on specific areas ensures continuous monitoring of critical weather phenomena such as developing cyclones and thunderstorms, enabling more precise and timely forecasting. Owing to their localized detail and meteorological interest, these images are particularly well-suited for this work and form the basis of our data selection.

2.2. Convolutional approaches to synthetic VIS generation

Kim et al. (2019) conducted the first experiments using deep convolutional networks to generate synthetic nighttime reflectance by employing the Pix2Pix architecture (Isola et al., 2018) to transform longwave infrared (LWIR) imagery into red-band visible (VIS) data. Their experiments revealed that longwave infrared (LWIR) bands in the $10.3\mu\text{m} - 11.3\mu\text{m}$ range exhibit the highest correlation with the visible red band ($0.55\mu\text{m} - 0.80\mu\text{m}$), establishing a foundational baseline for this task. While their model demonstrates the potential of generative modeling for this task, its performance was hindered by seasonal variability, particularly due to changes in the spatial distribution of clouds and low-level fog. A further shortcoming of the Pix2Pix model was its difficulty in distinguishing cloud cover from bare land at night, leading to frequent misclassification of cold desert or sparsely vegetated regions as cloud-filled (Han et al., 2022).

By incorporating four additional LWIR bands (Kim et al., 2020), advanced the initial approach and helped formalize the use of multi-band inputs as a standard strategy for improving generation fidelity. Expanding the multi-band approach (Cheng et al., 2022) integrate ERA5-derived scale and depth priors to enhance RGB VIS generation. While this improves image quality, it introduces computational overhead by increasing input size to 83 channels and requiring a Squeeze-and-Excitation preprocessing step (Hu et al., 2019). More recently, Yao et al. (2024) introduced solar positioning parameters, azimuth and zenith angles, into the generation process, resulting in synthetic red VIS imagery with highly realistic lighting that reflects variations across different times of day. Diverging from prior efforts that focus on data enrichment, our work centers on architectural design, specifically the development of a context-rich latent space. To this end, we employ a multi-band approach with three input bands and assess the performance of a Vector Quantized Variational Autoencoder (VQVAE) (van den Oord et al., 2018; Esser et al., 2021), trained adversarially.

2.3. Multi-layer perceptron approaches to synthetic VIS generation

While (Harder et al., 2020) explore convolutional generative models, including UNet and UNet+, and identify Pix2Pix CGAN as the most effective for realistic generations, alternative approaches take a fundamentally different route by modeling the task as a regression problem using MLPs. NightDNN (Yan et al., 2023) employs six LWIR wavelength bands that capture key atmospheric features such as water vapor, cloud cover, and surface temperatures. These bands are flattened into 1D arrays and subsequently passed through a multi-layer perceptron (MLP). Since pixel-by-pixel training tends to lose surface texture and spatial structure, the authors mitigate this by training a second, shallow MLP to encode geographic features from NASA's Blue Marble images. This additional input helps the model retain the local and global details lost by focusing solely on minimizing pixel-wise mean square error (MSE). The method is applied to full-disk imagery, which benefits from the added spatial encoding to maintain high-quality image generation. The model outputs a 1D array of values, which is reshaped into an RGB image. This technique effectively removes solar lighting artifacts, including glint, and shows how modeling pixel interactions can also lead to high-quality results.

In building on this prior work, Pasillas et al. (2024) introduces a deep MLP framework, called machine learning nighttime visible imagery (ML-NVI). This model is trained on Day Night Band (DNB) data from the Visible Infrared Imaging Radiometer Suite (VIIRS), producing consistent DNB-derived imagery with accurate cloud representation throughout the lunar cycle, thus enabling robust nighttime cloud detection. The study demonstrates that MLPs are effective in generating consistent imagery by modeling pixel-level interactions, as opposed to relying solely on the local interactions

facilitated by convolutions. However, the “curse of dimensionality” makes MLP-based approaches impractical for satellite imagery, given the large number of pixels typically involved. In response, our approach integrates pixel-level interactions using a self-attention layer (Vaswani et al., 2017) within the VQVAE framework at the encoder’s lowest resolution.

3. Methodology

This section outlines our experimental approach by detailing how we process the data, design our model architecture for image-to-image translation, and implement the training protocol with focused hyperparameter studies. The methodology is organized into three main parts: data processing, model architecture, and training strategy.

3.1. Data processing

The study uses the ABI-L2-MCMIPM product from the GOES West satellite (NOAA, n.d.). This dataset comprises co-registered satellite images with a spatial resolution of 2 km^2 , resampled to 500×500 pixels. The data, originally provided in NetCDF format, is converted to NumPy (`.npy`) arrays for compatibility with deep learning frameworks. Min-max normalization is applied using valid minimum and maximum values for the ABI sensor at each wavelength band, ensuring consistent scaling. Since the ABI sensor lacks a dedicated green band, a synthetic green band is generated using the Cooperative Institute for Meteorological Satellite Studies formula:

$$\text{Green_Band} = 0.45(\text{Red_Band}) + 0.1(\text{Veggie_Band}) + 0.45(\text{Blue_Band}). \quad (1)$$

To reduce dataset size and computational complexity, a subset of spectral bands is chosen based on prior research (Park et al., 2021). The selected bands are 1, 2, 3, 8, 9, 10, 11, 13, 14 (see Table A1), along with the constructed green band, forming a 3D array of shape $10 \times 500 \times 500$. The ground truth (target domain \mathbf{Y}) comprises RGB images from the Red, Green, and Blue bands, and the input domain \mathbf{X} consists of three bands selected from $\{8, 9, 10, 11, 13, 14\}$. The study evaluates various band combinations to optimize the translation task.

After preprocessing, each sample includes:

- *Input domain*: Three selected LWIR bands.
- *Target domain*: Corresponding RGB images from VIS bands.

The dataset is partitioned as follows:

- *Training*: 6,000 examples (daytime imagery between 10 am and 4 pm PST, from 2019 to 2021).
- *Validation*: 600 examples (daytime imagery between 10 am and 4 pm PST, from 2019 to 2021).
- *Testing*: 360 nighttime images (1 am–4 am PST), 356 daytime images (10 am–2 pm PST), 20 daytime land-only images, and 20 daytime ocean-only images (all testing samples collected from 2022 imagery).

3.2. Model architecture

Image-to-image translation involves transforming an image from one domain to another while preserving its essential content (Isola et al., 2018). In our case, the task involves converting satellite imagery across different spectral representations, ensuring spatial consistency and accurate domain-specific characteristics. Our approach is framed as an image-to-image translation problem where the goal is to learn a function $F: X \rightarrow Y$ that maps an input image $x \in X$ to its corresponding output $y \in Y$. To accomplish this, we leverage a combination of discrete variational autoencoders (VQVAEs) and generative adversarial networks (GANs). The following components make up our model architecture.

3.2.1. Discrete variational autoencoders (VQVAEs)

VQVAEs compress high-dimensional data into a discrete latent space, facilitating efficient image reconstruction and translation. Their architecture consists of three main parts:

Encoder. The encoder $E(x)$ maps the input x to a lower-dimensional latent representation \hat{z} . Unlike standard VAEs that learn continuous latent representation, the VQVAE encoder outputs are served to a quantization layer which discretizes the latent representation $\hat{z} = E(x)$.

Quantizer. The quantizer discretizes the encoder's output by mapping each index vector to the closest one among K predefined embedding vectors in the space Z . This is formalized as

$$z_q = q(\hat{z}) := \left(\arg \min_{z_k \in Z} \|\hat{z}_{ij} - z_k\| \right). \quad (2)$$

Since the quantization step is non-differentiable, we use the straight-through estimator to approximate gradients and facilitate end-to-end training.

Decoder. The decoder reconstructs the input image from the quantized latent variable z_q . The overall training objective for the VQVAE integrates three loss components: reconstruction loss, vector quantization loss (aligning the embedding vectors), and commitment loss (ensuring the encoder commits to specific embeddings). The combined loss is given by

$$L_{VQ}(E, G, Z) = \|x - \hat{x}\|_2^2 + \|\text{sg}[E(x)] - z_q\|_2^2 + \|\text{sg}[z_q] - E(x)\|_2^2, \quad (3)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

3.2.2. Generative adversarial networks (GANs)

GANs are employed to enhance the perceptual realism of the generated images. The discriminator $D(x)$ distinguishes between real and generated examples, while the generator $G(\hat{z})$ produces realistic outputs. The GAN loss is defined as

$$L_{GAN}(\{E, G, Z\}, D) = \mathbb{E}_x[\log D(x) + \log(1 - D(\hat{x}))]. \quad (4)$$

Following the framework of Esser et al. (2021), adversarial training is combined with the VQVAE losses to yield the VQGAN model, with the overall learning objective formulated as

$$Q^* = \arg \min_{E, G, Z} \max_D \mathbb{E}_{x \sim p(x)} [L_{VQ}(E, G, Z) + \lambda L_{GAN}(\{E, G, Z\}, D)], \quad (5)$$

where λ balances the perceptual reconstruction and adversarial losses. A schematic of the complete VQGAN model used in this pipeline is presented in Figure 1.

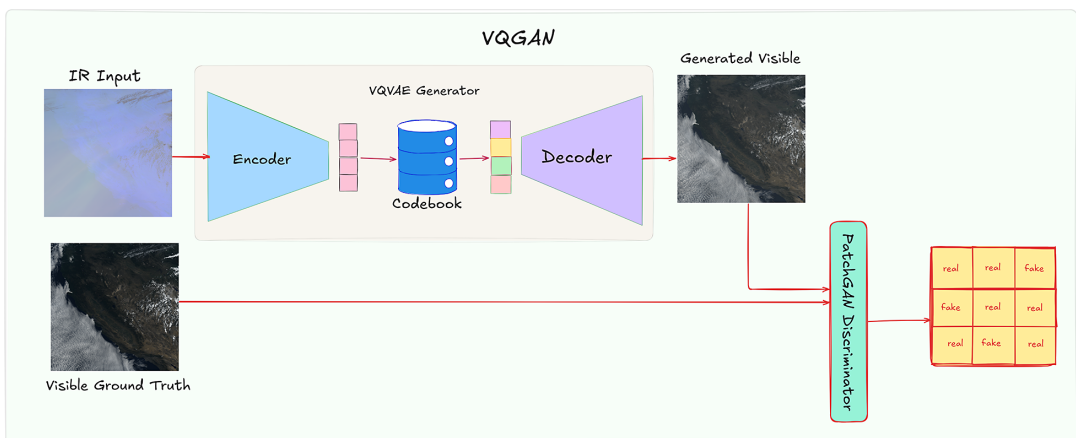


Figure 1. Diagram of the VQGAN model for image-to-image translation in the training pipeline.

3.3. Training protocol and hyperparameter tuning

Our training strategy follows the framework introduced by Esser et al. (2021), focusing on constructing a robust latent space for effective image-to-image translation. Key aspects include the following:

3.3.1. Overall training setup

- *Data preparation:* Training images are resized to 512×512 and normalized to the $[-1, 1]$ range.
- *Baseline configuration:* For fair comparison, the baseline employs the default VQGAN settings with a codebook size of 1024, 2 residual blocks, embedding dimension $Z^D = 256$, discriminator start at 10,000 steps, and discriminator loss weight of 0.8. This baseline achieves LPIPS=0.537, PSNR = 12.329, RMSE=0.244, and SSIM=0.240.
- *Training duration:* All models are trained for 100 epochs with a learning rate of 4.5×10^{-6} .

3.3.2. Training warm-up

Initial experiments set the latent embedding dimension $Z^D = 4$, utilize 1,024 embeddings, and fix the discriminator loss weight at 0.8. We varied the discriminator activation step between 2,500 and 12,000 steps and tested encoder-decoder depths ranging from 2 to 5 residual blocks. While the 2-block model showed early promise with a short warm-up, it suffered from sharp spikes in perceptual loss once the discriminator was introduced. In contrast, the 3-block model exhibited more stable training, especially when paired with a longer warm-up. Deeper variants (4 and 5 blocks) provided no significant improvements. Based on these findings, we adopted the 3-block configuration with a longer generator warm-up for most experiments, striking a balance between training stability and representation quality.

Interestingly, this regime also revealed an evaluation-time discrepancy: the model's output during training appeared visually inverted compared to the test-time output. Further inspection suggests this may stem from band misalignment during training. This issue is known to occur in models like VQGAN and Pix2Pix when handling multi-channel outputs, yet it is seldom highlighted in training literature, where such inversion artifacts are often overlooked or undocumented. While the baseline model continued to produce inverted outputs at test time, our final trained models successfully generated correctly ordered RGB imagery.

3.3.3. Codebook size exploration

We explore various codebook sizes to assess their impact on performance, with the discriminator start fixed at 5,000 steps and the latent embedding dimension $Z^D = 4$. The number of residual blocks in the encoder/decoder is increased from 3 to 4, while evaluating codebook sizes $K \in \{1024, 2048, 4096, 8192\}$. While larger codebooks are theoretically better at capturing a broader range of features, results show that the 2,048 and 4,096 variants perform comparably across most metrics, with 4,096 offering moderate improvements in SSIM (see Table A2). Interestingly, the deepest model (4 blocks) with the largest codebook (8,192) did not outperform the 2,048 baseline, which we attribute to insufficient generator warm-up.

3.3.4. Model depth and discriminator loss weight

To further stabilize training and improve image quality, we reduce the discriminator loss weight to 0.2 and delay its activation to 50,000 steps. This adjustment allows the generator to focus on learning meaningful representations without being prematurely penalized by the discriminator. Unlike earlier configurations, this setup results in a steady decrease in perceptual loss rather than the sharp spikes observed previously (Figure A3). These results highlight the benefits of pairing longer warm-up periods with a reduced discriminator influence, especially in deeper models trained with larger codebooks.

3.3.5. Varying input bands

Building on the stabilized training regime and insights from prior ablation studies, we further evaluate the effect of varying LWIR wavelength bands and embedding dimensions $Z^D \in \{4, 6\}$. These experiments

Table 1. Final configurations of evaluated models

Run	Z^D	No. of embeddings	z-channel	ResBlock	Disc start	Disc Wt	Steps
Baseline	256	1024	256	2	10,000	0.8	100,000
Embed 4	4	8192	256	2	70,000	0.2	100,000
Embed 6	6	8192	256	2	70,000	0.2	100,000

Note. All models were trained for 100 epochs with consistent architecture and loss settings.

retain two residual blocks in the encoder-decoder architecture and introduce water vapor imagery as an additional input channel. The discriminator activation step is extended to 70,000 steps to allow the generator more time to learn stable representations.

Although two band combinations were considered, Bands 11, 13, and 14 versus Bands 10, 11, and 14, all results reported here are from the latter configuration. This combination was chosen based on its more consistent output quality during exploratory testing. The 4- and 6-dimensional embeddings both yield high-quality outputs (Figure A2), with no significant advantage observed for the higher-dimensional case aside from a longer convergence time. Notably, the inclusion of water vapor imagery substantially improves image realism (Table 1).

4. Results and discussion

4.1. Evaluation of land, ocean, and nighttime image generation

The top two models were evaluated across land and ocean imagery to assess their performance over different surface types. Interestingly, land imagery consistently outperformed ocean imagery across all evaluation metrics (Table 2), contrary to initial expectations. This discrepancy may be due to a bias in the training data as most mesoscale scenes were centered over the western United States, comprising predominantly land. Consequently, the model may have become more adept at reconstructing daytime land scenes due to their relative abundance during training. Visual comparisons underscore this difference in performance (Figure A1).

Despite the lower numerical scores for ocean scenes, these examples revealed promising qualitative results. A key success lies in the model's ability to reconstruct subtle atmospheric features from longwave infrared (LWIR) inputs. Features like low-level fog—which typically blend into the background in LWIR—were clearly generated in the visible output. Similarly, the model successfully highlighted low-altitude cumulus clouds associated with the early stages of tropical cyclones. While the model did not fully capture the structured vortex of a developing storm, it did reconstruct the general circular cloud pattern and a perceptible center of rotation, showing promise in its ability to infer complex weather dynamics from thermal data (Figure A1).

Nighttime evaluations further confirmed the model's strengths and limitations. Over ocean regions, the model retained high levels of detail and contrast (Figure A2). Land imagery, however, often defaulted to overwhelmingly white land cover (Figure A4). This behavior is consistent with Han et al. (2022) and likely arises because land surfaces cool more rapidly than oceans at night. The model appears to infer cloud presence over cooled land surfaces based on their temperature similarity, suggesting it is learning surface temperature rather than implicitly modeling cloud or atmospheric structure.

These findings point to a deeper limitation in current image-to-image translation models when applied to geophysical data: they lack an understanding of the physical world. The persistent nighttime artifact over land, for instance, signals the model's reliance on statistical correlations rather than physical reasoning. This highlights the need for incorporating physically-informed representations. Embedding techniques such as Sphere2Vec (Mai et al., 2023), or broader frameworks that acknowledge the spatiotemporal and physical structure of atmospheric data, may offer a way forward. Treating satellite imagery as generic natural images overlooks domain-specific constraints—and this evaluation underscores the need for dedicated methods tailored to geoscientific image modeling.

Table 2. Performance metrics comparison

Model	Config	Scene	LPIPS↓	PSNR↑	RMSE ^a ↓	SSIM↑
Ours	Dim = 4	Land/Ocean	0.251	21.788	0.084	0.637
Ours	Dim = 6	Land/Ocean	0.268	21.244	0.090	0.614
Ours	Dim = 4	Land	0.258	22.108	0.081	0.635
Ours	Dim = 6	Land	0.275	21.543	0.086	0.605
Ours	Dim = 4	Ocean	0.330	18.956	0.117	0.525
Ours	Dim = 6	Ocean	0.328	18.856	0.121	0.522
Previous work						
Yao et al. (2024)	Pix2Pix	Tropical cyclones	–	28.3	0.043	0.885
Kim et al. (2019)	Pix2Pix	Winter	–	–	0.129	–
	Pix2Pix	Summer	–	–	0.145	–
	Pix2Pix	–	–	–	0.105	–
Kim et al. (2020)	Pix2Pix	–	–	–	0.11	0.77
Harder et al. (2020)	U-Net	–	–	–	0.09	0.85
	U-Net++	–	–	–	0.07	0.86
	Pix2Pix	Red Band	–	–	0.061	–
Han et al. (2022)		Green Band	–	–	0.050	–
		Blue Band	–	–	0.047	–
Cheng et al. (2022)	Pix2Pix	–	–	25.5	0.082	0.480
Yan et al. (2023)	DNN	–	–	–	8.38	–

Note. Best-performing metrics (highlighted in yellow) for our model evaluated on combined land/ocean, land-only, and ocean-only scenes. Best overall performance per category is highlighted in green.

^aRMSE values are not directly comparable across methods due to differences in data domains, pre-processing pipelines, and evaluation setups.

4.2. Input bands and the learned latent space

The performance improvements observed with the inclusion of a water vapor imagery channel may stem from the additional vertical distribution information it provides. Specifically, the low-level water vapor band introduces information from a distinct atmospheric layer, effectively acting as a weak depth prior for the model (Figure A5). While the other LWIR channels offer integrated signals across the full atmospheric column, water vapor imagery tends to be more level-specific, allowing the model to better localize features in its latent representation.

Despite still exhibiting some of the same challenges as the baseline, such as color inversion in reconstructed VIS output, the qualitative detail in the generated images was noticeably improved. Fine-grained structures and spatial coherence were better preserved, suggesting that the multi-level atmospheric context helped the model disentangle relevant features more effectively. A comparison between the input bands, baseline model, ground truth, and final generation is shown in Figure A6.

These results indicate the model learning a robust, generalizable representation of the earth and atmosphere that is, to some extent, independent of the specific input bands presented. This suggests the potential for a refined approach, where the latent space could be designed to be invariant to the wavelength inputs. If we view the different bands as multimodal inputs, we can then imagine that the learned latent space would be conditioned on all these multimodal inputs. For tasks such as latent diffusion, it could be possible to train a model on a single satellite with a sufficiently broad variety of wavelengths and transfer the learned latent space to other tasks. In such cases, the only requirement would be to train a new encoder model on the specific inputs for a given task. The model would replace the inputs with their nearest neighbor vector embeddings from the learned latent space, allowing the model to operate fully in the latent space for objectives like diffusion or super-resolution. This approach would make computation more efficient, as it would leverage a common embedding space, enabling future research to focus on improving interpretability across multiple models that share this latent space.

5. Conclusion and limitations

In this work, we evaluated the potential of VQVAEs to generate synthetic RGB visible imagery from LWIR data during nighttime. Our model successfully captures key spatial features and even high-frequency details, such as small surface clouds, but several challenges remain.

A primary limitation we observed is the mischaracterization of colder nighttime ground surfaces as cloud-filled. This shortcoming suggests that the current CNN-based approaches may lack sufficient encoding of spatial information. In future work, incorporating explicit positional encodings (e.g., Sphere2Vec; Mai et al., 2023) or adopting a multimodal approach, where additional inputs like SAR imagery inform the model, may improve differentiation between land and cloud features.

Another critical challenge lies in our data processing pipeline. Although the conversion of raw GOES data from NetCDF format into deep learning—ready NumPy arrays was necessary for training, the resulting files were extremely large and inefficient. These issues, combined with the limitations of NetCDF for random access, significantly constrained training scalability. As a result, we were forced to use only a small fraction, roughly 6,000 out of over 56,000 available NetCDF files, due to storage and compute constraints. This highlights a broader bottleneck in the field: the absence of pre-processed, full-spectrum cloud datasets tailored for deep learning applications.

Recent work (Pasillas et al., 2024) has emphasized that, while quantitative metrics provide some insight, they often fall short in assessing spatial coherence and operational utility. In this context, qualitative evaluation by domain specialists, who are the end users of such imagery, is essential. These expert assessments can reveal subtleties and real-world usability that automated metrics may overlook, especially in safety, critical applications like weather forecasting or disaster response. Nonetheless, the lack of formal spatial evaluation metrics remains an open issue. Metrics such as Moran's I, Geary's C, or Kullback–Leibler divergence along with consistent comparisons against prior models like (Harder et al., 2020), whose evaluations were limited to SSIM and RMSE, are needed to more rigorously assess spatial bias and distributional fidelity. We consider this an important direction for future work, as such evaluations are currently underutilized across the literature.

Overall, our study provides valuable insights into the challenges and potential of applying VQVAEs for nighttime VIS synthesis from LWIR data. The identified limitations and future directions lay the groundwork for developing more physically informed, multimodal approaches in geoscientific image modeling.

Open peer review. To view the open peer review materials for this article, please visit <https://doi.org/10.1017/eds.2025.10015>.

Acknowledgements. The authors are grateful for the technical assistance of the Data Analytics for Canadian Climate Services (DACCS) for allowing the use of the Marble platform for data storage and exploration during this project. The authors would also like to thank the School of Electrical Engineering and the Lowy International School of Tel-Aviv University, for allowing David Tomarov to participate in the student exchange program at the University of Toronto, where this research was conducted.

Author contribution. Conceptualization: M.A.; Methodology: M.A. and D.T.; Data curation: M.A.; Data visualization: M.A., D.T.; Writing—original draft: M.A., D.T.; Writing—review and editing: M.A., D.T., S.E.; Supervision: S.E. All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. Our code publicly available at *nighttime_vqgan*. Note that the model code is identical to that from Esser et al. (2021); the primary modifications reside in the custom dataloader and data processing scripts included in our repository. The experimental data originates from NOAA Geostationary Operational Environmental Satellites (GOES) 16, 17 & 18 and is accessed via the University of Toronto's Marble RedOak THREDDS server. However, due to ongoing additional processing of the GOES dataset, the final processed data is not yet publicly available but will be provided at a later date on RedOak.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada with grant number RGPIN-2019-07042.

Provenance statement. This article was accepted into the Climate Informatics 2025 (CI2025) Conference. It has been published in Environmental Data Science on the strength of the CI2025 review process.

References

- Cheng W, Li Q, Wang ZG, Zhang W and Huang F (2022) Creating synthetic night-time visiblelight meteorological satellite images using the GAN method. *Remote Sensing Letters*. 13(7), 738–745. <https://doi.org/10.1080/2150704X.2022.2079016>.
- Chirokova G, Knaff JA, Brennan MJ, DeMaria RT, Bozeman M, Stevenson SN, Beven JL, Blake ES, Brammer A, Darlow JA, DeMaria M, Miller SD, Slocum CJ, Molenar D and Hillger DW (2023) ProxyVis – A proxy for Nighttime visible imagery applicable to geostationary satellite observations. *Weather and Forecasting* 38(12), 2527–2550. <https://doi.org/10.1175/WAF-D-23-0038.1>. <https://journals.ametsoc.org/view/journals/wefo/38/12/WAF-D-23-0038.1.xml>
- Conway ED (1997) *An Introduction to Satellite Image Interpretation*. Baltimore: Johns Hopkins University Press, Apr
- Esser P, Rombach R and Ommer B (2021) *Taming Transformers for High-Resolution Image Synthesis*. arXiv:2012.09841 [cs]. <http://arxiv.org/abs/2012.09841>.
- Han K-H, Jang J-C, Ryu S, Sohn E-H and Hong S (2022) Hypothetical visible bands of advanced meteorological imager Onboard the geostationary Korea multi-purpose satellite -2A using data-to-data translation. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing* 15, 8378–8388. <https://doi.org/10.1109/JSTARS.2022.3210143>. <https://ieeexplore.ieee.org/document/9904830/>.
- Harder P, Jones W, Lguensat R, Bouabid S, Fulton J, Quesada-Chacón D, Marcolongo A, Stefanović S, Rao Y, Manshausen P and Watson-Parris D (2020) *NightVision: Generating Nighttime Satellite Imagery from Infra-Red Observations*. arXiv: 2011.07017 [cs, eess]. <http://arxiv.org/abs/2011.07017>.
- Hu J, Shen L, Albanie S, Sun G and Wu E (2019) *Squeeze-and-Excitation Networks*. arXiv: 1709.01507 [cs.CV]. <https://arxiv.org/abs/1709.01507>.
- Isola P, Zhu J-Y, Zhou T and Efros AA (2018) *Image-to-Image Translation with Conditional Adversarial Networks*. arXiv: 1611.07004 [cs]. <http://arxiv.org/abs/1611.07004>.
- Kim K, Kim J-H, Moon Y-J, Park E, Shin G, Kim T, Kim Y and Hong S (2019) Nighttime reflectance generation in the visible band of satellites. *Remote Sensing* 11(18), 2087. <https://doi.org/10.3390/rs11182087>.
- Kim J-H, Ryu S, Jeong J, So D, Ban H-J and Hong S (2020) Impact of satellite sounding data on virtual visible imagery generation using conditional generative adversarial network. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing* 13, 4532–4541. <https://doi.org/10.1109/JSTARS.2020.3013598>. <https://ieeexplore.ieee.org/document/9154540/>.
- Mai G, Xuan Y, Zuo W, He Y, Song J, Ermon S, Janowicz G and Lao N (2023) Sphere2Vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing* 202, 439–462. <https://doi.org/10.1016/j.isprsjprs.2023.06.016>.
- Menzel WP and Purdom JFW (1994) Introducing GOES-I: The first of a new generation of geostationary operational environmental satellites. *Bulletin of the American Meteorological Society* 75(5), 757–782.
- NOAA (n.d.) *NOAA Geostationary Operational Environmental Satellites (GOES) 16, 17 & 18 – Registry of Open Data on AWS – registry.opendata.aws*. <https://registry.opendata.aws/noaa-goes/>.
- Park J-E, Kim G and Hong S (2021) Green band generation for advanced baseline imager sensor using Pix2Pix with advanced baseline imager and advanced Himawari imager observations. *IEEE Transactions on Geoscience Remote Sensing* 59(8), 6415–6423. <https://doi.org/10.1109/TGRS.2020.3032732>.
- Passilas CM, Kummerow C, Bell M and Miller SD (2024) Turning night into day: The creation and validation of synthetic night-time visible imagery using the visible infrared imaging radiometer suite (VIIRS) day-night band (DNB) and machine learning. *Artificial Intelligence for the Earth Systems*. <https://doi.org/10.1175/AIES-D-23-0002.1>. <https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0002.1/AIES-D-23-0002.1.xml>.
- Ronneberger O, Fischer P and Brox T (2015) *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv: 1505.04597 [cs]. <http://arxiv.org/abs/1505.04597>.
- Schmit TJ and Gunshor MM (2024) *GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Cloud and Moisture Imagery Product*.
- Schmit TJ, Griffith P, Gunshor MM, Daniels JM, Goodman SJ and Lehair WJ (2017) A closer look at the ABI on the GOES-R Series. *Bulletin of the American Meteorological Society* 98(4), 681–698. <https://doi.org/10.1175/BAMS-D-15-00230.1>.
- van den Oord A, Vinyals O and Kavukcuoglu K (2018) *Neural Discrete Representation Learning*. arXiv:1711.00937 [cs]. <http://arxiv.org/abs/1711.00937>.
- Vandal TJ, McDuff D, Wang W, Duffy K, Michaelis A and Nemani RR (2022) Spectral synthesis for geostationary satellite-to-satellite translation. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–11. <https://doi.org/10.1109/TGRS.2021.3088686>. <https://ieeexplore.ieee.org/document/9462910/>.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is all you need. *CoRR* abs/1706.03762. arXiv: 1706.03762. <http://arxiv.org/abs/1706.03762>.
- Yan J, Qu J, An H and Zhang H (2023) Simulation of visible light at night from infrared measurements using deep learning technique. *Geocarto International* 38(1), 2227610. <https://doi.org/10.1080/10106049.2023.2227610>.
- Yao J, Du P, Zhao Y and Wang Y (2024) *Simulating Nighttime Visible Satellite Imagery of Tropical Cyclones Using Conditional Generative Adversarial Networks*. en. arXiv:2401.11679 [physics]. <http://arxiv.org/abs/2401.11679>.

A. Appendix

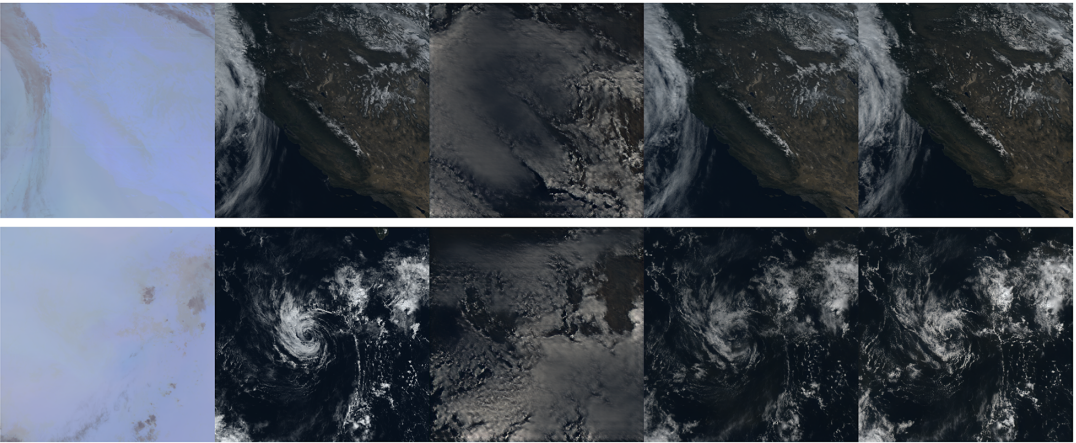


Figure A1. Comparison of model outputs for land and ocean cover. Row 1 shows land cover, while Row 2 shows ocean cover. Columns, from left to right, represent: input, ground truth (GT), baseline model, embedding dimension $Z^D = 4$, and embedding dimension $Z^D = 6$.

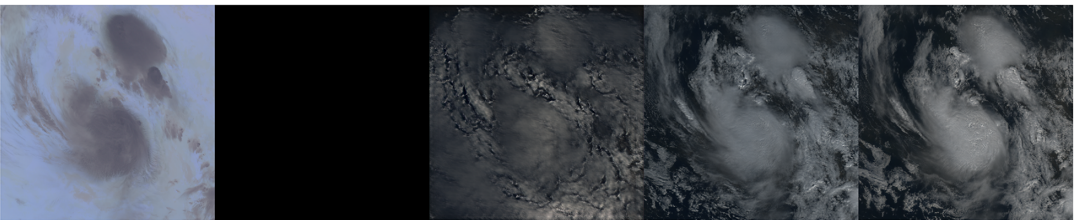


Figure A2. Comparison of model outputs for nighttime. From left to right: input, ground truth (GT), baseline model, embedding dimension 4, and embedding dimension 6.

Table A1. Spectral and spatial characteristics of GOES ABI sensor bands. Reflective bands operate during daytime; radiance bands are thermal and operate continuously

Band	Wavelength (μm)	Resolution (km ²)	Type	Descriptive name
Reflective bands				
1	0.47	1.0	Reflectance	Blue
2	0.64	0.5	Reflectance	Red
3	0.86	1.0	Reflectance	Vegetation
4	1.37	2.0	Reflectance	Cirrus
5	1.61	1.0	Reflectance	Snow/ice
6	2.24	2.0	Reflectance	Cloud particle size
7	3.90	2.0	Reflectance (daytime)/ radiance (nighttime)	Shortwave window
Infrared radiance bands				
8	6.19	2.0	Radiance	Upper-level water vapor
9	6.93	2.0	Radiance	Mid-level water vapor
10	7.34	2.0	Radiance	Low-level water vapor
11	8.44	2.0	Radiance	Cloud-top phase
12	9.61	2.0	Radiance	Ozone
13	10.33	2.0	Radiance	Clean longwave window
14	11.21	2.0	Radiance	Longwave window
15	12.29	2.0	Radiance	Dirty longwave window
16	13.28	2.0	Radiance	CO2 (carbon dioxide)

Table A2. Evaluation of reconstruction quality across different codebook sizes. Highlighted values (yellow) indicate best performance for each metric

Codebook size	Residual blocks	LPIPS↓	PSNR↑	RMSE↓	SSIM↑
1024	3	0.587	10.749	0.293	0.234
2048	3	0.527	11.622	0.267	0.276
4096	3	0.539	11.905	0.259	0.311
8192	4	0.537	11.267	0.276	0.263

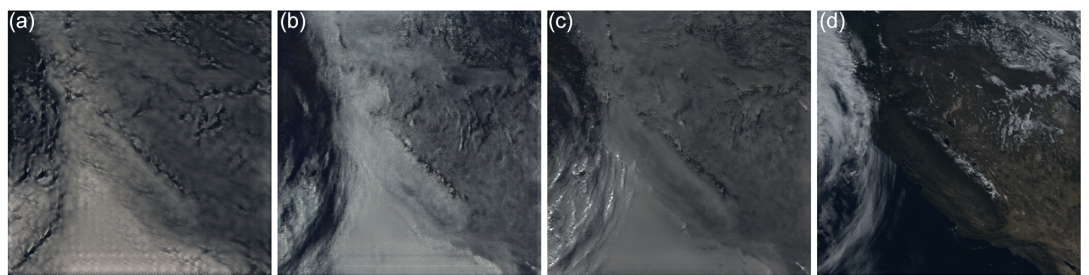


Figure A3. Comparison of the effect of reduced discriminator influence on image reconstruction quality. (a) Default model. (b) Model with 2 residual blocks before each downsample and upsample block. (c) Model with 4 residual blocks before each downsample and upsample block. (d) Ground truth image.

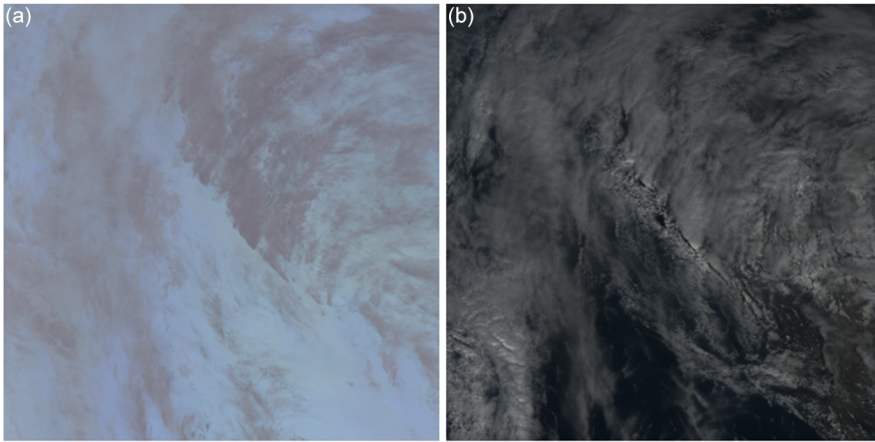


Figure A4. Nighttime visible images over land. (a) 3 Band nighttime IR image. (b) Generated nighttime visible image.

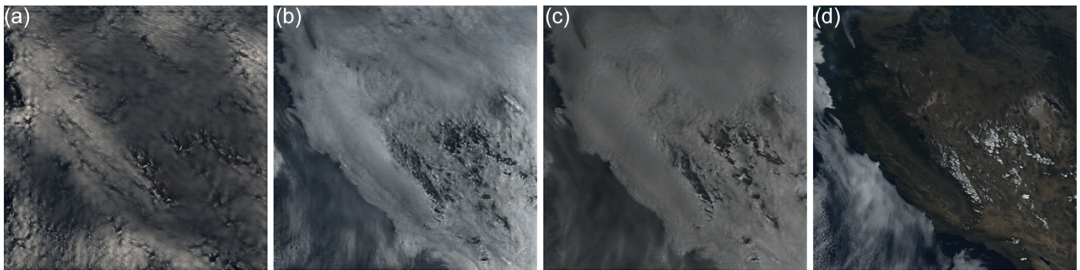


Figure A5. Comparison of the impact of embedding dimension on models trained with ABI bands 10, 11, and 14. (a) Baseline model. (b) Model with 4-dimensional embedding. (c) Model with 6-dimensional embedding. (d) Ground truth VIS.

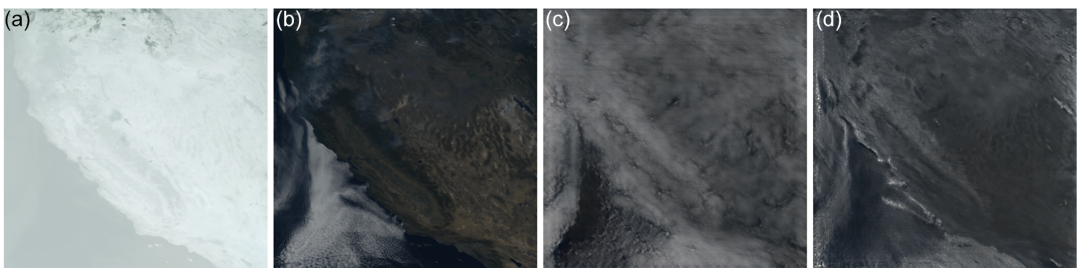


Figure A6. Visual comparison of model outputs when evaluated on inputs it was not trained on—specifically, the Band 11, 13, and 14 combination. (a) Input LWIR imagery. (b) Ground truth VIS image. (c) Baseline model output. (d) Output from the trained model using $Z^D = 4$.