Seeing Through the Clouds: Cloud Gap Imputation with Prithvi Foundation Model

Denys Godwin, Hanxi (Steve) Li, Michael Cecil & Hamed Alemohammad
Graduate School of Geography and Clark Center for Geospatial Analytics
Clark University
Worcester, MA 01610, USA
`dgodwin@clarku.edu`

Abstract

Filling cloudy pixels in multispectral satellite imagery is essential for accurate data analysis and downstream applications, especially for tasks which require time series data. To address this issue, we compare the performance of a foundational Vision Transformer (ViT) model with a baseline Conditional Generative Adversarial Network (CGAN) model for missing value imputation in time series of multispectral satellite imagery. We randomly mask time series of satellite images using real-world cloud masks and train each model to reconstruct the missing pixels. The ViT model is fine-tuned from a pretrained model, while the CGAN is trained from scratch. Using quantitative evaluation metrics such as structural similarity index and mean absolute error as well as qualitative visual analysis, we assess imputation accuracy and contextual preservation.

1 Introduction

The enormous scale of satellite image datasets along with the proliferation and sophistication of machine learning models has led to the development of Geospatial Foundation Models (GFM) which use self-supervised learning to encode patterns from terabytes of data for which high-quality ground truth labels do not exist (e.g. Lacoste et al. (2023); Jakubik et al. (2023); Klemmer et al. (2023); Tseng et al. (2023); Cha et al. (2023); Cong et al. (2022)). These patterns of encoded spatial, temporal, and band relationships can then be decoded with decoder heads that can be fine-tuned for individual use-cases. This not only has the potential to reduce the training time, cost, and resource use of machine learning inference, but also to improve model performance.

Self-supervision for satellite imagery may be achieved using masked autoencoders learners (He et al., 2022). Missing data imputation is therefore an inherent ability of FMs trained in this way. Given that Missing data due to clouds is a persistent issue in multispectral satellite imagery, self-supervised models that excel at cloud gap imputation and can transfer this ability to other geographic regions with little fine-tuning could assist in downstream tasks such as land use change detection or crop monitoring.

The utility of cloud gap imputation, whether done by artificial neural networks or by algorithmic interpolation methods, is dependent on its downstream application. Any reconstructed pixel value is a "best guess" based on learned representations of spatial and temporal context, and does not represent the true pixel value at a time and location. However, the augmentation of training datasets with synthetic data can improve the performance of computer vision models, as shown in the medical field by Mahmood et al. (2020). Additionally, the accuracy of a generative model in pixel reconstruction is a useful metric in assessing its ability to encode deep representations of the input data.

In adapting the concept of FMs to satellite imagery, it is important to benchmark against current state-of-the-art models and assess the unique modalities and challenges of the remote sensing domain (Rolf et al., 2024). Here, we measure the performance of Prithvi GFM against a Conditional Generative Adversarial Network (CGAN) for imputing cloud gaps in time series of Harmonized Landsat and Sentinel-2 (HLS) imagery using a range of sample sizes.
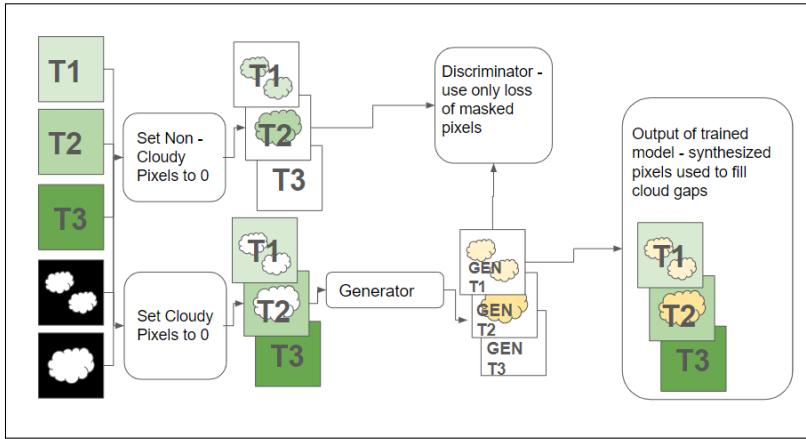
Figure 1: Training the CGAN to impute cloudy pixels was accomplished by masking out clouds from input data, using this as the condition on which to generate, then comparing the generated data against the unmasked ground truth.

2 MODEL ARCHITECTURES

Prithvi is a GFM based on the multi-temporal vision transformer (ViT) architecture (with 100M parameters) and has been trained on 1TB of multispectral satellite imagery from HLS dataset (Jakubik et al., 2023). The model is pretrained using a masked autoencoder learner on a dataset where each scene consists of 3 time steps of 6 bands each.

The CGAN architecture contains two encoder-decoder convolutional neural networks which are trained against each other to generate realistic outputs based on input data. We use randomly masked time series scenes as the input condition (Figure 1). The CGAN model was adapted from Baier et al. (2022) which is based on the landmark CGAN architecture Pix2Pix (Isola et al., 2017). We use a multi-scale patch-based discriminator to evaluate plausibility at two physical scales, for both small detail and wider spatial context. The use of the CGAN for inpainting gaps in input images is established using non-geospatial data (Demir & Unal, 2018). This technique in the geospatial context benefits from our time-series data, as it allows values from each time step to inform guesses about missing values, thanks to the persistence, between different time scenes in a given year, of features on the ground.

3 METHODOLOGY

In order to compare the performance of Prithvi to the baseline CGAN model, both models were trained using subsets of a cloud-free time-series dataset with randomly added real life cloud masks and tested on a reserved validation set which remained constant and identical across all experiments with no shuffling or random selection.

3.1 DATASET AND EXPERIMENT SETUP

Imagery from HLS dataset was collected across the Contiguous United States (CONUS) with diverse land cover classes (details in Jakubik et al. (2023)). Each chip covers a 224 x 224 pixel region, with a spatial resolution of 30 meters and six spectral bands from three temporal snapshots stacked together for 18 total channels. The three scenes are selected between Mar and Sep 2022 with time difference between scenes varying between 1 and 200 days. After filtering for missing values and cloudy pixels, a total of 7,852 cloud-free chips evenly distributed across the CONUS were generated. This set was randomly partitioned into 6,231 training chips and 1,621 validation chips. The same training and validation sets were used for all experiments. The training dataset was then reduced to subsets of 3,200, 1,600, 800, and 400 for successive experiments.

The dataset for these experiments was collected independently of the training dataset used to train Prithvi to prevent leakage. Prithvi was trained using 2017 HLS data, while these experiments were run using 2022 HLS data. While both datasets were randomly sampled from CONUS, sampling from different years ensures that image chips used in pretraining Prithvi are not part of these fine-tuning experiments.

Cloud masks were generated from the same region of CONUS using HLS cloud mask quality flag and exported as a binary layer of cloudy and non-cloudy pixels. This yielded 21,642 cloud masks, of which 1,600 were randomly selected and reserved for validation. The validation set was balanced such that there were an equal number of samples in each of 10 equal divisions between 1 and 100 percent coverage.

We setup two experiments: in experiment 1 (E1) only the middle scene in each chip was masked, and in experiment 2 (E2) all combination of all time steps were used for masking (e.g. all three scenes, or just the first two scenes, etc). All experiments were run for 200 epochs.

3.2 PRITHVI FINE-TUNING

The pretraining loop for Prithvi was modified to accept real-life cloud masks as inputs, replacing the random masking that had been used in training the model. As the ViT architecture only reconstructs missing patches, and incorporates no data from patches which are masked out, any patch which included a cloudy pixel was therefore removed entirely. The Mean Squared Error (MSE) of reconstructed patches as compared to their equivalent ground truth patches is used as the loss metric for training. In calculating metrics for comparison with the CGAN, only missing pixels are replaced with generated values, rather than replacing the entire patch that the missing pixels belong to.

3.3 CGAN TRAINING

For CGAN training, first the generator weights are frozen and the discriminator is trained using hinge loss. Then, the discriminator weights are frozen and the generator weights are updated based on hinge loss compared to the discriminator outputs and to MSE calculated between true and generated values with hyperparameter α as the weight for MSE:

$$\text{Total Loss} = \text{hinge loss} + \alpha \cdot \text{MSE} \tag{1}$$

MSE is only calculated for pixels that are masked out, and discriminator loss is only calculated for patches which contain at least one cloudy pixel. Figure 1 shows how the CGAN was trained using cloud-masked time series as a condition and the ground truth as the target. For all experiments, discriminator learning rate was $1.0e - 4$ and generator learning rate was $5.0e - 4$, following Heusel et al. (2017) and preliminary experiments in learning rate selection. α is set to 5 for all experiments.

4 RESULTS AND DISCUSSION

We use Structural Similarity Index Measure (SSIM) as a quantitative metric to compare outputs of Prithvi and CGAN against the ground truth, but SSIM is not used as a loss metric during training. SSIM was chosen to compliment pixel-wise Mean Absolute Error (MAE) as it is sensitive to changes in texture, spatial patterns, and contrast, which are important in downstream tasks such as semantic segmentation. Reconstructed time series, where generated values were imputed for cloud masks but all unmasked data was preserved, were compared against the original complete time series using SSIM. Unlike pixel-wise error, SSIM was not adjusted to account for cloud coverage, so values for this metric are highly dependent on cloud cover.

In comparing results from E1 and E2 experiments, it is important to note that SSIM was calculated for only the center time step in E1, and across all time steps for E2. Therefore, SSIM between E1 and E2 experiments cannot be rigorously compared, and should serve only in comparing the results of Prithvi and CGAN model from the same round of experiments.

For E1 experiments, Table 2 and Figure A.3 show that Prithvi outperforms CGAN in nearly all cases in terms of both MAE and SSIM. Only CGAN trained on the full training set marginally outperforms

Prithvi with no fine-tuning. Table A.5 shows that Prithvi is also less affected by increases in time gap and cloudy percentage than the CGAN model, shown in Table A.6.

For E2 experiments, Prithvi outperforms the CGAN model as is shown in Table 1 and Figure 1. Zero-shot inference by Prithvi achieves an MAE of 0.03 on all masked pixels in the validation dataset, which is reduced to below 0.025 in 10 epochs or under for all training data subsets. As mean pixel reflectance values for all bands of the validation dataset were 0.151, MAE of 0.03 would denote an average error of approximately 20 percent.

Overall, the CGAN model consecutively shows improvement in both MAE and SSIM with larger subsets of data, suggesting that 6,231 scenes does not achieve saturation for the model. However, even with the full dataset, it is outperformed by Prithvi with no fine-tuning. Prithvi outperforms the CGAN despite being unable to exploit all available information due to discarding any input patch with any masked pixels during training and inference.

Table 1: MAE and SSIM for best epoch of 200 for E2 experiments (applying masks in all time steps). Best epoch is the best performance out of 5 runs for all experiments across all epochs.

| Sample Size | SSIM | | | | MAE | | | |
| | Training | | Validation | | Training | | Validation | |
	Prithvi	CGAN	Prithvi	CGAN	Prithvi	CGAN	Prithvi	CGAN
6231	**0.949**	0.919	**0.960**	0.937	**0.024**	0.033	**0.022**	0.032
3200	**0.949**	0.917	**0.959**	0.931	**0.024**	0.035	**0.023**	0.035
1600	**0.950**	0.916	**0.958**	0.926	**0.024**	0.035	**0.023**	0.037
800	**0.950**	0.905	**0.957**	0.916	**0.024**	0.038	**0.024**	0.042
400	**0.947**	0.894	**0.956**	0.912	**0.025**	0.045	**0.024**	0.045
0 (zero-shot)	**-**	-	**0.946**	-	**-**	-	**0.030**	-

In fine-tuning for E1, Prithvi experiences loss spikes before finding a new, higher performance local minimum, as is seen in Figure A.8. Similarly, for E2, Prithvi's best performance is achieved before 10 epochs on all fine-tuning experiments with a collapse in performance after that and a convergence to a new, suboptimal local minima (Figure A.10). This differs from fine-tuning E1 experiments because masking at multiple time steps is closer to the original pretraining method of Prithvi.

Visual inspection of true color reconstructed scenes show that both Prithvi and the CGAN are capable of filling cloudy pixels in challenging combinations of large time gaps and large chunks of missing data. The CGAN tends to preserve small details but outputs unrealistic reflectance values and salt and pepper noise, especially where there is insufficient data. For example, in Figure 2, the CGAN outputs anomalously high reflectance values. Prithvi, on the other hand, does not tend to preserve details such as roads and buildings, but outputs reflectance values that are more constrained within the bounds of true observations. When there is insufficient data, the images are subject to a checkerboarding effect, as in Figure A.14.

5 CONCLUSION

We find that Prithvi, even without fine-tuning, outperforms the CGAN on cloud gap imputation with all subsets of data. This shows the strength of GFMs on gap-filling tasks, given that this mimics the training loop of models such as Prithvi. The model's exposure to over 1TB of diverse images helps constrain its predicted values compared to the bespoke CGAN (Figures A.16-A.19). The gap-filled imagery can be used to augment time series data for training other downstream applications which benefit from complete coverage and multi-temporal scenes (like crop type segmentation or crop yield estimation). Further research is needed to assess the extent to which cloud gap imputation as a data augmentation technique is effective in training better downstream models.

Given that the models perform better in E1 experiments (masked middle scene), we recommend this configuration for cloud gap imputation. Prithvi has no time encoding for scenes; therefore, non-cloudy scenes may be selected from across the range of the year.
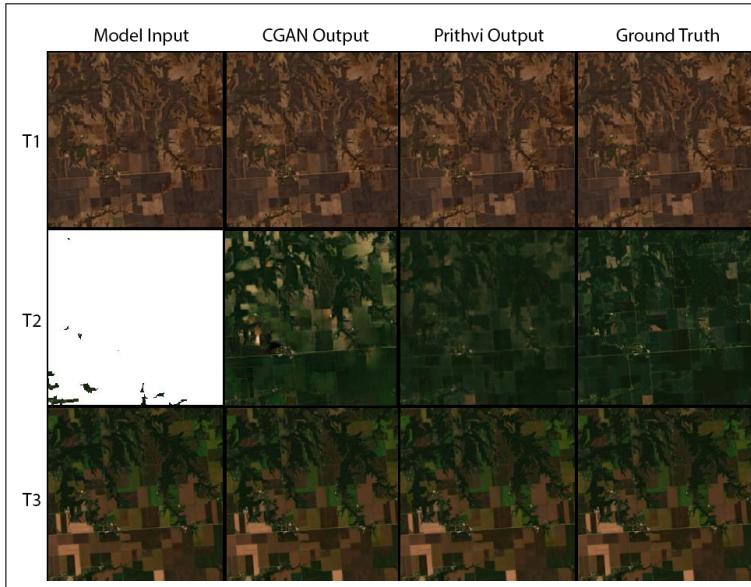
Figure 2: Reconstruction of a high-coverage image using CGAN and Prithvi, both trained using 6,231 images from E1 experiments (applying mask to the middle scene)

One strength of time series satellite imagery for cloud gap imputation is the relative stationarity of features on the ground over time. This strength contributes to the ability of FMs to generate realistic pixel values in masked areas. Future research can incorporate non-imagery datasets such as Digital Elevation Models (DEMs) and land cover classification layers that can improve model performance.

REFERENCES

Gerald Baier, Antonin Deschemps, Michael Schmitt, and Naoto Yokoya. Synthesizing optical and sar imagery from land cover maps and auxiliary raster data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. ISSN 1558-0644. doi: 10.1109/tgrs.2021.3068532. URL http://dx.doi.org/10.1109/TGRS.2021.3068532.

Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=WBhqzpF6KYH.

Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

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. *Advances in neural information processing systems*, 30, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. GEO-bench: Toward foundation models for earth monitoring. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=IptxZvA3at.

Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. Mckay, Kevan J. Salimian, Alexander Baras, and Nicholas J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3257–3267, 2020. doi: 10.1109/TMI.2019.2927182.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024.

Gabriel Tseng, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

A APPENDIX

Table 2: MAE and SSIM for best epoch of 200 for E1 experiments (applying mask in the middle scene).

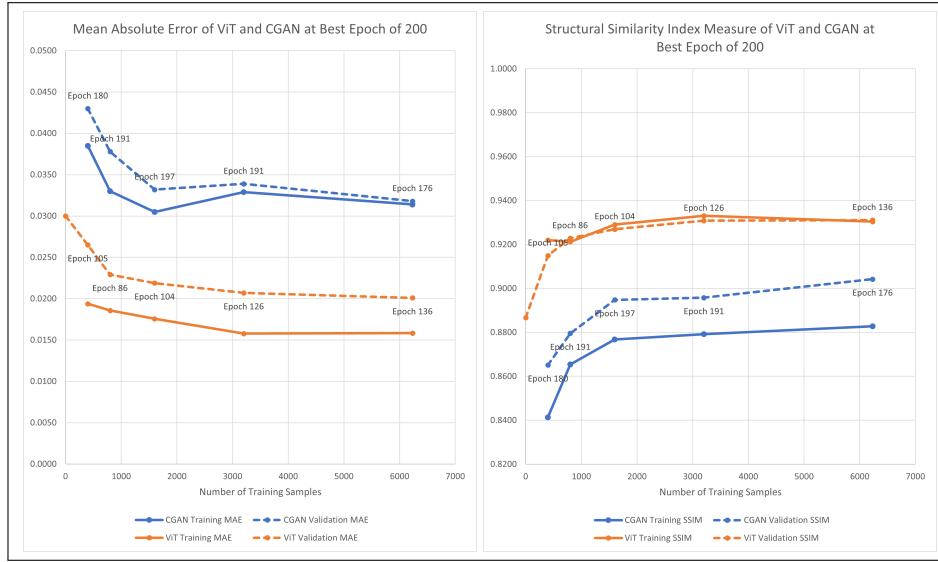| Sample Size | SSIM | | | | MAE | | | |
| | Training | | Validation | | Training | | Validation | |
	Prithvi	CGAN	Prithvi	CGAN	Prithvi	CGAN	Prithvi	CGAN
6231	**0.930**	0.883	**0.931**	0.904	**0.016**	0.031	**0.020**	0.032
3200	**0.933**	0.872	**0.931**	0.896	**0.016**	0.033	**0.021**	0.035
1600	**0.929**	0.877	**0.927**	0.895	**0.017**	0.031	**0.022**	0.037
800	**0.921**	0.866	**0.923**	0.876	**0.019**	0.033	**0.023**	0.042
400	**0.922**	0.841	**0.915**	0.865	**0.019**	0.039	**0.027**	0.045
0 (zero-shot)	**-**	-	**0.887**	-	**-**	-	**0.030**	-

Figure A.3: MAE and SSIM for best epoch of 200 for E1 experiments (applying mask to the middle scene). Best epoch is the best performance out of 5 runs for all experiments across all epochs.
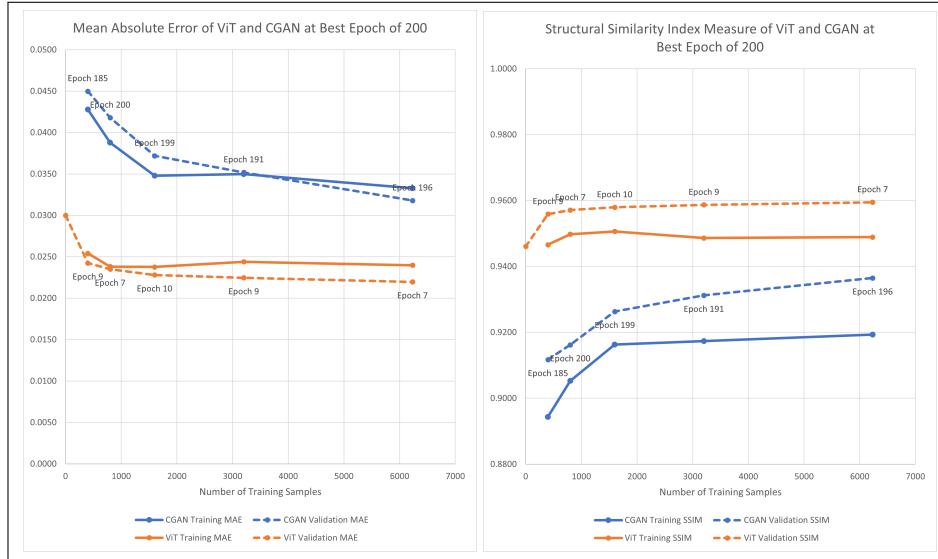

Figure A.4: MAE and SSIM for best epoch of 200 for E2 experiments (applying masks in all time steps). Best epoch is the best performance out of 5 runs for all experiments across all epochs.
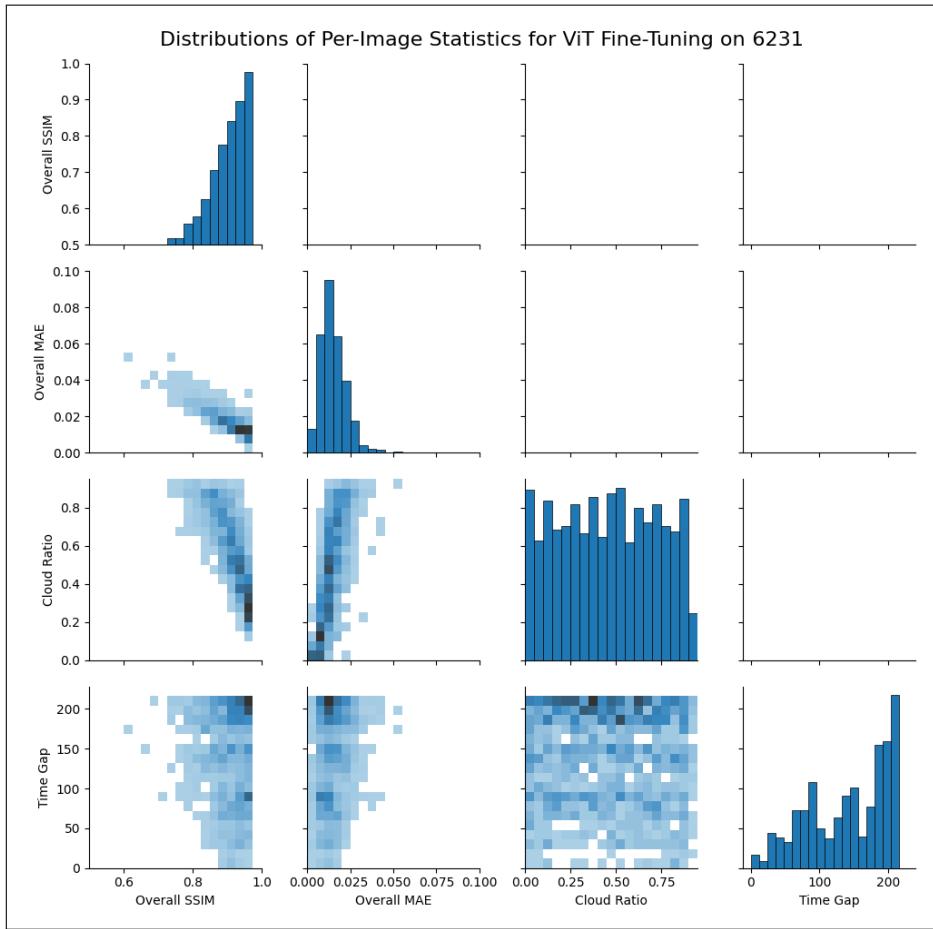
Figure A.5: Relationships between cloud cover, time gap, SSIM, and MAE of validation chips for best results of E1 experiments using the full dataset to fine-tune Prithvi. Statistics are calculated for each validation chip.
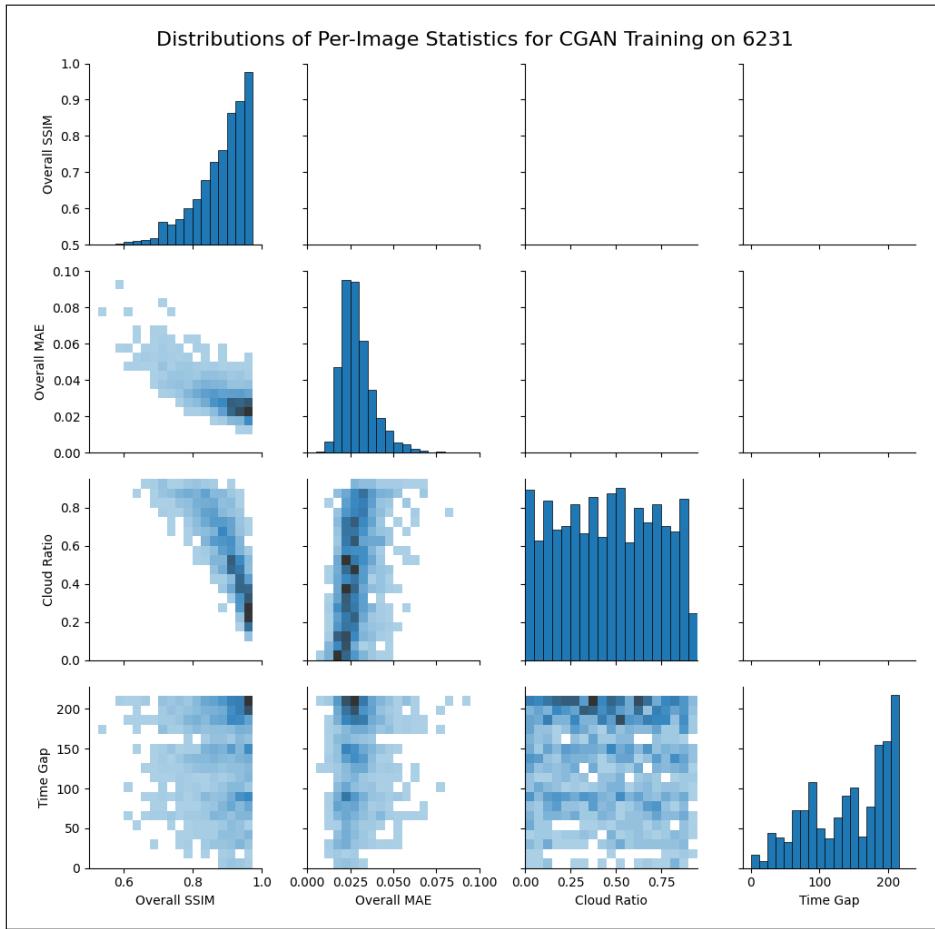
Figure A.6: Relationships between cloud cover, time gap, SSIM, and MAE of validation chips for best results of E1 experiments using the full dataset to train the CGAN. Statistics are calculated for each validation chip.
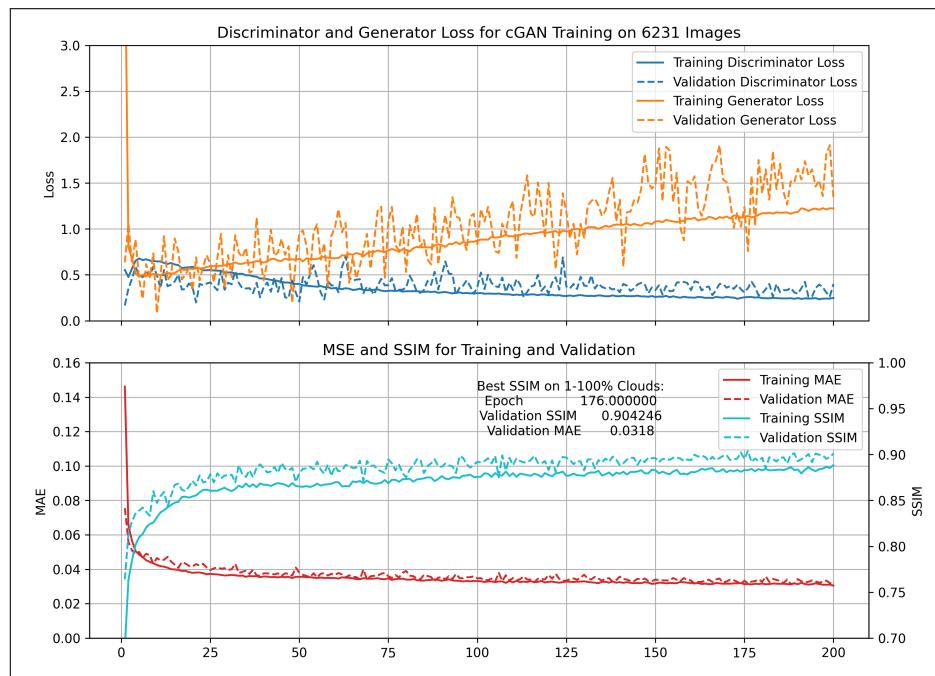
Figure A.7: Performance of the CGAN over the course of training on 6231 images in E1 experiments (applying mask to the middle scene)
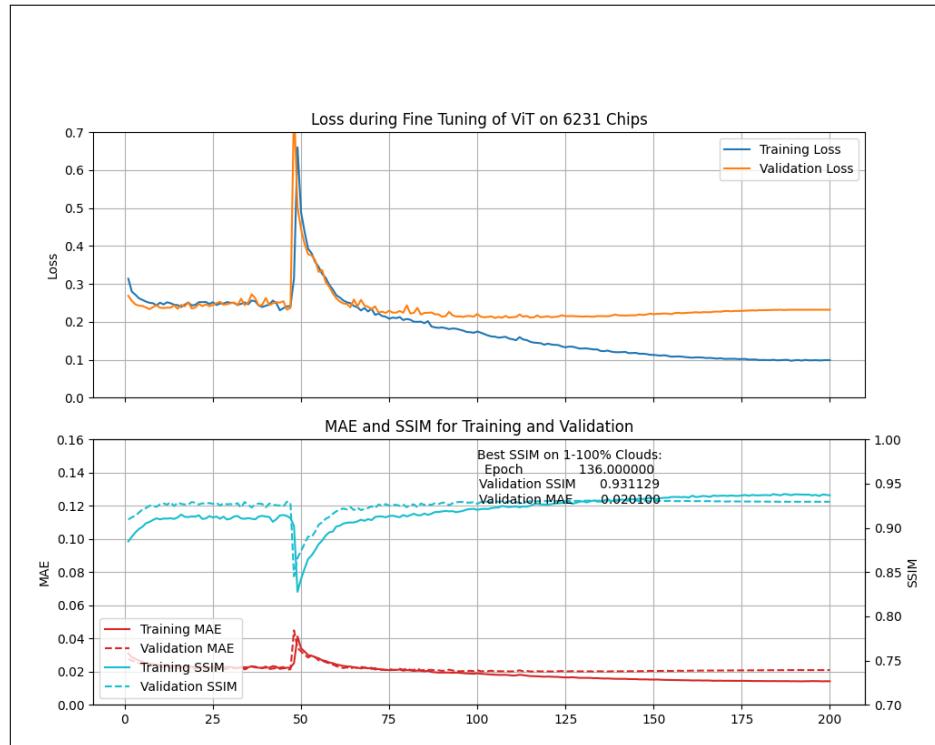
Figure A.8: Performance of Prithvi over the course of training on 6231 images in E1 experiments (applying mask to the middle scene)
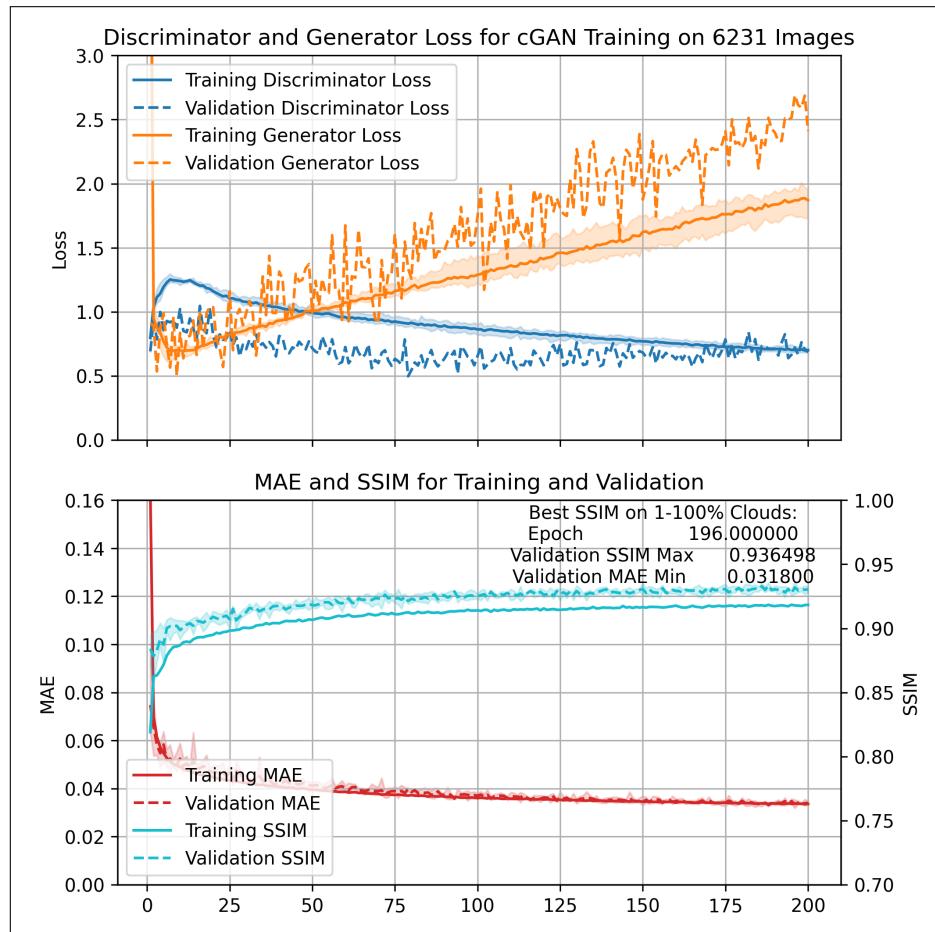
Figure A.9: Performance of the CGAN over the course of training on 6231 images in E2 experiments (applying masks in all time steps)
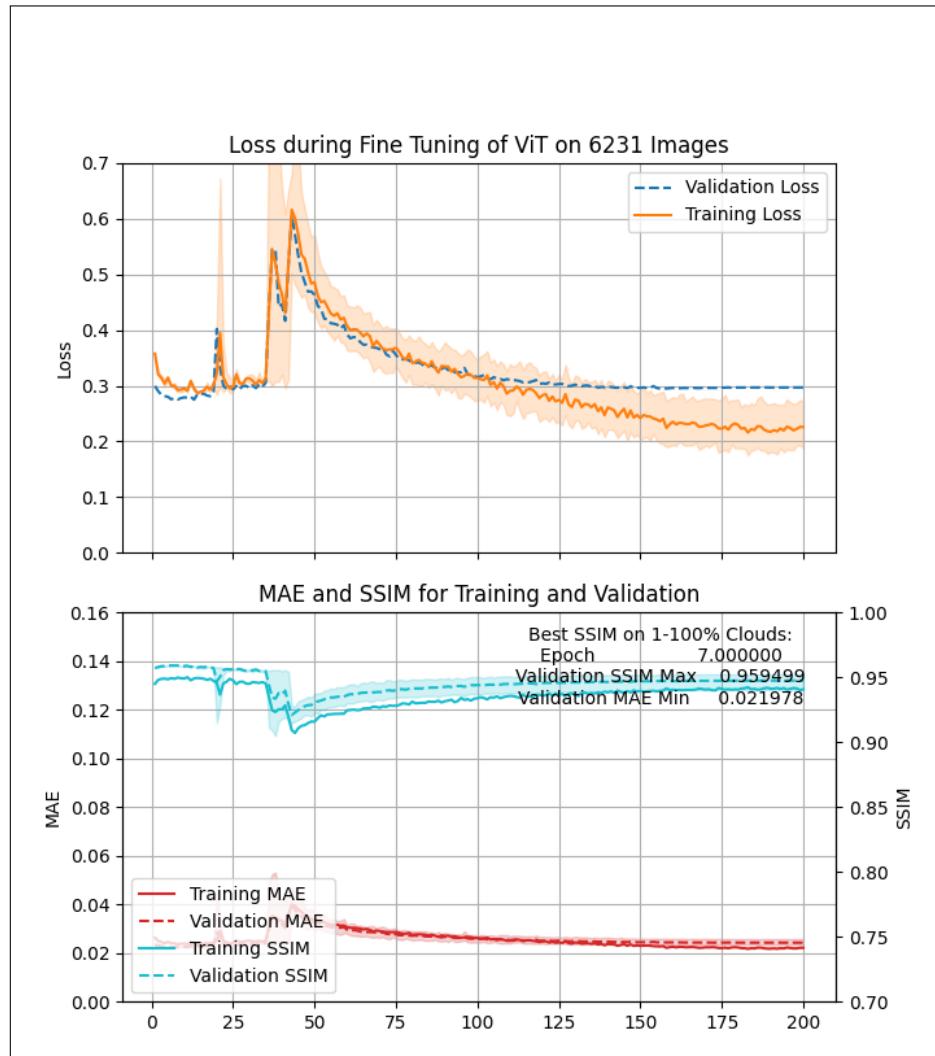
Figure A.10: Performance of Prithvi over the course of training on 6231 images in E2 experiments (applying masks in all time steps)
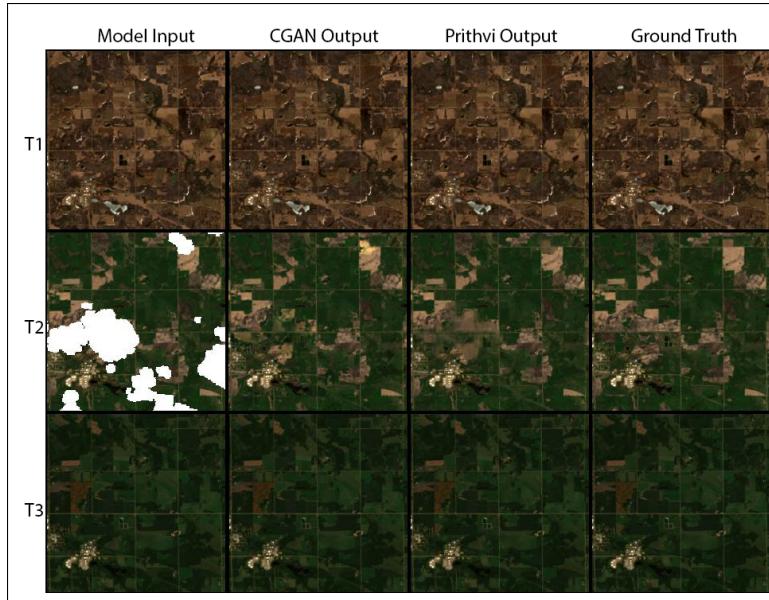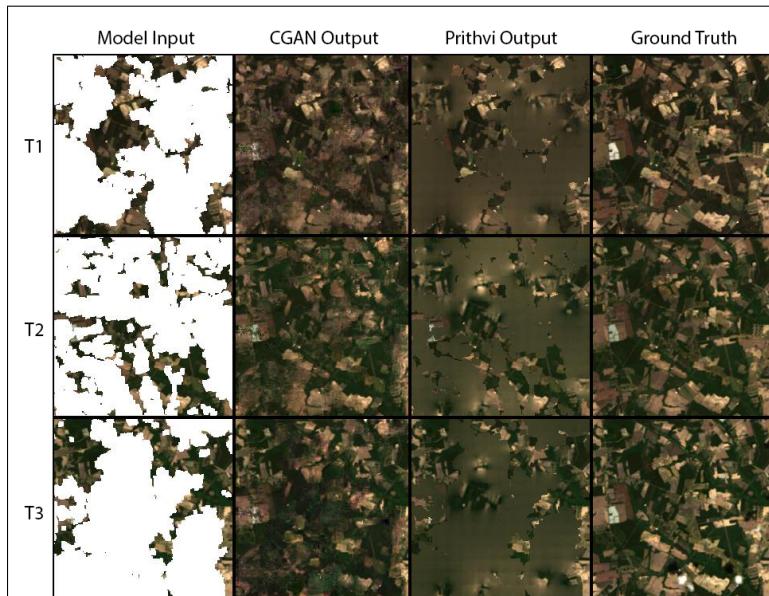
Figure A.11: Reconstruction of a low-coverage image using CGAN and Prithvi, both trained using 6,231 images from E1 experiments (applying mask to the middle scene)
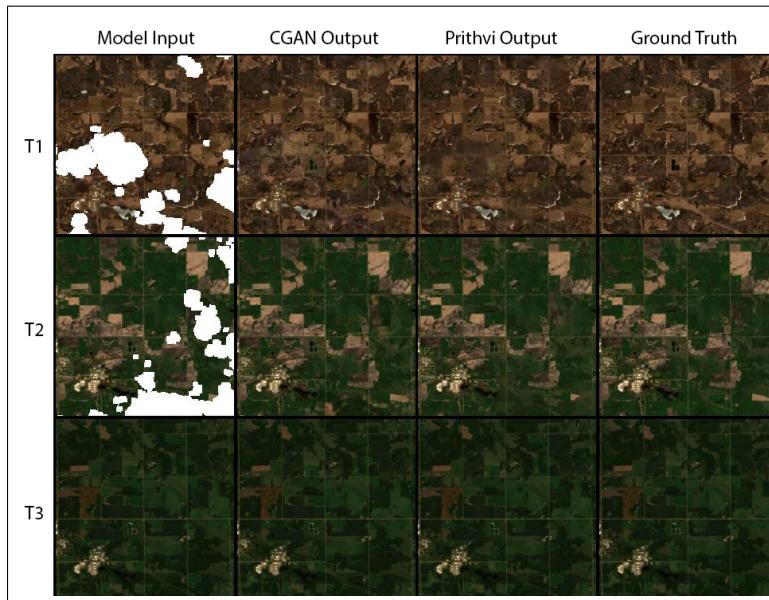
Figure A.12: Reconstruction of a low-coverage image using CGAN and Prithvi, both trained using 6,231 images, from E2 experiments (applying masks in all time steps)
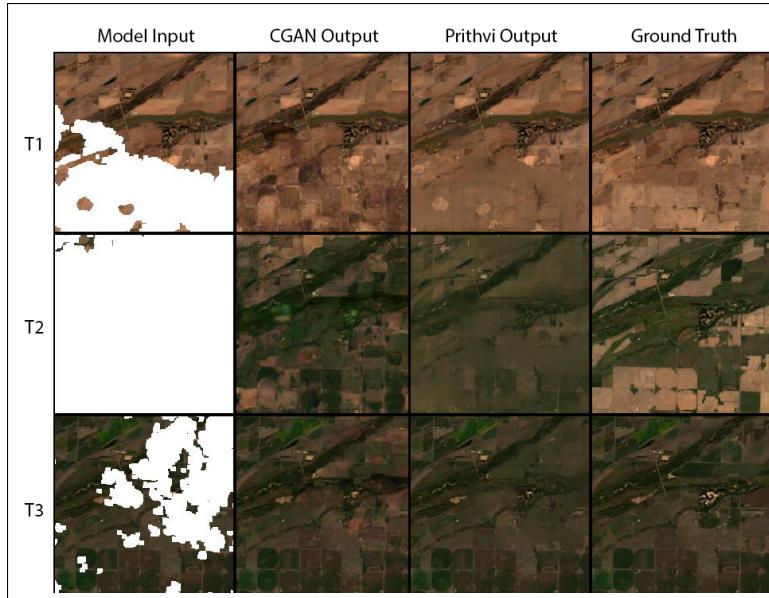
Figure A.13: Reconstruction of high-coverage image using CGAN and Prithvi, both trained using 6,231 images, from E2 experiments (applying masks in all time steps)

Figure A.14: Reconstruction of a nearly full-coverage image using CGAN and Prithvi, both trained using 6,231 images from E2 experiments (applying masks in all time steps)
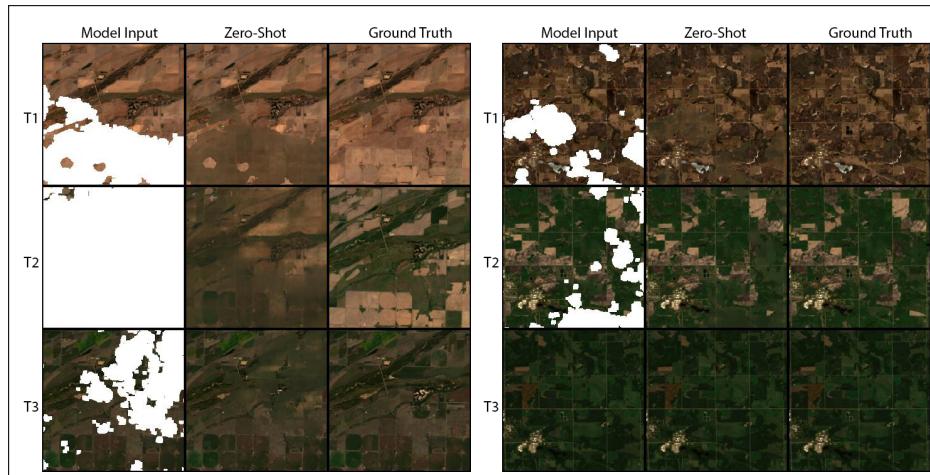
Figure A.15: Reconstruction of a high and low coverage image using Prithvi with no fine-tuning from E2 experiments (applying masks in all time steps)
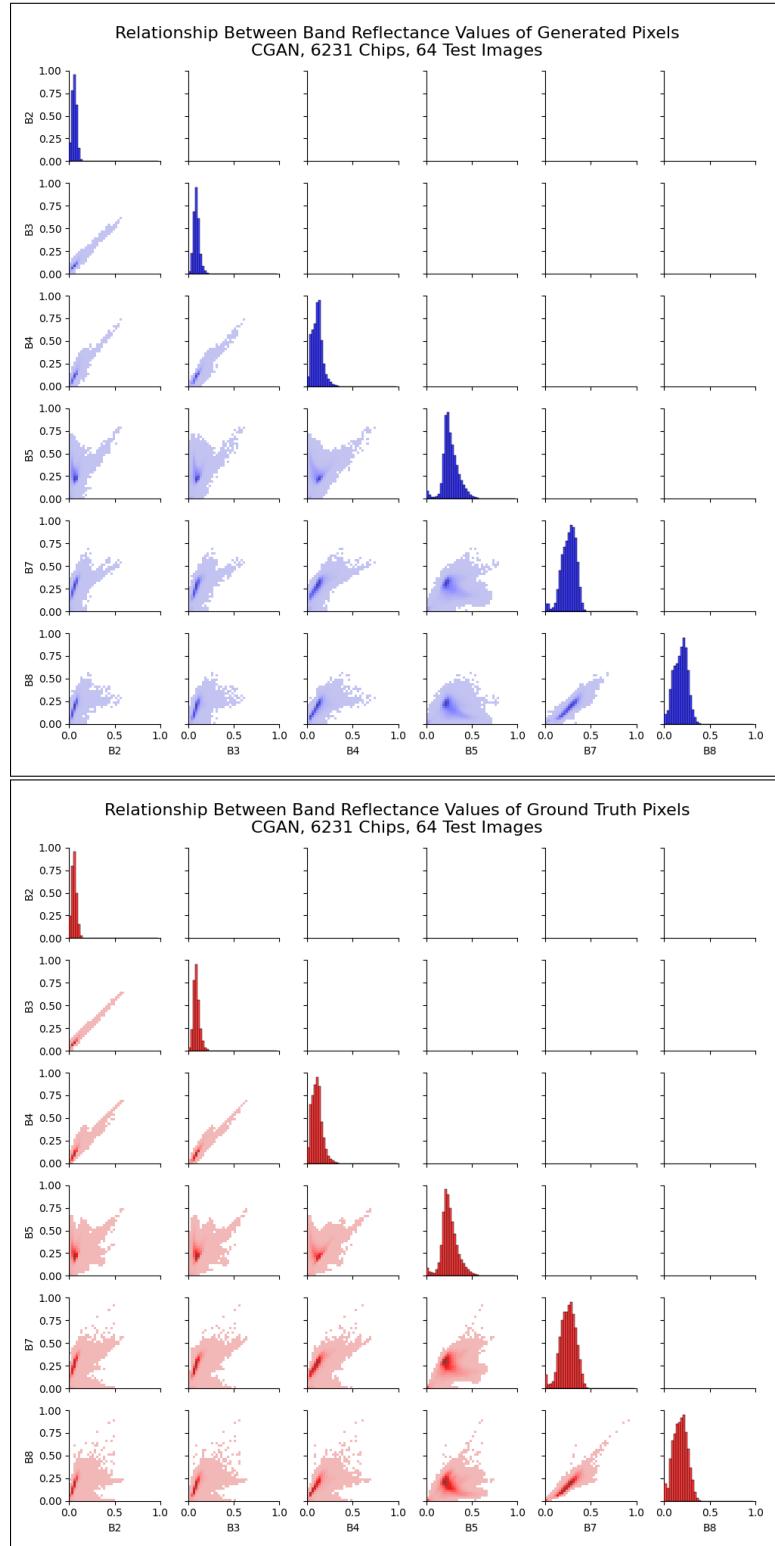
Figure A.16: Band correlation matrices for the best CGAN model trained on the full dataset for E2 experiments (applying masks in all time steps). Band indices are shown for the first 64 image/mask combinations from the validation dataset, with generated values (top) and ground truth values (bottom).
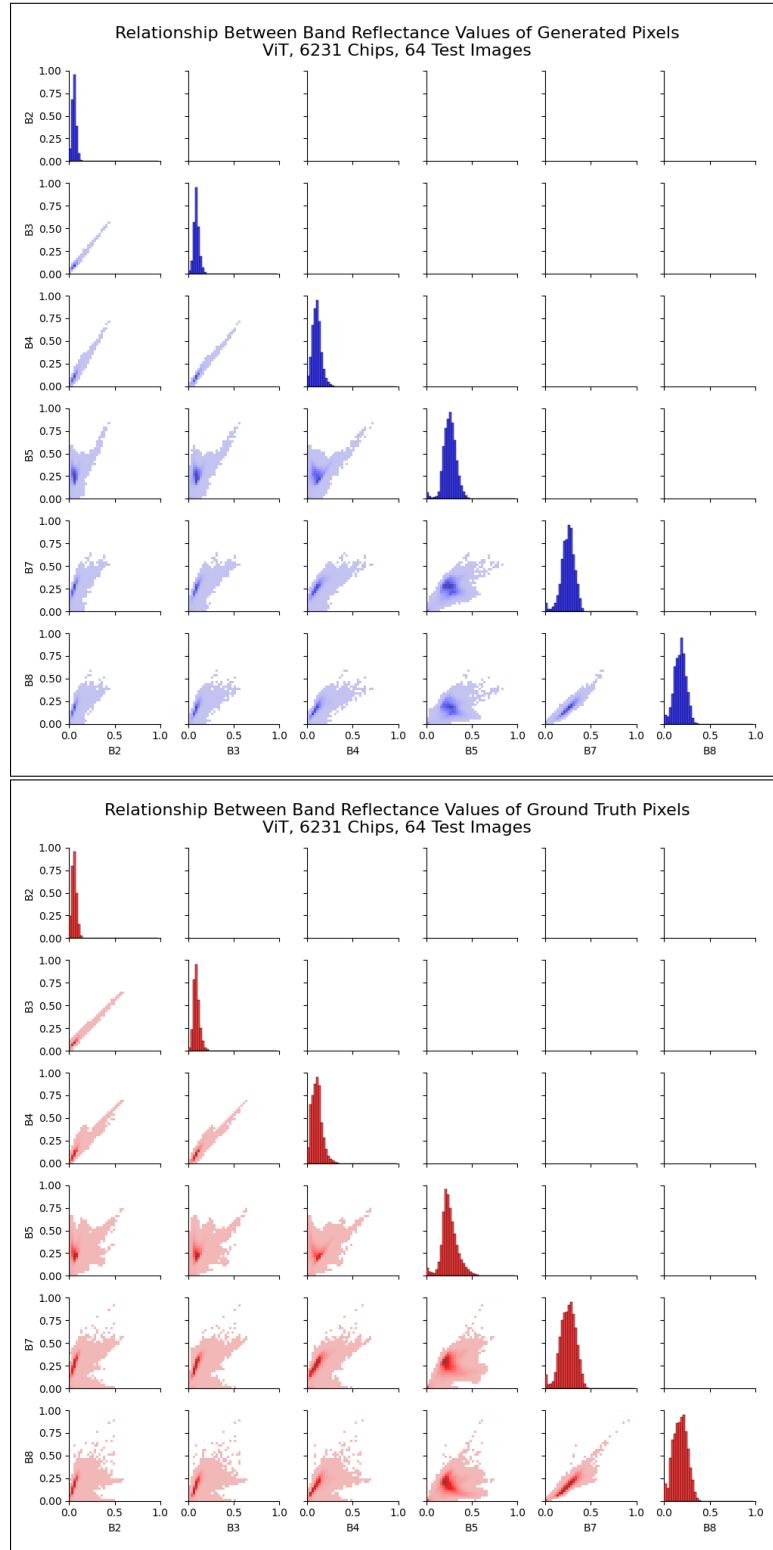
Figure A.17: Band correlation matrices for the best fine-tuned Prithvi model trained on the full dataset for E2 experiments (applying masks in all time steps). Band indices are shown for the first 64 image/mask combinations from the validation dataset, with generated values (top) and ground truth values (bottom).
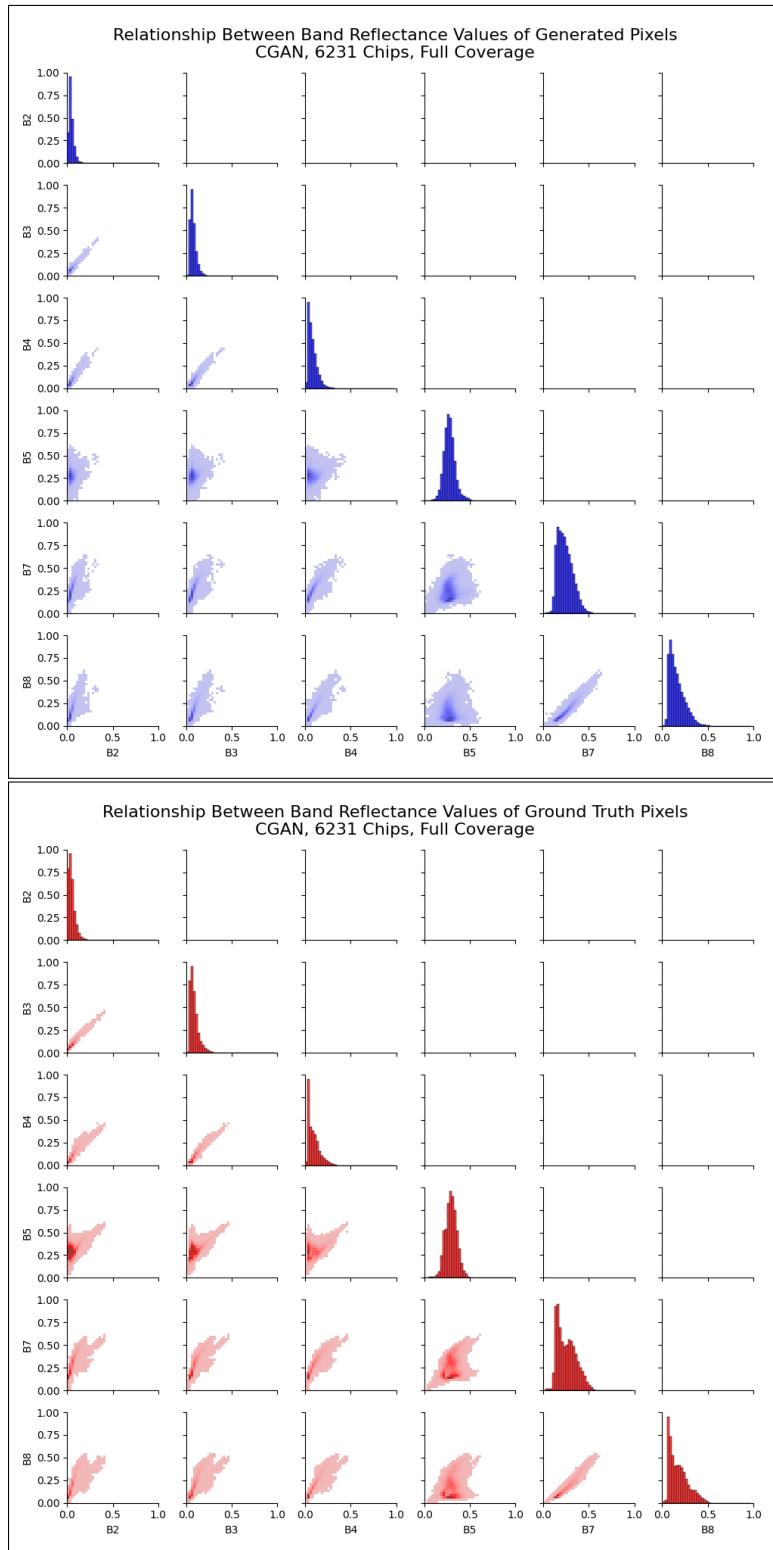
Figure A.18: Band correlation matrices for the best fine-tuned CGAN model trained on the full dataset for E2 experiments (applying masks in all time steps). Band indices are shown for the sample image in A.14, with generated values (top) and ground truth values (bottom).
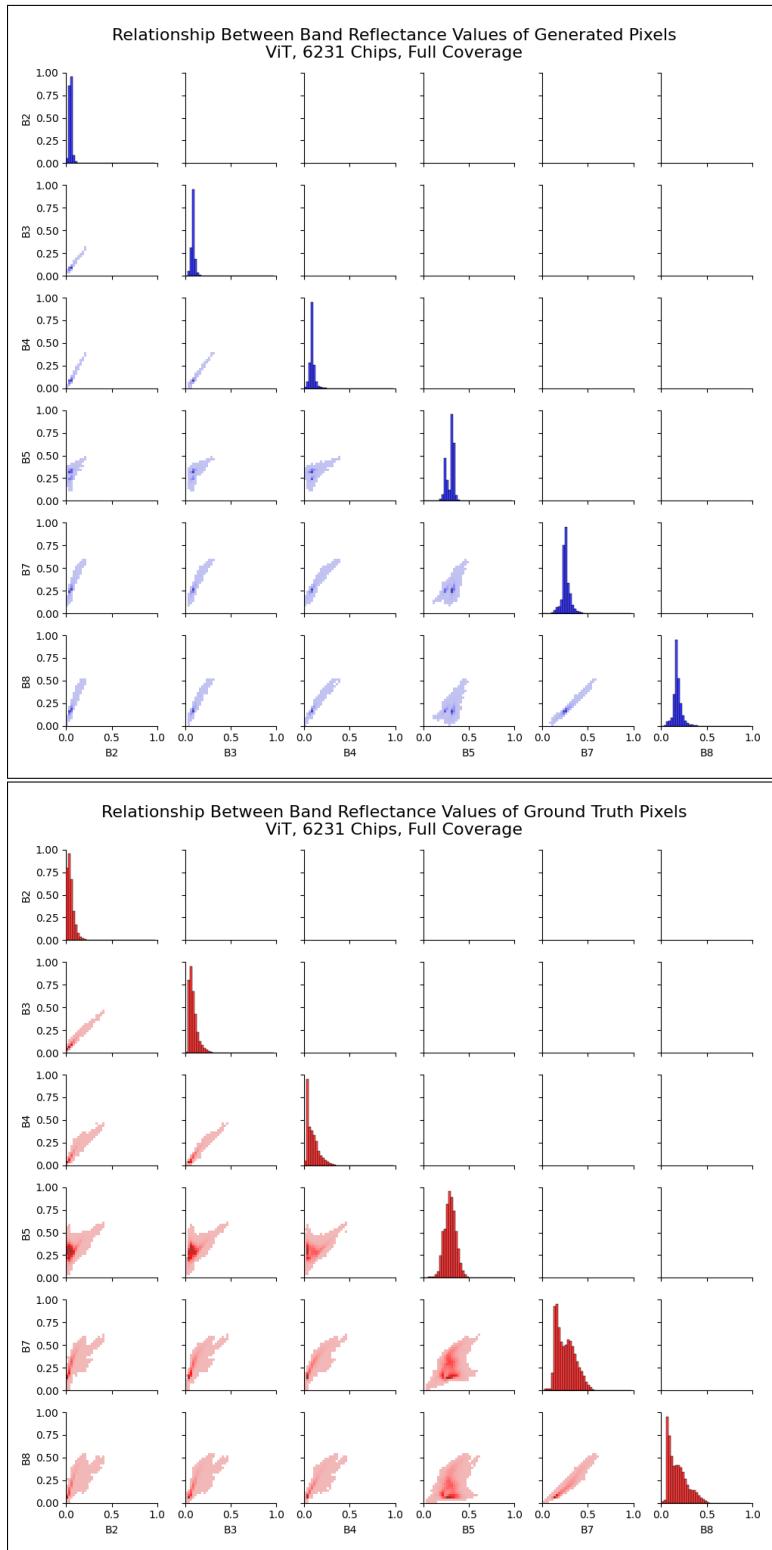
Figure A.19: Band correlation matrices for the best fine-tuned Prithvi model trained on the full dataset for E2 experiments (applying masks in all time steps). Band indices are shown for the sample image in A.14, with generated values (top) and ground truth values (bottom).