A self-supervised multi-source framework and architecture for generative cross-sensor harmonization.

Clément Dauvilliers
INRIA Paris
Paris, Île-de-France, France
clement.dauvilliers@inria.fr

Claire Monteleoni
INRIA Paris
University of Colorado Boulder

Abstract

In this work, we train a generative model to reconstruct satellite images from observations from other satellites within the same constellation. We train on observations that only partially geographically overlap, and occur at different times. Since a single image does not contain the full information required to reconstruct another, we use a multi-source architecture that combines images from multiple satellites as input to for the reconstruction. We use the reconstruction task to train the model in a self-supervised manner, where one image is randomly noised and needs to be reconstructed from the other ones. We show that in this setting, training with images occurring up to 6h apart leads to a better CRPS than training only on maximum time differences of 1h, despite the former being a harder task. Finally, we show that the model benefits from using multiple sources as inputs.

1 Introduction

Remote Sensing (RS) data and in particular satellite imagery plays a crucial role in many modern applications, such as land use estimation, climate monitoring or weather forecasting, among others. A large fraction of those come from constellations, i.e. ensembles of satellites coupled to produce a common product, which allow more frequent and larger scale observations than lone satellites. Nevertheless, the sensors on different satellites of a constellation do not necessarily have the exact same characteristics. For example, the passive microwave (PMW) sensors within the GPM constellation [2] differ in ground sampling distance, measured frequencies and brightness sensitivity, among others. Besides, each satellite possesses its own orbit and thus its own geometry over a specific area at any given time. For these reasons, applications using observations need to be adapted for each instrument specifically. This notably includes deep learning models trained on data from a specific sensor, whose performance drops when applied to a different instrument from the same constellation [8, 7].

However, training machine learning models on this so-called problem of *cross-sensor harmonization* faces several challenges. Some are due to the nature of satellite imagery [6]: dealing with multi-modal spatio-temporal data, features at multiple scales, and managing irregular geometries, among others. A challenge specific to cross-sensor harmonization is the lack of paired images. A usual setting is training on co-located observations, where "co-located" designates both space and time alignment: given K sensors $S_1, ..., S_K$, one would train a model on pairs of images from two sensors $(S_i, S_j)_{i,j \in \{1,...K\}}$ where the two images occur at the same time ("co-timed") and are geographically aligned. For some pairs of sensors, this configuration may never occur, or may occur only over specific areas and times, biasing the dataset [7].

In this preliminary work, we propose to train and evaluate a model on "close but non-co-located" observations: we allow training on images that cover only partially overlapping areas and do not exactly occur at the same time, but within a pre-defined time window. While this increases the

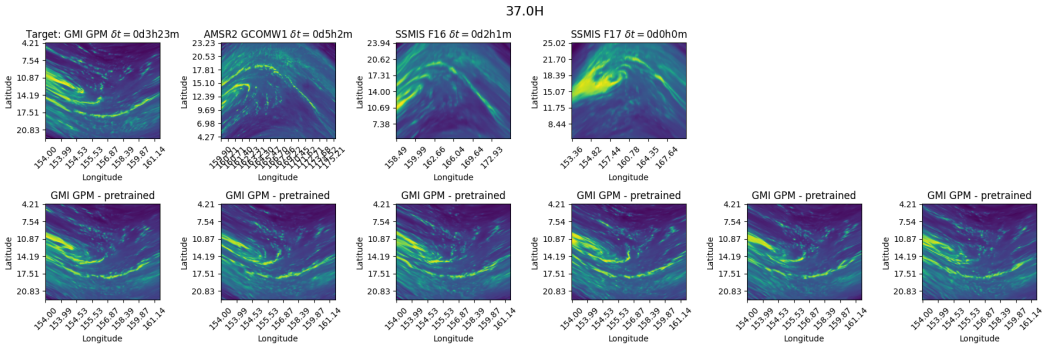

Figure 1: Visualization of an example of reconstruction, in the 37.0GHz frequency and horizontal polarization. Top row: (left) target image from the GPM satellite, which the model was tasked to reconstruct; (second, third and fourth images) observations from three different satellites given as input to the model. Each image has its own geometry and area. The δ_t value above each image indicates its time delta with the target image. Bottom row: six realizations sampled with the generative model trained with a time window of 6 hours.

amount of samples, the problem becomes ill-posed: if the inputs and target observations occur over slightly different areas and at different times, the inputs do not necessarily contain the full information required to predict the target image. To counter this, we employ two methodological aspects: first, we train a generative model, which allows to sample from the probability distribution of the target. Training a deterministic model that uses for example the root mean squared error (RMSE) as loss function would lead to predicting the mean of the target distribution; however in this case the task contains intrinsic uncertainty, and thus the mean would be a blurry, out-of-distribution image [1]. Second, our model can use multiple sources (i.e. images from different sensors) as input to predict its target. The hypothesis supporting this choice is that if each input observation only gives partial information regarding the target, the model may benefit from using their combination. In the following, we list the points that together distinguish this work within the field of cross-sensor harmonization:

Spatio-temporal misalignment. The satellite observations fed to the model and used as targets cover different geographical areas, and occur at irregular time intervals within a time window.

Training on non-co-timed data. We show that training on observations that do not occur at the same time leads to an improvement on co-timed ones, thanks to the larger amount of available data.

Generative framework. We train a generative model based on Flow Matching [4]. This choice is natural given the uncertainty inherent to the problem, and as downstream applications of the GPM sensors require in-distribution samples (e.g. for precipitation estimation [8]).

Multi-source inputs and outputs. Our model is trained on and can be applied to a flexible number of observations, corresponding to different sensors, each with its own characteristics.

Satellite geometry. Our model is applied to satellite observations directly in their original swath geometries (see Figure 1), without being interpolated to a regular grid. This lets the model learn any transformation between the sensors, without any loss of information due to re-gridding.

Self-supervised training. We train our model in a self-supervised setting, in which the pre-training task is also the downstream application. Given a set of images from different sensors, we randomly noise one of them and train the network to reconstruct it from the other sources.

2 Methods

2.1 Dataset

We use the TC-PRIMED dataset [5], which contains passive microwave overpasses over global tropical and extratropical storms between 1987 and 2023. We use observations from the following sensor-satellite pairs: AMSR2-GCOMW1, GMI-GPM, TMI-TRMM, SSMI-F11,F13,F14,F15,

SSMIS-F16,F17,F18,F19. We use the following frequencies and polarizations: near 37GHz horizontal and vertical, near 89GHz horizontal and vertical. For sensors that do not include exactly those frequencies, we use the closest (36.5, 36.64, 85.5 and 91.665GHz). The dataset is randomly split into a train, a validation and a test subset by storms, so that images from a common storm cannot be placed in the same subset.

2.2 Definition of a sample

Given a dataset of observations from a set of K sensors $S^1, ..., S^K$, a sample is defined as all observations of a common storm that fall within a time window $[t_0 - \delta_{max}, t_0]$, where t_0 is the reference time of the sample and δ_{max} is the size of the time window. δ_{max} is a hyperparameter, set to either 6 hours or 1 hour depending on the experiment.

2.3 Self-supervised flow matching training

Our training task is the following: at each iteration and for each sample in the mini-batch, one image x_1^k is randomly chosen to be noised, with uniform probability. A noise tensor x_0^k of same dimensions as x_1^k is sampled from a standard multivariate Gaussian distribution, and the chosen image x_1^k is replaced with a noised version x_t^k following the linear OT path often used in flow matching models:

$$x_t^k = tx_1^k + (1-t)x_0^k$$

where t follows a standard lognormal distribution. We refer the reader to the original flow matching paper [4] for the full details on flow matching training and sampling.

2.4 Architecture

We use the MOTIF architecture [1], which can be applied to non-co-located multi-source geospatial data. This architecture notably uses the geographic coordinates (latitude and longitude) at each pixel within the model, which allows the model to represent any geometry, including in our case the irregular geometries of the satellite swaths. Finally, the sources are treated symmetrically inside the architecture, which lets us noise different sensors within a common batch during training (see section 2.3). As the architecture was originally defined for deterministic training, we modify it to incorporate the ingredients specific to flow matching: the [MASK] token that normally replaces the masked source is replaced by the noised image; the flow matching timestep t is embedded with a linear layer and summed into the conditioning embedding alongside the source characteristics and the land-sea mask.

3 Results

We evaluate the predictions using the Continuous Ranked Probability Score (CRPS), which is one of the most commonly used metrics for probabilistic predictions [9]. For every configuration, we use 10 realizations per sample. We evaluate different training and evaluation configurations, and present the results in Table 1.

First, we compare two training strategies by letting the time window vary: $\delta_{max} = 1$ is the model trained on a time window of 1 hour, i.e. the images occur nearly at the same time ("co-timed"), as the features of tropical storms barely evolve within a single hour. $\delta_{max} = 6h$ is a configuration where the model is trained on a window of 6h, which is enough for cyclones to significantly evolve. Those two configurations are evaluated on exactly the same set with a time window of 1h, i.e. nearly co-timed. Our results show an improvement on the mean CRPS, with disjoint 95% confidence intervals. We obtain a large standard deviation, which given the tight confidence intervals we interpret as the result of heavy outliers. The conclusion is that even when evaluating on co-timed images only, training on images that are further away in time (here up to 6h) brings a small but significant improvement. A clear factor is that the training set with $\delta_{max} = 6h$ is a super-set of that with $\delta_{max} = 1h$ ($\sim$40,000 samples against $\sim$11,000). In other words: allowing non-co-timed observations increases the amount of training data, compensating the fact that it is harder task than training on purely co-timed images.

Second, we compare the same training strategy against different evaluation configurations by varying two factors: (1) the number of *minimum available sources within a sample*: we only keep samples

Table 1: Results of different training configurations and number of available sources during inference (mean values, in Kelvin). The δ_{max} duration is in hours. Top section: comparison between training on nearly co-timed images ($\delta_{max} = 1h$) and training on both co-timed and non-co-timed images ($\delta_{max} = 6h$). Middle section: comparison between limiting the input to 2 sources against 3 or more. Bottom section: comparison between limiting the input to 2 sources against 4 or more. Within each section, all experiments are evaluated over exactly the same samples.

Training config.	Eval config.			CRPS↓
	δ_{max}	Min. available sources	Sources used as input	
$\delta_{max} = 1h$	1	2	All	5.09[5.04,5.14](2.56)
$\delta_{max} = 6h$	1	2	All	**4.76** [4.71,4.81](2.37)
$\delta_{max} = 6h$	6	3	2	4.74[4.70,4.78](2.17)
$\delta_{max} = 6h$	6	3	All	**4.45**[4.42,4.49](1.96)
$\delta_{max} = 6h$	6	4	2	4.70[4.62,4.77](2.09)
$\delta_{max} = 6h$	6	4	All	**4.29**[4.22,4.35](1.84)

Brackets show 1000-bootstrap 95% confidence intervals. Parentheses show standard deviation.

for which at least a given minimum number of different sources are available within the 6h-time window. (2) the number of *maximum sources actually passed as input to the model*: if more than a given number of sources are available in the time window, only that maximum amount is actually fed into to the model, and the rest of the sources are ignored. Throughout all experiments, the sources that are closest to t_0 are kept in priority. The CRPS values show that when at least 3 or 4 sources are available, constraining the model to only using two leads to worse performance on the same samples, supporting the hypothesis that multiple sensors can bring complementary information as input and that the model can take advantage of it.

4 Conclusion

Those early results suggested that training on non-co-located satellite observations can help tackle the problem of cross-sensor harmonization. We showed that this hypothesis is verified, at least for the specific case of the GPM constellation over tropical cyclones. We pointed out that the task of cross-sensor harmonization itself can be used as a pretext task for self-supervised training, and that it is compatible with using generative model. We also show on this specific data that using multiple sources as input does improve the performances.

5 Discussion

While the results encourage us to keep experimenting with our setting, there are shortcomings that should be accounted for. To begin with, we only evaluate our method on a single dataset which, while it contains a reasonable amount of data and perfectly matches the task of cross-sensor harmonization, only concerns a precise subset of the global microwave imagery (tropical and extra-tropical storms). In addition, we do not here evaluate the choice of architecture, as we focus on comparing different training strategies. The main reason for that is that are very few deep learning architectures that allow multi-source data with the flexibility required here (flexible number of inputs, satellite geometry, irregular time intervals). A possible choice other than MOTIF is the Perceiver IO architecture [3], which would still need to be adapted to use spatio-temporal coordinates and a flow matching step.

Regarding the long-term, that type of work could hopefully lead to deep learning methods that combine a large amount of individual observations into a global, harmonized product. This could notably bring improvements in the field of data assimilation, essential for example in weather forecasting.

References

[1] Clément Dauvilliers and Claire Monteleoni. MoTiF: a self-supervised model for multi-source forecasting with application to tropical cyclones. *Environmental Data Science*, 4:e36, January 2025. ISSN 2634-4602. doi: 10.1017/eds.2025.10014. URL https://www.cambridge.org/core/journals/environmental-data-science/article/motif-a-selfsupervised-model-for-multisource-forecasting-with-application-to-tropical-cyclone/489085C5B41E673D1480273C2F55D01C.

[2] GPM constellation. The GPM Constellation | NASA Global Precipitation Measurement Mission. URL https://gpm.nasa.gov/missions/GPM/constellation.

[3] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A General Architecture for Structured Inputs & Outputs, March 2022. URL http://arxiv.org/abs/2107.14795. arXiv:2107.14795 [cs].

[4] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023. URL http://arxiv.org/abs/2210.02747. arXiv:2210.02747 [cs].

[5] Muhammad Naufal Razin, Christopher J. Slocum, John A. Knaff, Paula J. Brown, and Michael M. Bell. Tropical Cyclone Precipitation, Infrared, Microwave, and Environmental Dataset (TC PRIMED). *Bulletin of the American Meteorological Society*, 104(11):E1980–E1998, November 2023. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-21-0052.1. URL https://journals.ametsoc.org/view/journals/bams/104/11/BAMS-D-21-0052.1.xml.

[6] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission Critical – Satellite Data is a Distinct Modality in Machine Learning. June 2024. URL https://openreview.net/forum?id=PQ0ERKKYJu.

[7] Vibolroth Sambath, Natanaël Dubois-Quilici, Nicolas Viltard, Audrey Martini, and Cécile Mallet. Unsupervised Domain Adaptation to Mitigate Out-of-Distribution Problem of Spatial Radiometer Images: Application to Quantitative Precipitation Estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3403373. URL https://ieeexplore.ieee.org/document/10535296.

[8] Nicolas Viltard, Vibolroth Sambath, Pierre Lepetit, Audrey Martini, Laurent Barthès, and Cécile Mallet. Evaluation Of Drain, A Deep-Learning Approach To Rain Retrieval From Gpm Passive Microwave Radiometer. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2024. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2023.3293932. URL https://ieeexplore.ieee.org/document/10177741/.

[9] Michaël Zamo and Philippe Naveau. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences*, 50(2):209–234, February 2018. doi: 10.1007/s11004-017-9709-7. URL https://hal.science/hal-02976423. Publisher: Springer Verlag.