Convolutional Neural Processes for Inpainting Satellite Images: Application to Water Body Segmentation

Alexander Pondaven[*] [1]
ap2619@ic.ac.uk

Märt Bakler[*] [1]
mb1221@ic.ac.uk

Donghu Guo [†] [1]
dg321@ic.ac.uk

Hamzah Hashim [†] [1]
hh2019@ic.ac.uk

Martin Ignatov [1]
mgi18@ic.ac.uk

Samir Bhatt [1] [2]
s.bhatt@ic.ac.uk

Seth Flaxman [3]
seth.flaxman@cs.ox.ac.uk

Swapnil Mishra [2]
s.mishra@ic.ac.uk

Elie Alhajjar [4]
elie.alhajjar@westpoint.edu

Harrison Zhu [1]
hbz15@ic.ac.uk

Abstract

The widespread availability of satellite images has allowed researchers to monitor the impact of climate on socio-economic and environmental issues through examples like crop and water body classification to measure food scarcity and risk of flooding. However, a common issue of satellite images is missing values due to measurement defects, which render them unusable by existing methods without data imputation. To repair the data, inpainting methods can be employed, which are based on classical PDEs or interpolation methods. Recently, deep learning approaches have shown promise in this realm, however many of these methods do not explicitly take into account the inherent spatio-temporal structure of satellite images. In this work, we cast satellite image inpainting as a meta-learning problem, and implement Convolutional Neural Processes (ConvNPs) in which we frame each satellite image as its own task or 2D regression problem. We show that ConvNPs outperform classical methods and state-of-the-art deep learning inpainting models on a scanline problem for LANDSAT 7 satellite images, assessed on a variety of in- and out-of-distribution images. Our results successfully match the performance of clean images on a downstream water body segmentation task in Canada.

1 Introduction

Monitoring climate change requires the analysis of land features at a granular spatio-temporal level. With the surge of computational methods using remote sensing data, satellite images have been widely used in instances such as flood detection [1] and crop yield modelling [2]. The LANDSAT 7 satellite [3] is an invaluable source of satellite images to understand these trends and take appropriate action based on accurate climate models due to its long temporal coverage and high spatial resolution. However, as a result of a mechanical fault in the satellite's scanline corrector (SLC), satellite images taken from May 31, 2003 onward suffer from lines of missing pixels (Figure 3). As they occupy a significant area of the satellite images (about 20% of the data), the images obtained from LANDSAT 7 lost much of their research use in climate-related downstream tasks as the scanlines significantly impair the performance of computational methods using the corrupted images. However, existing ML or traditional satellite image models are able to process imputed images as opposed to retraining

[*] Equal contribution. [†] Order decided by a coin toss.

[1] Imperial College London. [2] University of Copenhagen. [3] University of Oxford.

[4] United States Military Academy, Workshop mentor

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

models or explicitly taking into account missing data, which is currently required for corrupted images.

Image inpainting (gap-filling) aims to fill the corrupted pixels in an image with values that resemble the original pixel values as closely as possible. Many deterministic methods have been proposed in the literature, that use higher order differential equations [4–6]. Moreover, recent advances in deep learning have shown promising results such as U-Net [7], which was initially used for biomedical image segmentation, and Partial Convolutions (PartialConv) [8], which is a modification to the classical convolutional layer to make it suitable for inpainting. One drawback of the traditional deep learning methods is that they treat all images as a single task and do not take into account spatio-temporal differences between images, where different predictive functions could better suit different tasks or images. This kind of problem is better suited to meta-learning methods, which learn task-specific representations and are better at capturing the differences between various inputs. Garnelo et al. [9] introduced a meta-learning approach called Conditional Neural Processes (CNPs) that uses an encoder-decoder architecture to learn a distribution over predictive functions. Gordon et al. [10] and Foong et al. [11] introduced Convolutional Conditional Neural Processes (ConvCNPs) and Convolutional Latent Neural Processes (ConvLNPs) respectively, which are better suited for image inpainting tasks due to their translational equivariance property. These Convolutional Neural Processes are shown to exhibit very good few-shot and zero-shot learning capabilities as well, which has been demonstrated for inpainting weather data [11, 12].

In this paper, we show that ConvNPs can be used for satellite image inpainting, particularly to correct the scanlines of LANDSAT 7 images. We use an MS-SSIM similarity score loss function [13] for sharper results, which preserves important land features for segmentation or regression tasks. We show that our ConvNPs outperform state-of-the-art image inpainting models with a relatively small dataset (training set of 800 images with dimensions 128x128 or 64x64). ConvNP models also show good performance for both in-distribution (inpainting images in the same country as the training dataset) and out-of-distribution (OOD) satellite images (considered as zero-shot tasks, inpainting images in a different country than the training dataset). In addition, via an OOD downstream water body segmentation problem, we show that ConvNP imputed images achieve similar performance as compared to the clean images in the training phase.

2 Methodology

We cast satellite inpainting as a meta-learning problem. The pixel locations on the grid and RGB pixel values at those locations are denoted as $x \in \mathbb{R}^2$ and $y \in \mathbb{R}^3$ respectively. Each image corresponds to a task, which could also be viewed as a 2D function [14, 15]. At prediction time, the observed set of pixels or "context set" is denoted by x_C, y_C. The aim is then to predict the target values y_T at locations x_T (in our case the entire image is predicted to avoid discontinuities). We argue that taking the meta-learning viewpoint allows to **explicitly** take into account the spatio-temporal variations for each task and thus promotes efficient learning as opposed to classical inpainting methods like U-Net and PartialConv which **implicitly** distinguish between different tasks and require enormous training sets with data augmentation.

Meta-learning methods [16, 17] aim to solve the problem of using a distinct function at inference time to predict target set values. The NP family [9, 18] architecture employs an encoder that outputs a task-specific representation of the context points, which can then be queried with a decoder network to give a task-specific output function distribution. Gordon et al. [10] introduced translational equivariance with ConvCNPs, making it more suitable for image data (on-the-grid data) with the use of CNNs. Foong et al. [11] presents Convolutional Neural Processes (ConvNPs, in this paper referred to as Convolutional Latent Neural process or ConvLNP) that utilise a latent variable to capture information from the context set. This results in a model that adjusts the predictor function depending on the context set of the task. One advantage of meta-learning is that the predictive functions use both the information from the current context set of the task as well as the information that is shared across tasks, making the method well-suited for OOD tasks. This allows modelling of heterogeneous function distributions and is a beneficial property for satellite image inpainting as they have multiple zero-shot tasks for different spatial locations and times, that are not seen during training. See Appendix D for in-depth model details.

3 Experiments

3.1 Inpainting LANDSAT 7 images

We study the performance of ConvCNPs and ConvLNPs for the task of inpainting LANDSAT 7 scanlines. We compare the results to the baseline models consisting of Navier-Stokes (NS) inpainting algorithm, U-Net and PartialConv, for which the latter two are vastly popular and yield state-of-the-art results on a variety of image inpainting problems. We train the ConvCNP, ConvLNP and U-Net using the MS-SSIM loss function, and for PartialConv we use the loss proposed in Liu et al. [8]. We conduct extensive experiments to measure the in-distribution performance of each model by inpainting satellite images of the same country, as well as OOD performance on a set of different countries (zero-shot prediction over unseen spatial locations). To evaluate the performance of all the models, we compute the MS-SSIM between the predictions and the ground truth images and the MSE score on just the scanline pixels. The MS-SSIM is bounded in $[0, 1]$, where values closer to 1 show that the images are more similar.

Data The training images are acquired from the LANDSAT 7 Satellite before the scanlines are present in the images. The training set consists of 1000 images from Kenya with dimensions 128x128 and 64x64. See Appendix C for a detailed description of the data collection process. The scanlines are acquired from 100 Kenya images post-SLC failure. We perform 5-fold cross validation with a 80%-20% train-test ratio for each split. During training, a scanline is applied to each image as a mask chosen randomly from the 100 scanlines extracted. In-distribution inpainting performance is reported using unseen images from Kenya and out-of-distribution inpainting performance is reported using unseen images from UK, Nepal, Brazil and Norway.

Results. The MS-SSIM and scanline RMSE scores for 128x128 images can be seen in Figure 1. Scores for 64x64 images and examples of 128x128 image imputations can be found in Appendix B. As can be seen both empirically and by the MS-SSIM and MSE scores, ConvNPs perform well both in-distribution and OOD with ConvLNP performing the best overall with its latent flexibility. U-Net achieve good results on the test set for the Kenya images it was trained on, but suffers in performance on other countries, thus lacking the zero-shot capabilities of ConvNPs. PartialConv inpainted results were blurry with discolouration and may need longer training time with a larger dataset to reflect the performance of the original paper [8]. The meta-learning approach treats input images as different tasks and hence the variability between images and their characteristics is better accounted for. Note that UK inpainting performance is artifically increased by the high cloud frequency, making it an easier task. NS generally results in blurry scanline imputations, especially between the border colours of neighbouring pixels. It is also notable that NS is not a machine-learning algorithm, hence it is location agnostic and there is no concept of in- and out-of-distribution datasets.

3.2 Climate Downstream task

OOD performance of imputed results has been evaluated on a downstream segmentation task to classify seasonality of water in Canada. This involves classifying all pixels into three classes: "not

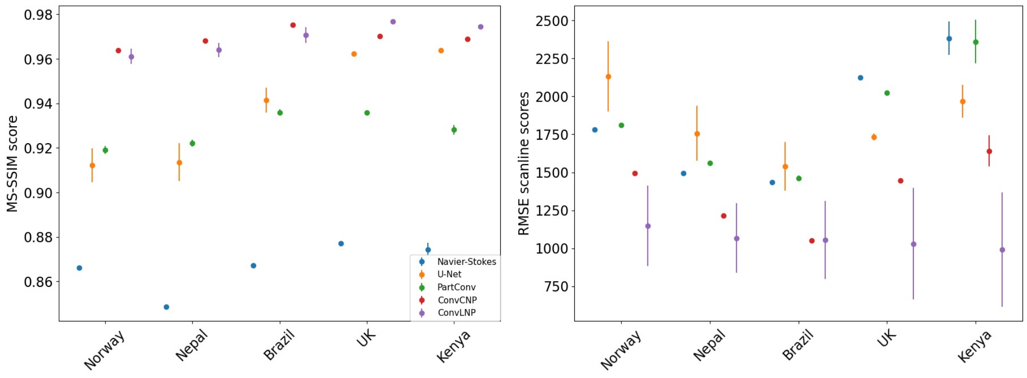


Figure 1: Mean and standard error of the MS-SSIM scores (left) and MSE scores (right) on scanlines over 5-fold cross validation for predicting over Kenya and OOD datasets for images of dimension 128x128 (64x64 results in Figure 6).

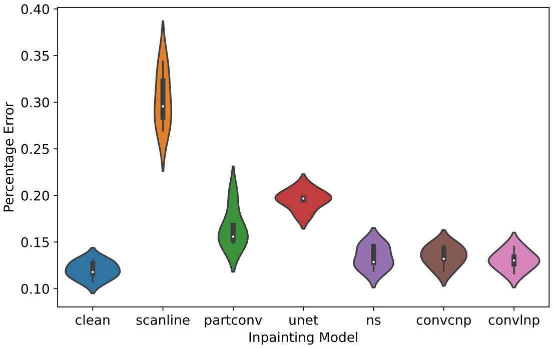


Figure 2: Water segmentation task percentage pixel error on 128x128 images.

water", "seasonal water" and "permanent water" based on three months of satellite images at each location in 2000. Due to lack of data around water sources, a masked binary cross entropy loss is applied to evaluate loss on only labelled pixels. A 3D convolution operation is applied to merge the temporal information, followed by a U-Net for segmentation. Clean images are downloaded from Google Earth Engine [19]. The full set of scanlines are applied randomly to corrupt images and are imputed by the different methods. The segmentation model is trained on clean images and evaluated over imputed images for each model, clean images, and corrupt images (scanline). The percentage pixel errors for segmentation over 5-fold cross validation is reported in Figure 2. ConvNPs can be seen to outperform other deep learning baselines and performs very similarly to clean images, which the segmentation models were trained on. ConvNP imputation performance is also comparable to that of NS (which is not trained on any particular location) in the OOD setting as it does not change non-scanline pixels and ConvNPs appear to have slight discolouration. This shows the use of imputed results for existing climate models and how imputed results can be used as a substitute for clean images when not available.

4 Discussion and Conclusion

We find that ConvNPs are successful at inpainting LANDSAT 7 satellite images corrupted by scanlines in both in-distribution and out-of-distribution tasks, outperforming classic and state-of-the-art inpainting methods. This enables the use of LANDSAT 7 data in existing models solving climate-related tasks just by imputing results as shown in the water body segmentation task. When models are trained on corrupted images, their performance is only slightly lower than that of models trained on clean images. A potential reason for such behaviour is the high entropy nature of these images, leading to CNNs having difficulty learning. In this work, the main success is in improving on existing inpainting models and allowing for easy integration of cleaned LANDSAT 7 data to monitor climate trends.

In an upcoming work, we intend to make improvements to LANDSAT 7 downstream models and measure the performance of models trained on imputed results. The main idea relies on training a more diverse dataset while fine-tuning inpainting models on new locations to enhance out-of-distribution imputations. Another line of effort involves trying other baselines like diffusion models [20], making use of recent advances in ConvNPs to improve expressiveness [21, 22, 12] and to more explicitly account for space-time considerations [23]. The current work may readily be extended to other inpainting tasks such as cloud removal and other downstream tasks to tackle climate change.

5 Acknowledgements and Funding Disclosure

Many thanks to Samir Bhatt for proposing this problem. The authors would like to thank Samir Bhatt, Seth Flaxman, Jevgenij Gamper, Iwona Hawryluk, Tom Mellan, Swapnil Mishra, Oliver Ratmann and Maxime Rischard for useful feedback and suggestions. Elie Alhajjar also helped as a workshop mentor, providing useful feedback and discussion leading up to the paper submission. The authors would also like to thank the Imperial College Data Science Society (ICDSS), especially Kenton Kwok, for organising and supporting this project as part of the Advanced Team. HZ is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1), the Department of Mathematics of Imperial College London and Cervest Limited. SB, AP and HH acknowledge The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345) that supports SB. We would also like to thank Andy Thomas for his endless support with using the NVIDIA4 GPU Compute Server.

References

[1] Sayak Paul and Siddha Ganju. Flood segmentation on sentinel-1 SAR imagery with semi-supervised learning. *CoRR*, abs/2107.08369, 2021. URL `https://arxiv.org/abs/2107.08369`.

[2] Harrison Zhu, Adam Howes, Owen van Eer, Maxime Rischard, Yingzhen Li, Dino Sejdinovic, and Seth Flaxman. Aggregated gaussian processes with multiresolution earth observation covariates, 2021. URL `https://arxiv.org/abs/2105.01460`.

[3] USGS. Landsat 7 courtesy of the u.s. geological survey. 1999.

[4] Martin Burger, Lin He, and Carola-Bibiane Schönlieb. Cahn–hilliard inpainting and a generalization for grayvalue images. *SIAM Journal on Imaging Sciences*, 2(4):1129–1167, 2009.

[5] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[6] Andrea Bertozzi and Carola-Bibiane Schönlieb. Unconditionally stable schemes for higher order inpainting. *Communications in Mathematical Sciences*, 9(2):413–457, 2011.

[7] 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*, pages 234–241. Springer, 2015.

[8] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *CoRR*, abs/1804.07723, 2018. URL `http://arxiv.org/abs/1804.07723`.

[9] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018.

[10] Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. *ICLR*, 2020.

[11] Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020.

[12] Stratis Markou, James Requeima, Wessel P Bruinsma, Anna Vaughan, and Richard E Turner. Practical conditional neural processes via tractable dependent predictions. *ICLR*, 2022.

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[14] Emilien Dupont, Yee Whye Teh, and A. Doucet. Generative models as distributions of functions. *NeurIPS*, 2021.

[15] Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo J. Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *CoRR*, abs/2201.12204, 2022. URL `https://arxiv.org/abs/2201.12204`.

[16] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[18] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[19] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, Dec 2016. ISSN 1476-4687. doi: 10.1038/nature20584. URL `https://doi.org/10.1038/nature20584`.

[20] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *CoRR*, abs/2201.09865, 2022. URL `https://arxiv.org/abs/2201.09865`.

[21] Wessel P Bruinsma, James Requeima, Andrew YK Foong, Jonathan Gordon, and Richard E Turner. The gaussian neural process. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

[22] Stratis Markou, James Requeima, Wessel Bruinsma, and Richard Turner. Efficient gaussian neural processes for regression. *ICML 2021 Workshop on Uncertainty and Robust- ness in Deep Learning*, 2021.

[23] Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. Sequential neural processes. *Advances in Neural Information Processing Systems*, 32, 2019.

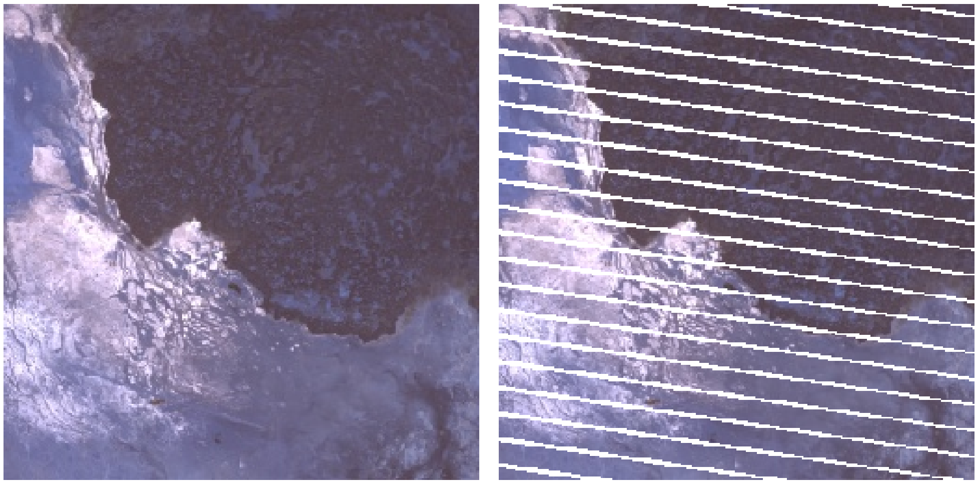
[24] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL `https://doi.org/10.1016/j.rse.2017.06.031`.

[25] 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, pages 1398–1402. Ieee, 2003.

[26] Yann Dubois, Jonathan Gordon, and Andrew YK Foong. Neural process family. `http://yanndubs.github.io/Neural-Process-Family/`, September 2020.

Appendix

A Additional Images



(a) Before 31st of May 2003

(b) After 31st of May 2003

Figure 3: LANDSAT 7 images before and after the scanline corrector failure.

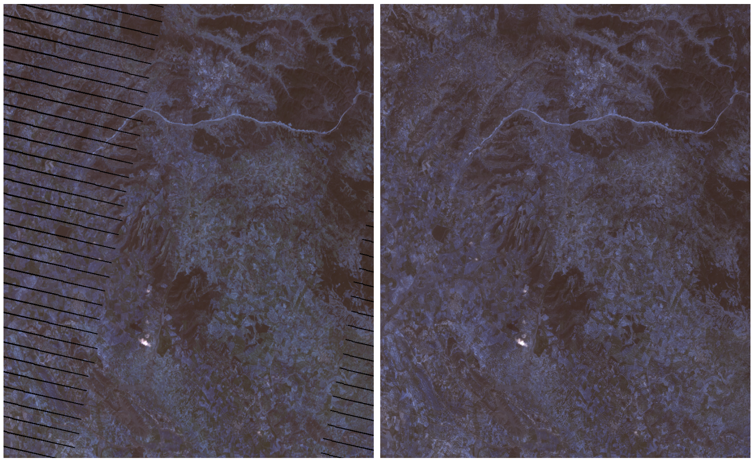


Figure 4: Kenya 1024x1024 by predicting on 64x64 patches with ConvCNP. (Left) Original image (Right) Inpainted image.

B Imputation Results

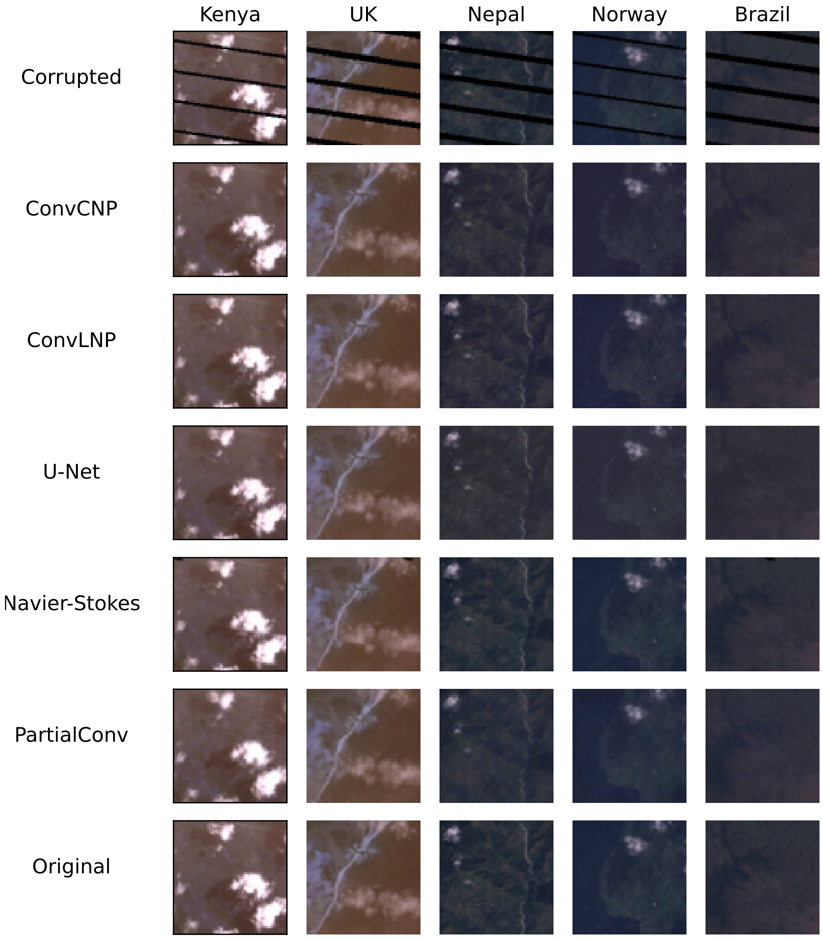


Figure 5: Inpainting predictions for all models on 128x128 images for all models over multiple regions with thinner scanline set applied.).

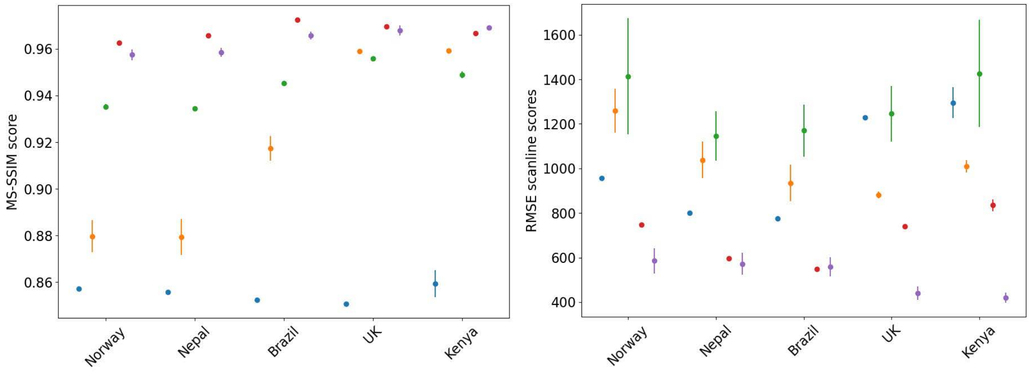


Figure 6: Mean and standard error of the MS-SSIM scores (left) and MSE scores (right) on scanlines over 5-fold cross validation for predicting over Kenya and OOD datasets for images of dimension 64x64. Note that standard errors lower than 0.01 have not been visualised.

C Data collection and preprocessing

LANDSAT 7 images are downloaded using the Google Earth Engine (GEE) API [24]. For this paper, we focus solely on the visible RGB bands of the LANDSAT 7 satellite (B3, B2, B1) with spatial resolutions of 30 meters. The images are sampled from a uniform spatial grid with a grid spacing of 0.4 degrees longitude and latitude. They are downloaded with a dimension of 256x256 pixels, corresponding to a land area of approximately $59km^2$. Due to computational limitations, these images are cropped to 64x64 and 128x128, and model results are reported using these sizes. Note that smaller satellite images could be patched together to perform a larger inpainting task as seen in Figure 4.

The images are extracted from specific dates and locations, and are divided into pre- and post-2003 (non-working SLC period). All pre-2003 data is from between 1999 to 2003. Post-2003 data is collected from between 2003 to 2004. Images with missing pixels (alpha channel is present) are filtered out for pre-2003 images. Satellite images collected from Kenya are used for training the models. Satellite images collected from UK, Norway, Brazil and Nepal are used to test the model's capabilities on out-of-distribution (unseen, location-wise) images. UK images are sometimes completely white due to the presence of clouds, which resulted in 'better' inpainting results across all models. To create more challenging out-of-distribution tasks, Norway, Brazil and Nepal images are filtered by only taking images where the middle 64x64 section had less than 90% white pixels (so the cropped 64x64 dimension dataset also had less clouds). Post-2003 data is used to extract a set of 100 scanline bit masks from Kenya data to apply to un-corrupted pre-2003 images during training to have access to the clean ground-truth. Some images had large sections of missing pixels, so the post-2003 images are filtered to have $< 20\%$ missing pixels, but also at least 100 missing pixels (set arbitrarily) to ensure the presence of scanlines.

D Model details

In the meta-learning setting, during training, we learn a global parameter θ, which, given contexts of a task, could also output a **task-specific** representation R_m. The global objective function is given by $\mathbb{E}_{m\sim\mathcal{M}}[\mathcal{L}(D_\eta(E_\xi(x_{C_m}, y_{C_m}))(x_{T_m}), y_T)]$, with $D_\eta(E_\xi(x_{C_m}, y_{C_m}))(x_{T_m}) \approx f_{\theta_m}(x_{T_m})$, where $\theta = (\eta, \xi)$, E_ξ encodes the context set (x_C, y_C) to a task-specific representation, D_η decodes the task-specific representation and target location to the output, and $\mathcal{L}$ is a loss function.

Convolutional Conditional Neural Processes: Gordon et al. [10] introduced translational equivariance to the NP family [9, 18] through ConvCNPs, making it more suitable for image data (on-the-grid data). With the same notation, we denote the original image as I and the context mask M_C, for which $[M_C]_{i,j} = 1$ if the pixel at location (i, j) is in the context set, and 0 otherwise. Our masked context set is thus given by $Z_C = M_C \odot I$. Concatenating the context mask and the masked context point, we thus get $\phi = [M_C, Z_C]$. Applying a convolution to ϕ, we obtain the functional representation $R = Conv_\theta([M_C, Z_C]^{)}$, where $Conv_\theta$ is the 2D convolution operator with positively-constrained kernel parameters θ. We then apply the normalisation $R^{(1:C)} = R^{(1:C)}/R^0$. This step is known as **SetConv** (when not evaluated at the target points). We can decode R using a CNN, which includes an absorbed MLP to map the output of the CNN at each location (i, j) to $\mathbb{R}^2$ and gives μ, the image prediction.

Convolutional Latent Neural Processes: Foong et al. [11] presents the Convolutional Neural Processes (ConvNPs, in this paper referred to as Convolutional Latent Neural process or ConvLNP) that utilise a latent variable to capture information from the context set. It is similar in architecture to the conditional neural process with the encoder-decoder architecture, but in the ConvLNP the encoder outputs a distribution over the latent variable $\mathbf{z}$ with the SetConv representations: $\mathbf{z} \sim p(\mathbf{z}|R)$. This enables ConvCNPs to learn 'richer joint predictive distributions' [11] and handle multimodalities. The full computational graphs for ConvCNP and ConvLNP are described in Figure 7.

Training objective: Following Foong et al. [11], we use the maximum likelihood training objective for the ConvNPs: $\mathcal{L} = \log p(y_T|x_T, C)$, but we instead use the MS-SSIM metric (Multi Scale Structural Similarity, Wang et al. [25]) between the mean predictions and ground truth images. MS-SSIM is a structural similarity metric for images, and is widely used in the field of signal processing, having shown empirically to increase sharpness of final prediction images. In the ConvLNP training objective, the maximum likelihood approach uses sample estimates to approximate the likelihood of the predictions: $\mathcal{L} \approx \frac{1}{L} \sum_{l=1}^{L} \log p(y_T | \mathbf{z}, x_T, C)$.

ConvNP training: Our implementation follows Dubois et al. [26]. Both ConvCNPs and ConvLNPs use Resnet blocks in the encoder and linear MLPs in the decoder. The ConvCNP has a 10-layer ResNet encoder with a representation size of 128 channels and the decoder MLP has 4 layers. It is trained for 400 epochs, with batch size 8 and learning rate 10^{-4}, which decays exponentially by a factor of 5. The ConvLNP model for 128x128 images is trained for 200 epochs with a batch size of 4 (low batch size due to computational limitations) and during training, 4 samples are obtained of the latent variable while during evaluation, 8 samples are used. For 64x64, the latent samples are increased for ConvLNP, namely 16 latent samples are used for training and 32 are used during inference. Both Resnets used in ConvLNP have 8 layers. Our code and data will be open-sourced upon publication.

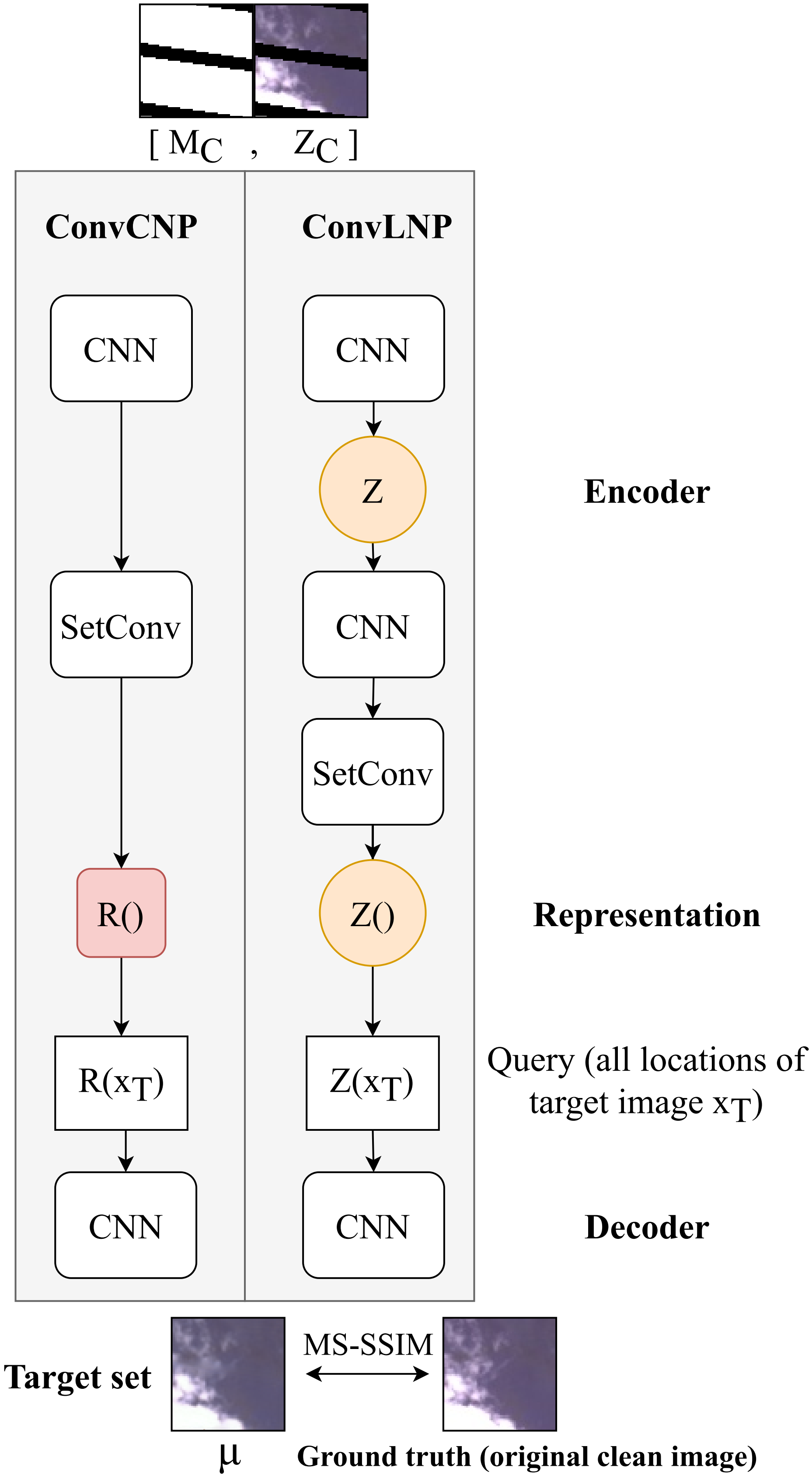


Figure 7: ConvCNP and ConvLNP on-the-grid architecture.