ESC-GAN: Extending Spatial Coverage of Physical Sensors

Xiyuan Zhang, Ranak Roy Chowdhury, Jingbo Shang*, Rajesh Gupta*, Dezhi Hong*
University of California, San Diego, La Jolla, CA, USA
xiyuanzh@ucsd.edu,{rrchowdh,jshang,gupta,dehong}@eng.ucsd.edu

ABSTRACT

Scientific discoveries and studies about our physical world have long benefited from large-scale and planetary sensing, from weather forecasting to wildfire monitoring. However, the limited deployment of sensors in the environment due to cost or physical access constraints has lagged behind our ever-growing need for increased data coverage and higher resolution, impeding timely and precise monitoring and understanding of the environment. Therefore, we seek to *extend the spatial coverage* of analysis based on existing sensory data, that is, to "generate" data for locations where no historical data exists. This problem is fundamentally different and more challenging than the traditional spatio-temporal imputation that assumes data for any particular location are only partially missing across time. Inspired by the success of Generative Adversarial Network (GAN) in imputation, we propose a novel ESC-GAN. We observe that there are *local* patterns in nearby locations, as well as trends in a *global* manner (e.g., temperature drops as altitude increases regardless of the location). As local patterns may exhibit at different scales (from meters to kilometers), we employ a multi-branch generator to aggregate information of different granularity. More specifically, each branch in the generator contains 1) randomly masked 3D partial convolutions at different resolutions to capture the local patterns and 2) global attention modules for global similarity. Next, we adversarially train a 3D convolution-based discriminator to distinguish the generator's output from the ground truth. Extensive experiments on three geo-sensor datasets demonstrate that ESC-GAN outperforms state-of-the-art methods on extending spatial coverage and also achieves the best results on a traditional spatio-temporal imputation task.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Data mining**; • **Applied computing**;

KEYWORDS

Spatio-temporal data; imputation; super resolution; self-attention; generative adversarial network

ACM Reference Format:
Xiyuan Zhang, Ranak Roy Chowdhury, Jingbo Shang*, Rajesh Gupta*, Dezhi Hong*. 2022. ESC-GAN: Extending Spatial Coverage of Physical Sensors. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3488560.3498461

1 INTRODUCTION

Wide-scale deployment of environmental geo-sensors have advanced our understanding of our ecosystem and its evolution. These sensors enable us to observe the very fabric of our surrounding physical world. For example, outdoor thermometers detect seasonal patterns and annual shift in temperature to help understand global warming [12]; rain gauges measure the precipitation level for hydrological modeling, flood forecasting, and agricultural purposes [1]; magnetometers monitor the earth's magnetic field, helping advance magnetosphere studies [13], just to name a few.

While tremendously valuable, a critical limitation of these geo-sensors is that each of them covers only a fraction of the total area, and it is impossible for a majority of them to cover the entire planetary surface. Constructing and maintaining sensing stations incurs high costs, and many locations are often inaccessible due to physical access constraints such as harsh environmental conditions and urban development. Consequently, sensors are usually sparsely distributed across the globe, limiting deeper understandings of large-scale phenomena. Figure 1a shows an example of the limited availability of magnetic field monitoring stations on the earth.

In this paper, we seek a cost-effective approach to extend the spatial coverage of existing planetary sensory data without deploying additional sensors. We name this task as *extending spatial coverage* (ESC) of sensor data. To be specific, our goal is to "generate" data at locations where no historical values are ever recorded and extend the spatial coverage to the entire globe (as illustrated in Figure 1b). To formulate the ESC problem, we assume the entire globe (denoted as D) is gridded into $M \times N$ cells. Given the data from a set of observed or partially observed grid cells D_O (i.e., where we have sensory measurements), we aim to generate data for the remaining unobserved grid cells $D_U = \{D \setminus D_O\}$.

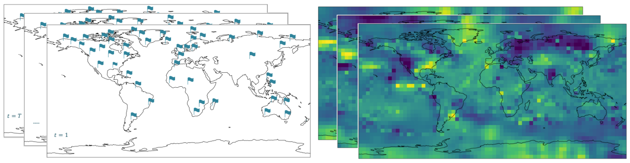
The ESC problem faces unique challenges and opportunities:

- *Complete lack of temporal dimension information.* In ESC task, we have no prior data or knowledge for unobserved grid cells. Therefore, traditional (spatio-)temporal imputation models, which assume data are partially missing in the temporal domain [6, 25, 26, 28, 47, 51], cannot solve the ESC problem. Spatial imputation methods could be applied to each time snapshot separately [10, 14, 17, 37]. They, however, miss the temporal trends from observed grid cells. Spatio-temporal interpolation methods [3, 20, 21, 38, 45], on the other hand, do not sufficiently exploit the spatio-temporal properties (e.g. global context and multi-scale structure) for imputation.
- *Existence of global context.* Sensors at a distance might exhibit similar readings due to similar geographical contexts (e.g., similar landform) [47]. This complements the first law of geography

[*] Corresponding authors.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA.

ACM ISBN 978-1-4503-9132-0/22/02.
https://doi.org/10.1145/3488560.3498461



(a) Original Data Coverage (b) Extended Data

Figure 1: Our task of extending spatial coverage: based on the sparse sensory measurements over time in (a), we aim to extend its coverage to the entire globe as shown in (b).

("...near things are more related than distant things") [41] and inspires us to explore global contexts. Solutions to image and video inpainting [8, 24, 44, 49] typically consider only local context (e.g., patterns in nearby pixels), and thus cannot capture global patterns. It is necessary to combine global and local context views of the spatio-temporal data.

- *Multi-scale structure.* Spatio-temporal data often exhibit multi-scale structures. Specifically, while fine-grained data of a particular cell can reveal accurate and detailed local patterns, coarse-grained data distributed across a large area present a "macro" view, which is less sensitive to local missing information. Therefore, we need to jointly consider coarse-level information for completeness and fine-level information for accurate modeling of planetary sensory data.

We propose a novel framework ESC-GAN to address all these challenges. Figure 2 illustrates our proposed model architecture. ESC-GAN comprises a generator and a discriminator following the Generative Adversarial Network (GAN) framework [15], as GAN has been widely used to model the complex distribution in spatio-temporal data [22, 25–27, 40, 52]. For the generator, we leverage local 3D partial convolutions together with global attention modules to focus on both local (e.g., patterns in nearby locations) and global (e.g., phenomena that exist regardless of location) contextual information. We also design a multi-branch encoder for aggregating information of different granularity. The discriminator is composed of 3D convolutional layers. Extensive experiments on three geo-sensor datasets have verified the effectiveness of ESC-GAN under different missing data scenarios. Moreover, as our model does not impose any additional constraints or assume any domain knowledge, it can also be applied to other spatio-temporal problems like urban environment monitoring, traffic estimation, etc.

In summary, we explore a challenging task of extending spatial coverage, where we attempt to generate data at locations with no historical observations. We have analyzed the unique challenges and opportunities, which guide our model design. Our main contributions are summarized as follows:

- We propose a novel ESC-GAN framework that can address all identified challenges.
- We leverage 3D partial convolutions to learn the local correlations in both spatial and temporal dimensions, global attention modules to capture global contextual information, and a multi-branch encoder to exploit information of different granularity.
- Extensive experiments on three real-world datasets have demonstrated the superiority of ESC-GAN over all the compared methods, including state-of-the-art spatio-temporal imputation methods, image and video inpainting methods.

2 ESC PROBLEM FORMULATION

In the extending spatial coverage (ESC) problem, we aim to learn to generate data at locations where data are completely missing at all timestamps $t = 1, \dots, T$. Formally, we denote a grid cell as S_{ij}, where $i = 1, \dots, m$ and $j = 1, \dots, n$. Let $\mathbf{X} \in R^{t \times m \times n}$ denote the data, and $\mathbf{M} \in \{0, 1\}^{t \times m \times n}$ denote the corresponding masking matrix. If $M_{t,i,j} = 1$, it means $S_{i,j}$ is valid at time t (i.e. we have data in grid cell $S_{i,j}$ at time t); otherwise, $S_{i,j}$ is invalid at time t (i.e. data is missing in cell $S_{i,j}$ at time t). Let Φ denote the set of missing timestamps. In the ESC problem, we do not have any data available in certain grid cells. Therefore, assuming $S_{i,j}$ is a grid cell that is unobserved, we have $M_{t,i,j} = 0$ for all timestamps $t \in \Phi = \{1, \dots, T\}$. Given data observed at other locations $\{X_{t,i,j} | M_{t,i,j} = 1\}$ for $t = 1, \dots, T$, our goal is to generate data for all the unobserved grid cells and output a complete set of grid cells $\{\tilde{X}_{t,i,j}\}$ at all timestamps $t = 1, \dots, T$. This is in contrast to the traditional spatio-temporal imputation problem, where a given missing grid cell S_{ij} is only *partially unobserved*, i.e., $M_{t,i,j} = 0$ for $t \in \Phi \subseteq \{1, \dots, T\}$.

3 OUR ESC-GAN FRAMEWORK

As illustrated in Figure 2, ESC-GAN comprises a generator of UNet-like structure [34] and a discriminator consisting of multiple 3D convolutional layers. The generator combines local partial 3D convolution with global attention module in multiple branches. We next detail our ESC-GAN framework.

3.1 Randomly Masked 3D Partial Convolution

Spatio-Temporal data display local similarity, so we first apply convolutions to model the local patterns. Vanilla convolutions on grid map would treat each grid cell equally as valid. To obtain the output $Y_{t,y,x}$, we calculate (for simplicity we omit the bias term in the formula):

$$Y_{t,y,x} = \sum_{k=-k'_t}^{k'_t} \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_{k'_t+k, k'_h+i, k'_w+j} \cdot X_{t+k, y+i, x+j},$$

where $X_{t,y,x}$ and $Y_{t,y,x}$ represent the input and output of a specific convolution layer, respectively; $\mathbf{W}$ represents convolution filter weights; and $k'_t = \frac{k_t - 1}{2}$, $k'_h = \frac{k_h - 1}{2}$, $k'_w = \frac{k_w - 1}{2}$ with k_t, k_h, k_w representing kernel size along the time, height, and width dimensions, respectively. From the equation we can see that $Y_{t,y,x}$ is based on all the grid cells within the receptive field, regardless of whether the corresponding cells contain valid values. These invalid values involved introduce bias into the training process.

One way to address this problem is to apply partial convolutions [24] only on valid grid cells, i.e., locations with data in our context. For each iteration, we apply a training mask $\mathbf{M}$ removing a random subset of locations, as shown in Figure 3. For visualization purpose, $\mathbf{M}$ is shown as rectangle in Figure 3, but in practice $\mathbf{M}$ is randomly scattered across the map. $\mathbf{M}$ randomly masks out locations over all the cells. If the original cell has no data, then

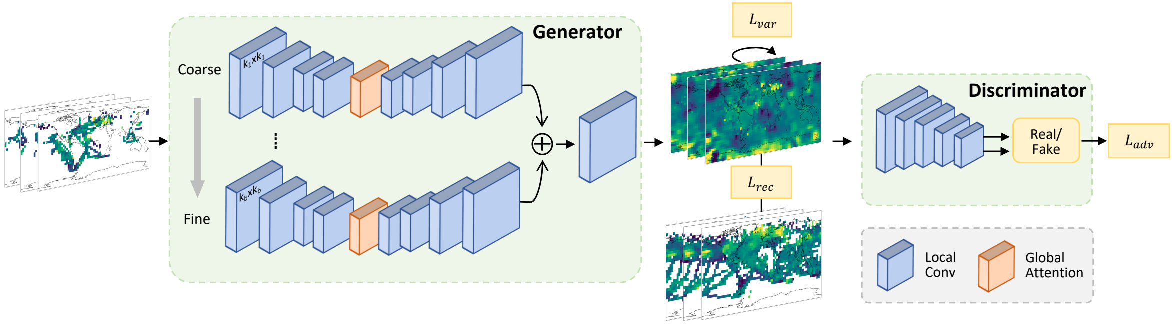
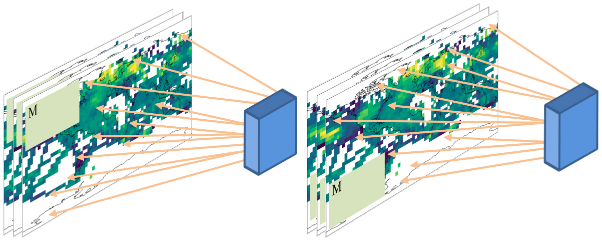


Figure 2: An overview of our ESC-GAN: The multi-branch generator takes as input grid maps with data missing at many locations, and the generator then produces grid maps with all the missing grid cells recovered. We feed the recovered maps together with ground-truth maps to the discriminator for a real or fake classification. We combine three loss functions, i.e., reconstruction loss, variation loss, and adversarial loss.



(a) iteration *i* **(b) iteration *j***

Figure 3: Randomly Masked 3D Partial Convolution. We randomly mask out parts of the grid maps on different iterations (shown as M). Blue rectangle denotes convolution filter, and orange arrows show that convolution operation is only performed on unmasked grid cells.

it does not change after being masked. We defer more complex masking strategy that coordinates missing distributions as future work. If $M_{ij} = 0$ then we mask out the corresponding grid cell S_{ij} at all timestamps; otherwise, we keep the data at this grid cell. Such masking during training helps the model learn how to recover data over time in the masked-out cell. Moreover, since the training mask is randomly generated during each iteration, it is able to cover the entire map and introduces minimum bias of region difference.

Formally, to generate data at missing locations, we extend partial convolutions to *randomly masked 3D partial convolutions*:

$$\mathbf{Y} = \begin{cases} \mathbf{W}^T(\mathbf{X} \odot \mathbf{M})\frac{|\mathbf{1}|_1}{|\mathbf{M}|_1} + b, & \text{if } |\mathbf{M}|_1 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Here, we keep the same notation such that $\mathbf{X}, \mathbf{Y}, \mathbf{W}$ represent the input, output, and convolution filter weights, respectively, in a specific convolution layer, and b is the bias term. $\mathbf{1}$ has the same size as $\mathbf{M}$ with all values being 1. Therefore, $\frac{|\mathbf{1}|_1}{|\mathbf{M}|_1}$ serves as a scaling factor to compensate for the number of valid grid cells. We also apply mask updating after each layer following prior study [24]: if the output is able to condition on at least one valid input, then the corresponding mask after updating would be 1.

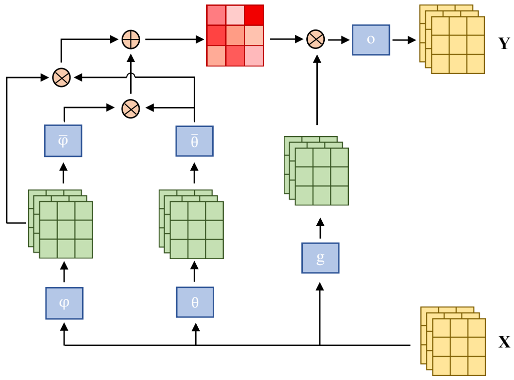


Figure 4: Global attention module architecture. The input X is transformed through three linear embeddings θ, ϕ, g. We calculate pairwise and unary scores, and multiply the scores with embeddings to obtain output Y.

3.2 Global Attention Module

In the real world, sensor values are not necessarily only correlated within a small window, in terms of both space and time. Two sensors afar could still have similar values if they share similar geographical contexts (e.g. landform) [47]. Moreover, sensor readings in different years might be similar too when they share similar temporal contexts (e.g., seasonal pattern).

The 3D partial convolution with random mask we introduce in the previous subsection only operates locally, since it mainly sums up product of kernel weights and input within local sliding windows. To take the global view into account, we incorporate global attention module as illustrated in Figure 4.

The global attention module aims to take both neighboring and distant grid cells into account for calculation. Similar to [42, 43, 50], we first apply linear embedding $\theta(\mathbf{X}) = \mathbf{W}_\theta \mathbf{X}, \phi(\mathbf{X}) = \mathbf{W}_\phi \mathbf{X}$ to embed the input. As found in [7, 48], vanilla non-local block often degenerates to purely unary term in some image recognition tasks. Therefore, we follow [48] to split the attention computation into a pairwise and a unary term, for better modeling both global pairwise similarity and global unary effect. We compute the attention scores of every two input regions $\mathbf{X}, \mathbf{X}'$ in the embedding space as

$$f(\mathbf{X}, \mathbf{X}') = \frac{e^{(\theta(\mathbf{X})-\mu_\theta)^T(\phi(\mathbf{X}')-\mu_\phi)+\mu_\theta^T\phi(\mathbf{X}')}}{\sum_{\mathbf{X}'} e^{(\theta(\mathbf{X})-\mu_\theta)^T(\phi(\mathbf{X}')-\mu_\phi)+\mu_\theta^T\phi(\mathbf{X}')}},$$

where μ_θ, μ_ϕ are average embedding values over all the regions from θ, ϕ. $(\theta(\mathbf{X}) - \mu_\theta)^T(\phi(\mathbf{X}') - \mu_\phi)$ captures pairwise long-range dependency, and $\mu_\theta^T\phi(\mathbf{X}')$ captures the global unary effect. Intuitively, $f(\mathbf{X}, \mathbf{X}')$ indicates the global context similarity between $\mathbf{X}'$ and $\mathbf{X}$. Then, to calculate the output value $O_{t,y,x}$, we calculate the weighted sum of attention scores from all input values $X_{t',y',x'}$:

$$O_{t,y,x} = \sum_{\forall t',y',x'} f(X_{t,y,x}, X_{t',y',x'})g(X_{t',y',x'}),$$

where $g(\mathbf{X}) = \mathbf{W_g}\mathbf{X}$ is also a linear embedding layer. Finally, we apply linear embedding $\mathbf{W_o}$ with residual link [16] to compute the output feature Y as

$$\mathbf{Y} = \mathbf{W_o}\mathbf{O} + \mathbf{X},$$

The above linear embedding layers $\mathbf{W}_\theta, \mathbf{W}_\phi, \mathbf{W_g}, \mathbf{W_o}$ are implemented as $1 \times 1 \times 1$ convolutions. Following [43], a batchnorm layer with scale parameter initialized to zero follows $\mathbf{W_o}$ so that the module starts from an identity mapping that relies on local information, then gradually learns the long range dependency. We follow the subsampling trick in [43] and apply max pooling after ϕ and g for computational efficiency. We leave a more efficient design for global attention module as future work.

3.3 Multi-Scale Structure Learning

Spatio-temporal data often demonstrate multi-resolution structures [25]. In our case, fine-grained data in a specific grid cell reflect accurate measurement of that cell, while coarse-grained data covering multiple grid cells are less sensitive to missing values. The one-branch encoder-decoder U-Net architecture is not able to effectively extract representations at different levels [44]. To better capture the dependencies across different scales, we adopt a multi-scale learning procedure. More specifically, we apply b parallel branches of U-Net structure in the generator. These b branches apply convolution filters of different receptive fields to extract multi-resolution features. Assume $\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_b}$ are hidden features computed after the last layer of decoders from different branches, then we concatenate these features and feed the concatenated feature into a convolution layer shared across branches to obtain the aggregated features $\mathbf{h}' \in R^{c \times t \times m \times n}$:

$$\mathbf{h}' = Conv([\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_b}]),$$

where c, t, m, n are the channel, time, height, and width dimension size, respectively. Here, the number of branches b could be decided by the input granularity, i.e., more branches when input data is fine-grained. The parallel U-Net structure overcomes limitation of the coarse-to-fine architecture where the fine stage is dependent on the coarse stage, thus being susceptible to upstream errors.

3.4 ESC-GAN Generator

As demonstrated in Figure 2, the overall architecture of ESC-GAN contains a generator G and a discriminator D. The generator contains 3D partial convolutions for learning local patterns and global attention module for learning global trends. The generator also adopts multiple branches to aggregate multi-level features. We combine grid reconstruction loss L_{rec}, variation loss L_{var}, and adversarial loss L_{adv} to optimize the generator.

To ensure grid reconstruction accuracy, we calculate Mean Square Error (MSE) between the ground-truth and generated grid maps. More precisely, let $\mathbf{X}$ and $\mathbf{Z}$ denote the ground truth and generated grid map, $\mathbf{M}$ denote the random training mask, we calculate MSE for these randomly masked out regions as

$$L_{rec} \quad = \quad \frac{1}{N_{masked}} \sum_t \sum_y \sum_x (1 - M_{t,y,x})(Z_{t,y,x} - X_{t,y,x})^2,$$

where N_{masked} denotes the number of masked out grid cells.

Apart from reconstruction loss, we also compute variation between the masked out region and valid region in the generated maps. The goal is to ensure smooth transition between masked out and valid portions. We first calculate the composite grid map $\tilde{\mathbf{X}}$, where valid regions keep the same value as the original grid map, and both randomly masked out regions and originally invalid regions are filled with generated values:

$$\tilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{M} + \mathbf{Z} \odot (1 - \mathbf{M}).$$

Then, we compute the variation over the composite grid map as

$$L_{var} = \frac{1}{N}\Big(\sum_{(y,x)\in R, (y+1,x)\in R} ||\tilde{X}_{t,y+1,x} - \tilde{X}_{t,y,x}||_1 + \sum_{(y,x)\in R, (y,x+1)\in R} ||\tilde{X}_{t,y,x+1} - \tilde{X}_{t,y,x}||_1\Big).$$

In the above equation, N is the number of grid cells in the map, and R is the 1-cell dilation of the masked region, similar to [24]. To compute the variation loss, we shift one grid cell in two spatial dimensions within R and penalize the shifted difference.

3.5 ESC-GAN Discriminator

Spatio-temporal data follow complex distributions and demonstrate high variations across time and space. They are influenced by a number of external factors (e.g. adverse weather), thus exhibiting irregular and stochastic forms. A model trained with only L_{rec} and L_{var} is not adequate to model these correlations, as it tends to output an average over different data [36]. Thus, we further train a discriminator with the generator using an adversarial strategy.

The discriminator $D(\cdot)$ is composed of 3D convolution layers. The generator $G(\cdot)$ generates grid maps $\mathbf{z} \sim P_\mathbf{Z}(\mathbf{z})$ that are indistinguishable from ground-truth grid maps $\mathbf{x} \sim P_\mathbf{X}(\mathbf{x})$, and the discriminator learns to classify feature maps as real or fake:

$$L_D = \mathbb{E}_{\mathbf{x}\sim P_\mathbf{X}(\mathbf{x})}[RELU(1 - D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim P_\mathbf{Z}(\mathbf{z})}[RELU(1 + D(\mathbf{z})],$$

$$L_{adv} = L_G = -\mathbb{E}_{\mathbf{z}\sim P_\mathbf{Z}(\mathbf{z})}[D(\mathbf{z})].$$

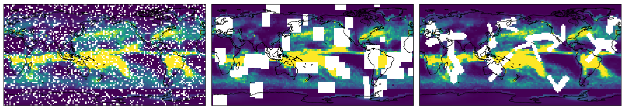
The overall training objective of ESC-GAN is a weighted sum of the reconstruction loss, variation loss, and adversarial loss, namely,

$$L = L_{rec} + \lambda_{var}L_{var} + \lambda_{adv}L_{adv}, \tag{1}$$

where λ_{var} and λ_{adv} are hyperparameters balancing between reconstruction loss, variation loss, and adversarial loss.

Table 1: Dataset Statistics

Dataset	Lat	Lon	Time	Granularity	#Grid Cells
HadCRUT	36	72	2004	5°× 5 °	5,194,368
CMAP	72	144	503	2.5°× 2.5 °	5,215,104
KDD CUP 2018	6	8	8736	0.0167°× 0.0175°	96096



(a) Scatter (b) Regular Cluster (c) Irregular Cluster

Figure 5: Three missing data distributions

4 EXPERIMENTS

We evaluate our ESC-GAN in this section. In particular, we investigate the following perspectives: (1) How does the proposed model perform compared with other baselines in the ESC task? (2) How effective are the different components in the proposed model? (3) How robust is the proposed model with respect to various missing region distributions and missing ratios? Finally, (4) How does the model perform when extended to traditional spatio-temporal imputation task with random missing data in the temporal domain? We provide both qualitative and quantitative analyses to verify the effectiveness and reliability of our model.

4.1 Datasets

We use three publicly available geo-sensory datasets to validate our proposed model. The third dataset is used to evaluate ESC-GAN for imputing random missing values in the temporal domain, which is a well-established task. Dataset statistics are summarized in Table 1.

HadCRUT[1] is a global temperature dataset, providing gridded temperature anomalies (measured by annual temperature shift) across the world [31, 32]. Temporally, the data contain monthly mean spanning from 1850 to 2020; spatially, the data covers grids of 5° latitude by 5° longitude globally (72 × 36 in total).

CMAP[2] consists of monthly averaged precipitation level values (mm/day) [46]. The data range is approximately 0 to 70mm/day. Values are obtained from 5 kinds of satellite estimates (GPI,OPI,SSM/I scattering, SSM/I emission, and MSU) and rainfall gauge. The data span from 1979 to 2020, and spatially, cover a 2.5° latitude by 2.5° longitude global grid (144 × 72 in total).

For HadCRUT and CMAP, we first normalize the dataset using z-normalization (i.e., subtracting the population mean from the individual raw stream and then dividing the difference by the population standard deviation). The whole grid map data are split into sequences of length 12, and each sequence corresponds to one-year worth of data. To simulate different real-world missing distribution, we study three types of common missing scenarios: (1) scattered locations, (2) regular clustered locations, and (3) irregular clustered locations, as visualized in Figure 5. These missing areas are left out for test set and are kept being masked out during training. The remaining area is further randomly split into 80% training set and 20% validation set. We tune the hyperparameters for both our model and baseline models on the validation set.

We also study how ESC-GAN performs on traditional spatio-temporal imputation task using the benchmark KDD CUP dataset. The **KDD CUP** 2018 dataset[3] measures hourly air quality and meterological data at city-scale. We follow previous study using this dataset [26, 28] to select 11 common locations in Beijing that measure both air quality and meterological values. Same as previous studies, we use 12 variables including PM2.5, PM10, temperature, weather, etc. To adapt the input to our model, we map the data from 11 stations into a 6 × 8 grid according to their geographical locations provided on the KDD CUP 2018 website. Then, prediction in the mapped grid cell is regarded as the prediction for the corresponding location.

We will discuss more details about the random spatio-temporal missing task in Section 4.7.

4.2 Compared Methods and Metrics

We compare ESC-GAN with four types of methods. (1) **Classical Imputation**: Mean/Zero imputation, Spatial K Nearest Neighbour (sKNN) [17], Inverse Distance Weighting (IDW) [10], Kriging [37], Matrix Factorization (MF) [29]; (2) **State-of-the-art Spatio-temporal Imputation**: Bayesian Temporal Tensor Factorization (BTTF) [11], Spatio-Temporal Multi-View Learning (ST-MVL) [47], Non-Autoregressive Multiresolution Imputation (NAOMI) [25], Inductive Graph Neural Network Kriging (IGKNN) [45]; (3) **State-of-the-art Image and Video Inpainting**: Partial Convolution (PConv) [18, 24], 3D Gated Convolution for video inpainting (3DGated) [8] ; (4) **Ablations of ESC-GAN**: single branch with only local convolution (ESC-GAN-vanilla), multiple branches with only local convolution (ESC-GAN-local), single branch with global attention (ESC-GAN-single). We provide more details about baseline implementation and settings in the supplementary material.

Following previous spatio-temporal imputation research [26, 47], we evaluate the performance of our model and baselines using $MSE(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$ (Mean Square Error) and $MAE(x, y) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|$ (Mean Absolute Error).

4.3 Experimental Setup

Our generator is a UNet-like architecture containing a four-layer encoder and a four-layer decoder with skip connections. We use partial convolutional layers instead of convolutional layers. We set the number of branches $b = 2$ for aggregating information from different granularity, based on hyperparameter tuning on the validation set. The first two layers in the coarse branch use larger filter sizes (3 × 7 × 7 and 3 × 5 × 5), the other layers use 3 × 3 × 3 filters. Both branches contain global attention modules after the last layer of decoder. Our discriminator contains five convolutional layers, with filter size 3 × 4 × 4. We optimize the model using the Adam optimizer [19] with a learning rate of 5e-3. Batch size is set to 16 for the CMAP and KDD CUP datasets, and 4 for the HadCRUT dataset. We set $\lambda_{var} = 0.1, \lambda_{adv} = 0.001$ based on hyperparameter tuning on the validation set.

4.4 Main Results and Analysis

Quantitative Results. We evaluate our model on HadCRUT and CMAP and report on the results in Table 2. On both datasets,

[1] available on Climate Research Unit website: https://crudata.uea.ac.uk/

[2] CMAP Precipitation data is provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov/

[3] KDD CUP Challenge 2018 dataset, available at: http://www.kdd.org/kdd2018/

Table 2: Experimental results of our ESC-GAN for different missing data patterns, along with compared methods, on HadCRUT and CMAP datasets. We mark the best results (in bold) and the second best.

Method	HadCRUT						CMAP					
	Scatter		Reg Cluster		Irr Cluster		Scatter		Reg Cluster		Irr Cluster	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Zero	1.0396	0.6638	0.8551	0.5983	0.7446	0.5879	1.0290	0.6830	1.2861	0.7620	1.2869	0.7390
Mean	0.9697	0.6375	0.7985	0.5744	0.6971	0.5677	1.0272	0.6804	1.3294	0.7548	1.2969	0.7364
sKNN	0.3756	0.3991	0.4645	0.4649	0.3791	0.4210	0.1120	0.1785	0.7159	0.4863	0.4888	0.3960
IDW	0.3524	0.3868	0.4440	0.4535	0.3596	0.4087	0.1042	0.1719	0.7036	0.4792	0.4658	0.3839
Kriging	0.9517	0.6308	0.7995	0.5767	0.6906	0.5703	0.8838	0.5863	0.9709	0.6279	1.1257	0.6545
MF	0.6181	0.5216	0.7669	0.5782	0.6111	0.5390	0.1721	0.2395	0.8583	0.5753	0.5942	0.4974
BTTF	0.5867	0.5225	0.6798	0.5553	0.5764	0.5332	0.2474	0.3137	0.9423	0.6237	0.5723	0.5154
ST-MVL	0.3648	0.3964	0.4710	0.4655	0.3581	0.4084	0.1162	0.1832	0.7202	0.4919	0.5039	0.4177
IGKNN	0.7214	0.5492	0.7212	0.5501	0.6405	0.5491	0.8474	0.6132	1.1710	0.7285	1.1254	0.7170
NAOMI	1.0391	0.6637	0.8550	0.5983	0.7442	0.5877	1.0288	0.6833	1.2859	0.7620	1.2863	0.7396
PConv	0.3908	0.4211	0.4759	0.4784	0.4122	0.4494	0.1008	0.1704	0.6469	0.4492	0.2969	0.3024
3DGated	0.3610	0.3907	0.4265	0.4454	0.3581	0.4066	0.1400	0.2071	0.5532	0.4381	0.2952	0.3033
ESC-GAN	**0.3354**	**0.3800**	**0.4097**	**0.4295**	**0.3418**	**0.3971**	**0.0802**	**0.1531**	**0.5441**	**0.4308**	**0.2739**	**0.3017**

ESC-GAN consistently outperforms all the baselines.

Classical zero filling and mean filling have high errors by both metrics, as they do not consider any neighborhood information. sKNN and IDW take into account local information based purely on distance weighting, and their performances degenerate under clustered missing patterns, where neighboring locations are simultaneously missing. Kriging and MF put strong assumptions on the input distribution, which may not be observed in real-world dataset.

For the state-of-the-art spatio-temporal imputation methods, BTTF assumes Gaussian spatial factor and does not sufficiently model spatial dependencies. ST-MVL is not fully data-driven and does not fully capture the underlying spatio-temporal correlations. IGKNN mainly leverages neighboring nodes for imputation without incorporating the global and multi-scale structure. NAOMI exploits the multi-resolution structure but heavily relies on temporal domain information. Although during training, locations that are randomly masked out in the temporal domain could well recover the missing values, it basically generates mean value of the dataset when extended to unobserved locations.

For image and video inpainting methods, Partial Convolution only uses local convolution operators to learn to recover missing information in two-dimensional space but does not leverage temporal information. 3D gated convolution combines temporal and spatial information in a three-dimensional space but does not model the global context information and the multi-scale structure of spatio-temporal data. By contrast, our ESC-GAN learns both local and global similarity and aggregates features at multiple scales in the three-dimensional spatio-temporal space. Therefore, ESC-GAN achieves the lowest errors compared to both classical and state-of-the-art imputation or inpainting methods.

Qualitative Results. In addition to quantitative improvements, we present example grid maps generated by ESC-GAN for qualitative comparison, as shown in Figure 6. We have zoomed in on the area with missing data for better view. The numbers above the figures are the MSE for the corresponding grid map. The first row of results in Figure 6 illustrates imputation for irregular clustered locations, and the second row is for regular clustered locations. Compared with all the baselines, our model generates values for the missing regions closest to the ground truth and shows smoother transition to the observed regions, qualitatively demonstrating the proposed model's effectiveness.

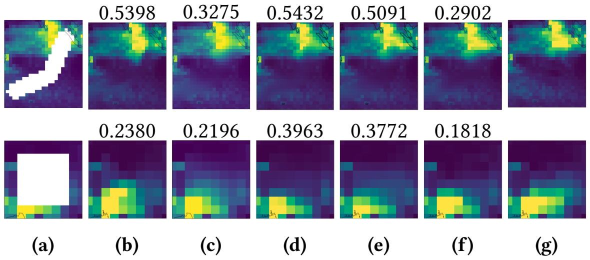
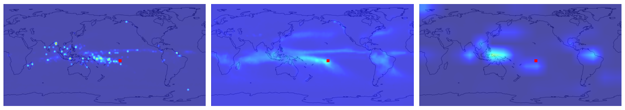


Figure 6: Case Study: Example grid maps for qualitative comparison. (a) Missing Area, (b) IDW, (c) ST-MVL, (d) PConv, (e) 3DGated, (f) ESC-GAN, and (g) GT. Numbers above the figures are the MSE for the corresponding generated grid maps compared with the GT grid map.

4.5 Ablation Studies



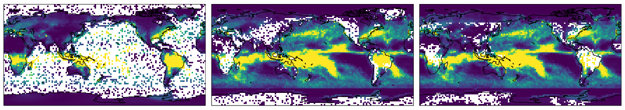
(a) Dot Product GT **(b) Cosine GT** **(c) ESC-GAN**

Figure 7: Ground-truth (GT) and generated attention maps.

Quantitative Analysis. We also conducted ablation study to separately examine the effect of our global attention module and multi-scale structure. We report MSE and MAE after removing global attention module (ESC-GAN-local), removing multi-scale structure (ESC-GAN-single), and removing both of these modules (ESC-GAN-vanilla). As shown in Table 3, the performance degenerates after removing either one or both of these components, which validates the

Table 3: Ablation Study of our global attention and multi-scale structure on CMAP, measured by MSE and MAE.

Method	Scatter		Reg Cluster		Irr Cluster	
	MSE	MAE	MSE	MAE	MSE	MAE
ESC-GAN-vanilla	0.0825	0.1557	0.5874	0.4433	0.2900	0.3099
ESC-GAN-local	0.0818	0.1545	0.5848	0.4362	0.2785	0.3032
ESC-GAN-single	0.0842	0.1588	0.5814	0.4388	0.2880	0.3107
ESC-GAN	**0.0802**	**0.1531**	**0.5441**	**0.4308**	**0.2739**	**0.3017**



(a) Ocean **(b) High-altitude** **(c) High-vegetation**

Figure 8: Patterned missing data distributions

Table 4: Evaluations for patterned missing distributions.

Method	Ocean		High-altitude		High-vegetation	
	MSE	MAE	MSE	MAE	MSE	MAE
sKNN	0.3162	0.3160	0.0865	0.1428	0.1857	0.2178
IDW	0.2738	0.2929	0.0786	0.1335	0.1687	0.2060
ST-MVL	0.2855	0.2972	0.0784	0.1347	0.1710	0.2061
PConv	0.2174	0.2517	0.0743	0.1303	0.1588	0.2004
3DGated	0.2989	0.3093	0.1231	0.1878	0.2254	0.2644
ESC-GAN	**0.1929**	**0.2512**	**0.0663**	**0.1234**	**0.1399**	**0.1911**

necessity of the proposed structure. The vanilla model ESC-GAN-vanilla is based only on local operators of randomly masked 3D partial convolution, and it does not model the global context similarity or leverage information at multiple scales. The local model ESC-GAN-local aggregates features from multiple granularity, but ignores the underlying global patterns exhibited in spatio-temporal data. The single model ESC-GAN-single considers the global trends in spatio-temporal data, but it learns such features at one single scale using one branch in the generator. Our proposed ESC-GAN jointly learns global and local dependencies, and aggregates multi-level features, thus producing more accurate estimations compared with different ESC-GAN ablations.

Attention Visualization. To provide more interpretable results, we randomly select query region and visualize the softmax attention score between query region and all the other regions on CMAP, as shown in Figure 7. We mark the query regions with red rectangles. Attention scores in three figures are calculated based on the average data from all timestamps. The left two figures are ground-truth attention scores measured by dot product and cosine similarity, both followed by softmax normalization. As CMAP dataset measures monthly precipitation, the query region exhibits patterns similar to regions near the equator and regions in the Pacific Ocean. Comparing ESC-GAN generated attention map with the ground truth, we could observe that ESC-GAN is able to accurately capture the global patterns through the attention mechanism.

4.6 Robustness Studies

Robustness to Non-Random Missing Shapes. Apart from the missing distributions in Table 2, extending sensor spatial coverage in real-world also encounters non-random realistic missing distributions, i.e. missing distribution follows a specific pattern as a result of land, elevation, vegetation, etc. For example, it is more difficult to deploy sensors on mountains than plains, so locations of higher elevation are expected to have a lower coverage of sensory data. Similar comparison also resides in ocean vs land, forests vs locations with lower vegetation cover rate. In light of this, we study the effectiveness of ESC-GAN with respect to three non-random missing data distributions (as shown in Figure 8), namely, missing data in the ocean, high-altitude area, and area with high vegetation cover. This also evaluates the model's transferability, as there exists a distribution gap between training regions and testing regions. We report on the results in Table 4. ESC-GAN achieves the best performance under different real-world non-random realistic missing distributions.

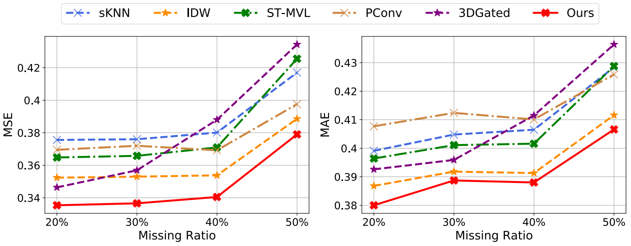


Figure 9: Evaluation for different missing ratio $|D_U|/|D|$ on HadCRUT, measured by MSE and MAE.

Robustness to Amount of Missing Data. We also study model robustness with respect to missing area size. Following previous notation, we use D to represent the whole map and D_U to represent the set of unobserved grid cells. For scattered missing distribution in Section 4.4, the missing ratio $|D_U|/|D|$ is 20%. We increase the missing ratio $|D_U|/|D|$ from 20% to 50%, and calculate the corresponding MSE and MAE of the best performing baselines in Figure 9. As shown in the figure, for different models, both MSE and MAE generally grow as the missing ratio increases. Moreover, under settings of all varying missing ratios, ESC-GAN is able to outperform all the baselines for both MSE and MAE, demonstrating the proposed model's robustness to varying missing area size.

4.7 Generalization to Traditional Spatio-Temporal Imputation

In addition to our proposed extension of the spatial coverage task, we also apply our model to the traditional spatio-temporal imputation task for random missing values, to evaluate its generalizability. For this, we conducted experiments on the KDD CUP 2018 dataset. We normalize the data using z-normalization and split the sequences into chunks of length 48, following previous studies [26, 28]. We compare the results with a list of methods for doing traditional spatio-temporal imputation, including both statistical imputation methods (filling with last available observation (Last) or mean value (Mean), k-Nearest Neighbors (KNN), Matrix Factorization (MF)) and deep learning-based models (MTSI [26], BRITS [6], DCRNN [23], CDSA [28]) following previous studies [26, 28].

In Table 5, our model outperforms all the other baselines at various missing data ratios from 20% to 90%. Compared with other methods, our model could jointly learn temporal and spatial dependencies at different scales. Without modification of the model

Table 5: Results of spatio-temporal imputation for random missing values on KDD 2018 Dataset, measured by MSE[4].

%Missing	20%	30%	40%	50%	60%	70%	80%	90%
Last	1.073	0.894	0.901	0.990	1.040	1.236	1.689	2.870
Mean	0.916	0.907	0.914	0.923	0.973	0.935	0.937	1.002
KNN	0.892	0.803	0.776	0.798	0.856	0.852	0.873	1.243
MF	0.850	0.785	0.787	0.772	0.834	0.805	0.860	1.196
MTSI	0.844	0.780	0.753	0.743	0.803	0.780	0.837	1.018
BRITS	0.455	0.421	0.372	0.409	0.440	0.482	0.648	0.725
DCRNN	0.579	0.565	0.449	0.506	0.589	0.622	0.720	0.861
CDSA	0.373	0.393	0.287	0.291	0.387	0.495	0.521	0.631
ESC-GAN	**0.207**	**0.229**	**0.232**	**0.231**	**0.274**	**0.299**	**0.326**	**0.434**

structure, ESC-GAN is directly applicable to spatio-temporal imputation tasks. This demonstrates the generalizability of ESC-GAN, indicating its potential in a broader range of applications.

5 RELATED WORK

Spatio-Temporal Data Imputation. For traditional spatio-temporal imputation, existing approaches mainly include statistical models and deep generative models. Statistical models include filling with zero, mean of existing values, filling with last observation, regression-based models [2], MICE [5], Matrix Factorization [33], k-nearest neighbours [17], tensor factorization [11], multi-view learning method [47]. Deep generative models have so far shown promising results with different sequential neural network-based models [6, 9, 25, 28] and generative adversarial network (GAN) [15]-based models [22, 26, 27, 40]. These methods typically assume partial missing in the temporal domain.

For spatial missing data imputation, existing approaches are mainly statistical, e.g. inverse distance weighting [10], matrix factorization [14, 29], variogram modeling [37]. These methods miss modeling the temporal trend from available locations.

Meanwhile, existing spatio-temporal interpolation methods do not sufficiently exploit properties of spatio-temporal data (e.g. global context and multi-scale structure) [3, 20, 21, 38, 45]. Miao et al. propose a pyramid dilated spatial-temporal network for learning crowd flow representations, but the model is designed for forecasting task and only learns temporal attention [30]. Tang et al. infer traffic volume of observed regions through joint modeling of dense and incomplete trajectories [39]. However, their method is uniquely focused on trajectory data and is not directly applicable to other domains. We propose to jointly model spatial and temporal dependencies, learning from both local and global patterns. Our model is applicable to a wide range of spatio-temporal problems.

Image and Video Inpainting. Image or video inpainting aims to reconstruct missing regions in an image or video frame. Unlike conventional convolution neural network that treats all the pixels equally as valid pixels, Partial Convolution- and Gated Convolution-based methods assign different weights to different input pixels to reduce color discrepancy and blurriness [8, 24, 49]. To synthesize different image components for image inpainting, Wang et al. propose a multi-column network to extract features at different levels [44]. However, these approaches, which are specifically designed for image and video inpainting, cannot achieve satisfactory results on our ESC task, as spatio-temporal data contain more complex correlations and stochastic properties compared with images. These methods also only exploit local features, while our model jointly models local and global patterns exhibited in spatio-temporal data.

Global Attention. Convolution neural network has been successful in image or video processing, but convolution by nature is a local operation. Attention mechanism [4] has enabled a model to gain a global view of the input data. In particular, self-attention [42] draws global dependencies between input and output based solely on the attention mechanism. Self-attention calculates response at one position based on weighted sum of values from all the other positions, and has shown impressive results in sequential task like machine translation and sequence generation. Non-local model is further introduced to bridge the gap of applying self-attention to image and video tasks [7, 43, 48, 50]. We follow the attention module in [48] to capture global context embedded in spatio-temporal data, and further combine it with multi-scale structure.

6 CONCLUSIONS AND FUTURE WORK

We address the challenge of extending the spatial coverage (ESC) of sensory data to locations without any historical values. Traditional spatio-temporal imputation methods do not work well as they rely on partial data availability for a location. This ESC task has far-reaching applications in geographical discovery, physical modeling, weather forecasting, urban planning, etc. It faces challenges related to collaborative use of spatial and temporal domains, local and global context and structure across multiple scales. To address these challenges, we devised a model to recover data for "new" locations leveraging both spatial and temporal dependencies. In view of the non-linear, multi-resolution and stochastic nature of spatio-temporal data, our method generates data considering both the global and local perspectives. We optimize the model with multi-scale and adversarial training to better capture the underlying patterns. We evaluated ESC-GAN on real-world geo-sensory datasets where results demonstrate that our model outperforms all the baselines under different missing scenarios.

There are limitations that we plan to address in future studies. As geometric distance on a sphere is not strictly preserved after being mapped to a 2D gridded map, we plan to incorporate spherical convolutions to better model spatial dependencies. We will also explore approaches to further extend the model to irregular super-resolution task for generating data at finer spatial granularity.

ACKNOWLEDGMENTS

This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This work was also sponsored in part by National Science Foundation Convergence Accelerator under award OIA-2040727 as well as generous gifts from Google, Adobe, and Teradata. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes not withstanding any copyright annotation hereon.

[4] We directly adopt the numbers for the compared methods from prior work [26, 28].

REFERENCES

[1] Lorenzo Alfieri, Peter Burek, Emanuel Dutra, Blazej Krzeminski, David Muraro, Jutta Thielen, and Florian Pappenberger. 2013. GloFAS–global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences* 17, 3 (2013), 1161–1175.

[2] Craig F Ansley and Robert Kohn. 1984. On the estimation of ARIMA models with missing values. In *Time series analysis of irregularly observed data*. Springer.

[3] Gabriel Appleby, Linfeng Liu, and Li-Ping Liu. 2020. Kriging Convolutional Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3187–3194.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] S van Buuren and Karin Groothuis-Oudshoorn. 2010. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* (2010), 1–68.

[6] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems* 31 (2018), 6775–6785.

[7] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[8] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. 2019. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision*. 9066–9075.

[9] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 1–12.

[10] Feng-Wen Chen and Chen-Wuing Liu. 2012. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment* 10, 3 (2012), 209–222.

[11] Xinyu Chen and Lijun Sun. 2021. Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[12] Kevin Cowtan and Robert G Way. 2014. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society* 140, 683 (2014), 1935–1944.

[13] JW Gjerloev. 2009. A global ground-based magnetometer initiative. *Eos, Transactions American Geophysical Union* 90, 27 (2009), 230–231.

[14] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, Bei Chen, and Xiangjun Dong. 2020. A Spatial Missing Value Imputation Method for Multi-view Urban Statistical Data.. In *IJCAI*. 1310–1316.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014), 2672–2680.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Andrew T Hudak, Nicholas L Crookston, Jeffrey S Evans, David E Hall, and Michael J Falkowski. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112, 5 (2008), 2232–2245.

[18] Christopher Kadow, David Matthew Hall, and Uwe Ulbrich. 2020. Artificial intelligence reconstructs missing climate information. *Nature Geoscience* (2020).

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Lixin Li and Peter Revesz. 2004. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems* 28, 3 (2004), 201–227.

[21] Lixin Li, Xingyou Zhang, James B Holt, Jie Tian, and Reinhard Piltner. 2011. Spatiotemporal interpolation methods for air pollution exposure. In *Ninth Symposium of Abstraction, Reformulation, and Approximation*.

[22] Steven Cheng-Xian Li and Benjamin Marlin. 2020. Learning from irregularly-sampled time series: A missing data perspective. In *International Conference on Machine Learning*. PMLR, 5937–5946.

[23] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[24] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 85–100.

[25] Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. 2019. Naomi: Non-autoregressive multiresolution sequence imputation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

[26] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. 2018. Multivariate time series imputation with generative adversarial networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

[27] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. 2019. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *AAAI Press*. 3094–3100.

[28] Jiawei Ma, Zheng Shou, Alireza Zareian, Hassan Mansour, Anthony Vetro, and Shih-Fu Chang. 2019. Cross-Dimensional Self-Attention for Multivariate, Geo-tagged Time Series Imputation. (2019).

[29] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* 11 (2010), 2287–2322.

[30] Congcong Miao, Jiajun Fu, Jilong Wang, Heng Yu, Botao Yao, Anqi Zhong, Jie Chen, and Zekun He. 2021. Predicting Crowd Flows via Pyramid Dilated Deeper Spatial-temporal Network. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 806–814.

[31] Colin P Morice, John J Kennedy, Nick A Rayner, and Phil D Jones. 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres* 117, D8 (2012).

[32] Colin P Morice, John J Kennedy, Nick A Rayner, JP Winn, Emma Hogan, RE Killick, RJH Dunn, TJ Osborn, PD Jones, and IR Simpson. 2020. An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset. *Journal of Geophysical Research: Atmospheres* (2020), e2019JD032361.

[33] Morten Morup, Daniel M Dunlavy, Evrim Acar, and Tamara Gibson Kolda. 2010. *Scalable tensor factorizations with missing data*. Technical Report. Sandia National Laboratories.

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

[35] Alex Rubinsteyn and Sergey Feldman. [n.d.]. *fancyimpute: An Imputation Library for Python*. https://github.com/iskandr/fancyimpute

[36] Divya Saxena and Jiannong Cao. 2019. D-GAN: Deep Generative Adversarial Nets for Spatio-Temporal Prediction. *arXiv preprint arXiv:1907.08556* (2019).

[37] Michael L Stein. 2012. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

[38] Koh Takeuchi, Hisashi Kashima, and Naonori Ueda. 2017. Autoregressive tensor factorization for spatio-temporal predictions. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1105–1110.

[39] Xianfeng Tang, Boqing Gong, Yanwei Yu, Huaxiu Yao, Yandong Li, Haiyong Xie, and Xiaoyu Wang. 2019. Joint modeling of dense and incomplete trajectories for citywide traffic volume inference. In *The World Wide Web Conference*. 1806–1817.

[40] Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. 2020. Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[41] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[44] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Image inpainting via generative multi-column convolutional neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

[45] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. 2020. Inductive graph neural networks for spatiotemporal kriging. *arXiv preprint arXiv:2006.07527* (2020).

[46] Pingping Xie and Phillip A Arkin. 1997. Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society* 78, 11 (1997), 2539–2558.

[47] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: filling missing values in geo-sensory time series data. (2016).

[48] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. 2020. Disentangled non-local neural networks. In *European Conference on Computer Vision*. Springer, 191–207.

[49] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *ICCV*.

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning*. 7354–7363.

[51] Lin Zhang, Alexander Gorovits, Wenyu Zhang, and Petko Bogdanov. 2020. Learning Periods from Incomplete Multivariate Time Series. In *ICDM*.

[52] Yingxue Zhang, Yanhua Li, Xun Zhou, Xiangnan Kong, and Jun Luo. 2020. Curb-GAN: Conditional Urban Traffic Estimation through Spatio-Temporal Generative Adversarial Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 842–852.

A COMPARED METHODS

(1) Classical Imputation. For classical imputation, we apply the methods for each timestamp separately.

- **Mean/Zero**: Mean/Zero filling extends the missing station values by filling mean value of the current timestamp/zero (which is the mean of all timestamps after normalization).
- **sKNN**: Spatial K Nearest Neighbour, which extends the readings with the average values of the missing station’s k nearest available spatial neighbors [17].
- **IDW**: Inverse Distance Weighting, a global spatial learning method that interpolates with weighted average of available data points as a function of inverse distance [10].
- **Kriging**: A geo-statistical interpolation method which assumes that the distance between points reflects spatial correlation that could be used to explain surface variation [37]. We implement Kriging method using PyKrige library[5].
- **MF**: Matrix Factorization, which iteratively replaces missing elements with those obtained from soft-thresholded SVD [29]. We implement MF method using fancyimpute library [35].

(2) State-of-the-art Spatio-temporal Imputation. We implement the state-of-the-art spatio-temporal imputation methods based on public code of the respective paper.

- **BTTF**: Bayesian temporal factorization framework for modeling spatio-temporal data with missing values. The method integrates low-rank matrix/tensor factorization and vector autoregressive (VAR) process into a single probabilistic graphical model [11].
- **ST-MVL**: A multi-view learning imputation method combining empirical statistical models for global view, with data-driven algorithms for local view [47]. In the original implementation, weights for combining four views are optimized for each location. However, in ESC task we do not have any available data to train on these unobserved locations. Therefore, we use optimized weights from their neighbors to predict the missing values.
- **NAOMI**: A non-autoregressive deep generative spatio-temporal imputation method. It also exploits the multi-resolution structure by decoding recursively from coarse to fine-grained resolutions [25]. The original NAOMI implementation mainly relies on temporal information for imputation. As in ESC task we have no temporal information for unobserved locations, we concatenate observed locations with unobserved locations channelwise in order to better generalize from observed locations.
- **IGKNN**: Deep spatio-temporal kriging method based on inductive graph neural network. The method learns spatial message passing mechanism through generating random subgraph and reconstructing subgraph signals [45].

(3) State-of-the-art Image and Video Inpainting. We implement the state-of-the-art image and video inpainting methods based on public code of the respective paper.

- **PConv**: Partial Convolution [18, 24] for imputing 2D data. The convolution is conditioned only on valid cells.
- **3DGated**: 3D video inpainting method, which uses 3D gated convolution as generator and proposes a Temporal Patch-GAN loss to enhance temporal consistency [8].

[5] https://github.com/GeoStat-Framework/PyKrige