Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ST-GAIN+: A Spatiotemporal Generative Adversarial Network for Missing Solar Irradiance Data Imputation

Hamood Alfahdi

hamood_24003614@utp.edu.om

Universiti Teknologi PETRONAS

Hitham Alhussain

Universiti Teknologi PETRONAS

Hilmi Hasan

Universiti Teknologi PETRONAS

Ahmed Omer Ismail

Applied College, King Khalid University

Sallam O. F. Khairy

University of Nizwa

Shamsuddeen Adamu

Universiti Teknologi PETRONAS

Ismail Said Almuniri

Universiti Teknologi PETRONAS

ALWaleed Sulaiman Alabri

Universiti Teknologi PETRONAS

Article

Keywords: generative adversarial networks, missing data imputation, solar irradiance, spatiotemporal modeling, uncertainty estimation

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10487457/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ST-GAIN+: A Spatiotemporal Generative Adversarial Network for Missing Solar Irradiance Data Imputation

Hamood Alfahdi[1], Hitham Alhussain[2], Hilmi Hasan[3], Ahmed Omer Ismail[4], Sallam O. F. Khairy[5], Shamsuddeen Adamu[6], Ismail Said Almuniri[7], ALWaleed Sulaiman Alabri[8]

[1]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia, Department of Information Systems, College of Economics, Management and Information Systems, University of Nizwa, Nizwa, Oman

[2]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

[3]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

[4]Unit of Specializations of Technology and Engineering, Applied College, King Khalid University, 61913 Muhayil, Saudi Arabia

[5]Department of Information Systems, College of Economics, Management and Information Systems, University of Nizwa, Nizwa, Oman

[6]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia
Ahmadu Bello University, Zaria, Nigeria

[7]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia, Department of Information Systems, College of Economics, Management and Information Systems, University of Nizwa, Nizwa, Oman

[8]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia, Department of Information technology, University of Technology and Applied Sciences, Oman

Corresponding authors:

Hamood Alfahdi, Email: hamood_24003614@utp.edu.om
Hitham Alhussain, Email: seddig.alhussian@utp.edu.my

ABSTRACT Missing data in solar irradiance datasets significantly degrades the performance of solar forecasting models and smart grid optimization systems. Traditional imputation methods often struggle to capture complex spatiotemporal dependencies and typically lack mechanisms for uncertainty estimation. To address these challenges, this paper proposes ST-GAIN+, a hybrid spatiotemporal generative adversarial network for missing solar irradiance data imputation. The proposed framework integrates Temporal Convolutional Networks (TCNs) to capture long-range temporal dependencies, dual attention mechanisms to model spatial and temporal correlations, and Bayesian uncertainty estimation based on Monte Carlo dropout to quantify predictive uncertainty. In addition, a GAIN-based adversarial learning framework with a hint mechanism is employed to improve realistic reconstruction of missing observations. The proposed model is evaluated under Missing Completely at Random (MCAR) conditions with missing rates of 30%, 50%, 70%, and 90% using multiple real-world solar irradiance datasets, including NASA POWER, NSRDB, and Nama_Oman. Experimental results demonstrate that ST-GAIN+ achieves stable and accurate reconstruction performance across multiple solar irradiance datasets under MCAR missing rates ranging from 30% to 90%. The framework consistently maintains low RMSE and MAE values even under severe data incompleteness. Comparative and ablation analyses indicate that temporal modeling, attention-based dependency learning, and uncertainty-aware estimation contribute to robust solar irradiance reconstruction. The findings suggest that the integration of adversarial learning, spatiotemporal modeling, and uncertainty-aware estimation provides a promising framework for reliable solar irradiance data imputation in renewable energy applications.

INDEX TERMS generative adversarial networks, missing data imputation, solar irradiance, spatiotemporal modeling, uncertainty estimation.

I. INTRODUCTION

Missing data is a pervasive challenge in modern data-driven systems and frequently arises due to sensor failures, communication interruptions, environmental interference, maintenance activities, and data transmission errors. In renewable energy monitoring systems, particularly solar irradiance measurement networks, missing observations are common because of harsh environmental conditions, cloud-induced variability, hardware degradation, and intermittent outages across geographically distributed stations [4]–[7], [27], [28], [30], [32]. The presence of missing values can significantly degrade the quality of downstream applications, including photovoltaic (PV) power forecasting, smart-grid operation, energy management, and renewable energy system optimization [5], [6], [10], [30], [32]. Consequently, accurate reconstruction of missing solar irradiance observations has become an important research problem for ensuring reliable decision-making in renewable energy systems.

Traditional imputation methods, including statistical interpolation, regression-based techniques, k-nearest neighbors, Multiple Imputation by Chained Equations (MICE), and MissForest, have been widely applied for missing-data reconstruction [14]–[17]. While these approaches are computationally efficient and effective for relatively simple datasets, they often struggle to capture nonlinear temporal dynamics, complex spatial correlations, and long-range dependencies that characterize solar irradiance time series. Moreover, their performance typically deteriorates under high missing rates, limiting their applicability to real-world renewable energy systems.

Recent advances in deep learning have enabled the development of more sophisticated imputation frameworks capable of learning complex temporal and spatial patterns directly from data. Among these approaches, Generative Adversarial Imputation Networks (GAIN) introduced adversarial learning for missing-data reconstruction by training a generator and discriminator in a competitive framework [1]. Building upon GAIN, several extensions have been proposed, including ImputeGAN [2], SolarGAN [4], ST-GAIN [3], WGAIN [9], Conv-GAIN [11], ImputeRNN [8], SAITS [19], BRITS [13], CC-GAIN [23], and IGANI [24], which incorporate temporal modeling, attention mechanisms, and adversarial learning to improve imputation performance across various application domains. These methods have demonstrated substantial improvements over conventional statistical techniques; however, several important limitations remain.

First, many existing approaches focus primarily on temporal dependency modeling while providing limited exploitation of spatial relationships among geographically distributed monitoring stations. Second, several models rely heavily on recurrent architectures, which may encounter difficulties in capturing long-range temporal dependencies efficiently [33], [34]. Third, most existing methods produce deterministic imputations and do not explicitly quantify uncertainty associated with reconstructed values. This limitation is particularly critical in risk-sensitive applications such as solar forecasting, renewable energy management, and smart-grid operation, where uncertainty awareness is essential for informed decision-making [37]–[40]. Finally, the combined effects of temporal modeling, spatial dependency learning, and uncertainty estimation remain insufficiently explored within a unified adversarial framework specifically designed for solar irradiance reconstruction.

To address these limitations, this paper proposes ST-GAIN+, a Spatiotemporal Generative Adversarial Network with uncertainty estimation for missing solar irradiance data imputation. The proposed framework extends the original GAIN architecture by integrating four complementary components: (1) adversarial learning for realistic reconstruction of missing values [1], [18]; (2) Temporal Convolutional Networks (TCNs) for hierarchical temporal feature extraction and long-range dependency modeling [33], [34]; (3) dual attention mechanisms for enhanced representation of informative temporal and spatial relationships [35], [36]; and (4) Bayesian uncertainty estimation through Monte Carlo dropout to quantify prediction confidence and improve reconstruction robustness under severe missing-data conditions [37], [38].

Unlike existing approaches that address temporal modeling, spatial dependency learning, or uncertainty estimation independently, ST-GAIN+ combines these capabilities within a unified adversarial learning framework specifically tailored for solar irradiance reconstruction. The interaction between adversarial learning, temporal feature extraction, dual attention mechanisms, and uncertainty-aware estimation enables the proposed framework to preserve temporal continuity, exploit spatial correlations, and improve reconstruction reliability simultaneously. Furthermore, the framework is designed to remain robust under highly incomplete data conditions commonly encountered in operational renewable energy systems.

The effectiveness of the proposed framework is evaluated using three real-world solar irradiance datasets representing both single-site and multisite scenarios: Nama_Oman_2016, NSRDB_2019, and NASA POWER. Experimental evaluation is conducted under Missing Completely at Random (MCAR) conditions with missing rates ranging from 30% to 90%. Results demonstrate that ST-GAIN+ consistently maintains low normalized reconstruction errors across all datasets, achieving normalized RMSE values below 0.023 and MAE values below 0.012 even under severe missing-data conditions. Furthermore, ablation analysis confirms the effectiveness of the proposed components and highlights the importance of uncertainty-aware learning in improving reconstruction robustness.

The main contributions of this work are summarized as follows:

1. A novel spatiotemporal generative adversarial imputation framework, ST-GAIN+, is proposed for reconstructing missing solar irradiance observations under MCAR conditions.
2. Temporal Convolutional Networks and dual attention mechanisms are integrated within a unified adversarial architecture to enhance temporal feature extraction and spatial dependency modeling.
3. A Bayesian uncertainty estimation mechanism based on Monte Carlo dropout is incorporated to provide confidence-aware imputation and improve reconstruction reliability.
4. Comprehensive experiments are conducted on multiple real-world solar irradiance datasets under varying missing rates, demonstrating stable performance and strong robustness under highly incomplete conditions.
5. Extensive ablation and comparative analyses are performed to evaluate the contributions of temporal modeling, attention mechanisms, and uncertainty-aware learning within the proposed framework.

The remainder of this paper is organized as follows. Section II reviews the related literature and background concepts. Section III presents the overall methodology, while Section IV introduces the mathematical formulation of the proposed approach. Section V describes the architecture of the proposed ST-GAIN+ framework, and Section VI details the training procedure. Section VII presents the experimental setup, performance evaluation, uncertainty calibration analysis, and ablation studies. Section VIII discusses the limitations of the current work and potential future research directions. Finally, Section IX concludes the paper.

II. BACKGROUND

Missing data is a common challenge in environmental monitoring, renewable energy systems, healthcare, transportation, and industrial applications. Depending on the relationship between the missing values and the underlying data generation process, missingness is generally categorized into three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [16], [17]. MCAR occurs when the probability of missingness is independent of both observed and unobserved variables, MAR occurs when missingness depends only on observed variables, and MNAR arises when missingness is related to unobserved variables or the missing values themselves. Among these mechanisms, MCAR is widely adopted in imputation research because it provides a controlled benchmark for evaluating reconstruction performance and enables fair comparison among competing algorithms [1], [3].

In solar irradiance monitoring systems, missing observations may result from sensor malfunctions, communication interruptions, equipment maintenance, environmental disturbances, and data acquisition failures [5], [6], [28], [32]. Such missing values can significantly affect photovoltaic power forecasting, energy management, and smart-grid operation. Consequently, accurate and reliable imputation techniques are essential for maintaining data quality and ensuring robust downstream analysis.

Traditional imputation approaches include statistical interpolation, regression-based methods, k-nearest neighbors, Multiple Imputation by Chained Equations (MICE), and MissForest [14]–[17]. These methods estimate missing values by exploiting local correlations and statistical relationships within the data. Although computationally efficient, their performance often deteriorates when faced with nonlinear dynamics, long-range temporal dependencies, high-dimensional feature spaces, and severe missing-data conditions commonly encountered in solar irradiance datasets.

Recent advances in machine learning and deep learning have led to significant improvements in missing-data reconstruction. Among these developments, Generative Adversarial Imputation Networkss (GAIN) introduced adversarial learning for missing-value estimation by training a generator and discriminator in a competitive framework supported by a hint mechanism [1]. Building upon GAIN, several extensions have been proposed, including ImputeGAN [2], SolarGAN [4], ST-GAIN [3], WGAIN [9], Conv-GAIN [11], ImputeRNN [8], IGANI [24], CC-GAIN [23], BRITS [13], and SAITS [19]. These methods incorporate adversarial learning, temporal modeling, attention mechanisms, and probabilistic representations to improve reconstruction accuracy across a variety of application domains.

Despite their success, several limitations remain. Many existing methods emphasize temporal modeling while providing limited exploitation of spatial relationships among geographically distributed sensors. Furthermore, recurrent architecture often encounters difficulties when modeling

long-term temporal dependencies and may suffer from increased computational complexity [33], [34]. Although attention-based approaches improve feature representation, most existing frameworks still generate deterministic imputations and do not explicitly quantify uncertainty associated with reconstructed values. This limitation is particularly important in renewable energy applications, where uncertainty information is critical for risk-aware decision making and operational planning [37]–[40].

To address these challenges, recent studies have explored the integration of adversarial learning, temporal feature extraction, and uncertainty-aware modeling for environmental and energy-related applications [4], [5], [10], [30], [32]. However, the combined effects of temporal convolutional modeling, spatial-temporal attention mechanisms, and Bayesian uncertainty estimation remain insufficiently investigated within a unified framework specifically designed for solar irradiance reconstruction.

Motivated by these observations, the proposed ST-GAIN+ framework extends existing adversarial imputation approaches by integrating four complementary components: adversarial learning for realistic reconstruction [1], [18], Temporal Convolutional Networks (TCNs) for hierarchical temporal feature extraction [33], [34], dual attention mechanisms for enhanced spatial-temporal representation learning [35], [36], and Bayesian uncertainty estimation through Monte Carlo dropout for confidence-aware reconstruction [37], [38]. Unlike conventional deterministic approaches, the proposed framework simultaneously models temporal continuity, spatial relationships, and predictive uncertainty, enabling robust imputation under severe missing-data conditions.

The effectiveness of ST-GAIN+ is evaluated using three real-world solar irradiance datasets representing both single-site and multisite environments. Experimental results demonstrate stable reconstruction performance across MCAR missing rates ranging from 30% to 90%, with consistently low normalized RMSE and MAE values. Furthermore, ablation analysis highlights the importance of uncertainty-aware learning, temporal feature extraction, and attention-based representation learning in improving reconstruction robustness and reliability.

III. METHODOLOGY

A. Problem Definition

Solar irradiance datasets collected from distributed monitoring stations frequently contain missing values due to sensor malfunctions, communication interruptions, calibration drift, and adverse environmental conditions. These missing entries degrade the quality of downstream tasks such as photovoltaic power forecasting, grid stability assessment, and energy management. Let the multivariate spatiotemporal dataset be denoted by

$$X \in \mathbb{R}^{T \times S \times F} \qquad (1)$$

where $T$is the number of time steps, $S$is the number of monitoring stations, and $F$is the number of measured features. A corresponding binary mask is defined as

$$M \in \{0,1\}^{T \times S \times F} \qquad (2)$$

where $M_{i,j,k} = 1$ indicates that $X_{i,j,k}$ is observed and $M_{i,j,k} = 0$ indicates that it is missing. The objective is to learn an imputation function

$$\hat{X} = f(X_{obs}, M) \qquad (3)$$

where $\hat{X}$denotes the reconstructed dataset, $f(\cdot)$represents the learned ST-GAIN+ imputation function, X_{obs}denotes the incomplete input containing the available observations, and $M$is the binary mask indicating the locations of observed and missing entries. The objective is to reconstruct the missing values accurately while preserving the temporal continuity, spatial consistency, and statistical characteristics of the original dataset.*B. Missing Data Mechanisms*

To ensure practical applicability, missing data mechanisms are commonly categorized into three standard types: missing completely at random (MCAR), where the probability of missingness is independent of both observed and unobserved values; missing at random (MAR), where missingness depends only on observed variables; and missing not at random (MNAR), where missingness depends on the missing values themselves or other unobserved factors [16], [17]. In renewable energy monitoring systems, solar irradiance datasets may exhibit different forms of missingness due to sensor degradation, communication failures, and environmental variability [5], [6], [30].

However, the present study focuses exclusively on the MCAR mechanism for experimental evaluation. Missing values are artificially generated under MCAR conditions, where missing observations are randomly distributed throughout the dataset without dependency on the underlying data distribution. The MCAR setting is widely adopted in imputation research because it provides a controlled and standardized benchmark for fair comparison between competing methods [1], [3], [17]. MAR and MNAR mechanisms are not considered in this study and are reserved for future investigation.

C. Limitations of Existing Methods

Existing imputation approaches remain insufficient for multi-station solar irradiance data for four main reasons. First, many methods treat observations as independent samples and therefore ignore temporal continuity and spatial interaction among stations. Second, shallow statistical or

machine learning models are limited in their ability to capture nonlinear dependencies and high-dimensional data distributions. Third, most approaches generate deterministic imputations without estimating predictive uncertainty, which reduces trustworthiness in risk-sensitive energy applications. Fourth, their performance deteriorates significantly as the missing rate increases, especially under block missingness or complex missingness mechanisms. These limitations motivate the development of a more expressive and uncertainty-aware spatiotemporal imputation framework.

D. Motivation for ST-GAIN+

To address the above challenges, this work proposes ST-GAIN+, a hybrid spatiotemporal generative imputation framework that extends adversarial imputation with enhanced temporal modeling, attention-based dependency learning, and Bayesian uncertainty estimation. The model integrates:
1) Temporal Convolutional Networks (TCNs) to capture long-range and multi-scale temporal dependencies [34], [35];
2) dual attention mechanisms to model both temporal relevance and inter-station spatial correlation [36]-[38];
3) a GAIN-based adversarial formulation to learn realistic imputations from incomplete data [1], [18]; and
4) Monte Carlo dropout to quantify predictive uncertainty [39], [40]. This combination enables robust and uncertainty-aware imputation under high missing rates and realistic missingness patterns.

IV. MATHEMATICAL FORMULATION

A. Input Representation and Masking

Given the complete spatiotemporal tensor $X$and its binary mask M, the corrupted input used for training is defined as

$$\tilde{X} = M \odot X + (1 - M) \odot Z \quad (4)$$

where Z is a random noise tensor having the same dimensionality as X, and $\odot$ denotes element-wise multiplication. This formulation preserves observed values and replaced missing entries with random noise, allowing the generator to learn how to reconstruct missing information rather than simply copying inputs.

B. Generator Function

The generator $G(\cdot)$is responsible for reconstructing the missing values while learning the underlying spatiotemporal structure of the incomplete data. Given the corrupted input $\tilde{X}$and the corresponding binary mask M, the generator first produces an estimated tensor, which is then combined with the observed values to construct the completed dataset:

$$\bar{X} = G(\tilde{X}, M)$$
$$\hat{X} = M \odot X + (1 - M) \odot \bar{X} \quad (5)$$

where $\bar{X}$denotes the raw generator output, and $\hat{X}$represents the completed dataset obtained by preserving the observed values while replacing only the missing entries with the corresponding generator estimates.
Unlike the conventional GAIN architecture, the proposed generator is enhanced with Temporal Convolutional Network (TCN) blocks and a dual attention mechanism to capture long-range temporal dependencies and dynamic spatial relationships among monitoring stations. Consequently, the generator produces imputations that are not only numerically accurate but also temporally coherent and spatially consistent. This design is motivated by previous studies demonstrating that convolutional and attention-based architectures provide superior long-range dependency modeling compared with conventional recurrent networks [34]–[38].

C. Discriminator Function

The discriminator $D(\cdot)$operates at the element level and estimates the probability that each entry in the imputed tensor is observed rather than generated. To guide adversarial learning, a hint matrix H is introduced as

$$H = B \odot M + 0.5(1 - B) \quad (6)$$

where $B \sim \text{Bernoulli}(p)$is a binary random matrix controlling the amount of mask information revealed to the discriminator. The discriminator output is defined as

$$\hat{M} = D(\hat{X}, H) \quad (7)$$

where $\hat{M}$is the discriminator's element-wise estimate of whether each entry is observed.

This hint-based formulation follows the adversarial identifiability principle of GAIN. It allows the model to operate effectively under missing data. The sequence corrupted input, hint fusion, generator output, and discriminator evaluation follows the methodological pattern adopted in this paper.

D. Loss Functions

The overall training objective consists of adversarial, reconstruction, and uncertainty-aware loss terms.

1) Adversarial Loss

The adversarial learning objective enables the generator and discriminator to compete in a minimax optimization process.

The discriminator is trained to distinguish between observed and imputed entries by minimizing the binary cross-entropy loss

$$L_D = -E\left[M \odot \log \hat{M} + (1 - M) \odot \log\left(1 - \hat{M}\right)\right] \quad (8a)$$

where $\hat{M}$denotes the discriminator's estimated probability that each entry is observed.
The generator is optimized to produce imputations that are indistinguishable from the observed data through the adversarial objective

$$L_{Gadv} = -E\left[(1 - M) \odot \log \hat{M}\right] \quad (8b)$$

Equation (8a) encourages the discriminator to correctly classify observed and imputed entries, whereas (8b) encourages the generator to reconstruct missing values that the discriminator is likely to classify as observed. Together, these adversarial objectives drive the generator to produce realistic and statistically consistent imputations while improving the discriminator's ability to distinguish genuine observations from generated values.

2) Reconstruction Loss

To preserve the observed measurements and stabilize adversarial training, a reconstruction loss is imposed on the known entries. This loss penalizes deviations between the generator predictions and the corresponding observed values:

$$L_{\text{rec}} = \| M \odot \hat{X} - M \odot \hat{X} \|_2^2 \quad (9)$$

where $\bar{X}$denotes the raw output of the generator. Since the loss is evaluated only over the observed entries indicated by M, it encourages the generator to preserve the known measurements while learning a consistent representation for reconstructing the missing values. Combined with the adversarial objective, this reconstruction loss improves training stability and enhances the accuracy of the imputed data.

3) Generator Optimization

The generator is trained by jointly optimizing the adversarial and reconstruction objectives. The overall generator loss is defined as

$$L_{\text{G}} = L_G^{adv} + \lambda L_{\text{rec}}, \quad (10)$$

where $L_G^{\text{adv}}$is the adversarial loss defined in (8b), $L_{\text{rec}}$is the reconstruction loss given in (9), and $\lambda$is a weighting coefficient that controls the contribution of the reconstruction term relative to the adversarial objective. This combined objective encourages the generator to produce realistic imputations while preserving the observed measurements and maintaining consistency with the underlying spatiotemporal structure of the data. The discriminator is optimized independently using the loss function defined in (8a).

V. PROPOSED ST-GAIN+ ARCHITECTURE

A. Architecture Overview

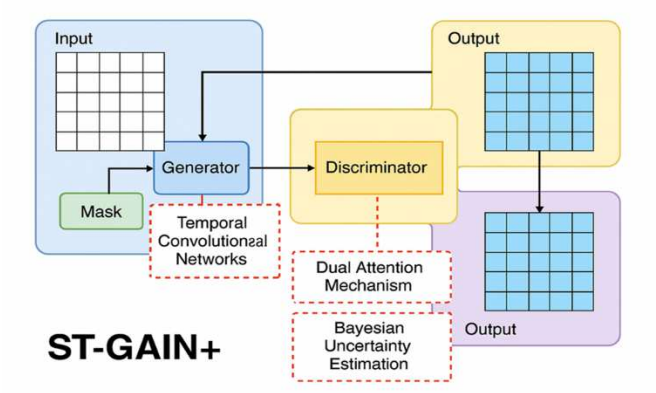


Fig. I. ST-GAIN+ framework.

The proposed ST-GAIN+ architecture is designed for missing data imputation in multivariate spatiotemporal solar irradiance datasets. The framework jointly integrates adversarial learning, temporal feature extraction, attention-based dependency modeling, and uncertainty estimation. The input to the model consists of the incomplete data tensor and the binary mask . The generator first produces an initial imputation, which is then evaluated by the discriminator using the hint matrix. To improve representation learning, the generator is equipped with TCN layers and dual attention modules. Finally, Monte Carlo dropout is used to quantify uncertainty in the imputed values. This section presents the architecture in a structured manner to clearly illustrate the interaction between the adversarial, temporal, attention, and uncertainty-aware components of the proposed framework.

B. Input and Masking Strategy

The corrupted input tensor is constructed using the masking strategy defined in (4), where the observed entries are preserved while the missing entries are replaced with randomly sampled noise. The binary mask Midentifies observed and missing values and is concatenated with the corrupted input before being fed into the generator. This masking strategy enables the generator to distinguish between available observations and values requiring

reconstruction while preserving the integrity of the observed measurements throughout the adversarial learning process.

C. Generator With Temporal Convolutional Networks

The generator is responsible for reconstructing missing values from incomplete observations. To effectively process long sequential solar irradiance data, the proposed model uses temporal convolutional networks rather than recurrent layers as the primary temporal backbone. TCNs employ causal convolutions to preserve temporal order and dilated convolutions to expand the receptive field, allowing the model to capture long-range dependencies across multiple time scales. Consequently, the generator can model daily periodicity, short-term fluctuations, and long-term seasonal variation more effectively than conventional recurrent structures.

D. Discriminator

The discriminator operates on the completed tensor $\hat{X}$and hint matrix H, producing an element-wise probability map indicating whether each value is observed or imputed. This data-point-level discrimination is particularly suitable for missing-value imputation because the model must decide the authenticity of each individual entry rather than classifying an entire sample. The adversarial interaction between the generator and discriminator pushes the imputed values toward the true data distribution.

E. Dual Attention Mechanism

To further enhance representation learning, ST-GAIN+ incorporates a dual attention mechanism composed of temporal attention and spatial attention. Temporal attention identifies the most relevant time steps for reconstructing a missing entry, allowing the model to emphasize informative historical patterns. Spatial attention captures dependencies among multiple monitoring stations, enabling the model to exploit inter-station similarities and spatial correlation. The attention-enhanced representation can be expressed as

Here is the formatted equation block:

$$
\begin{aligned}
&Q_t = XW^{Q_t}, \quad K_t = XW^{K_t}, \quad V_t = XW^{V_t} \\
&H_t = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_t}}\right) V_t \\
&Q_s = H_t W^{Q_s}, \quad K_s = H_t W^{K_s}, \quad V_s = H_t W^{V_s} \\
&X_{\text{att}} = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_s}}\right) V_s
\end{aligned} \tag{11}
$$

where W_{Q_t}, W_{K_t}, and $W_{V_t}$are the temporal projection matrices, while W_{Q_s}, W_{K_s}, and $W_{V_s}$are the spatial projection matrices. The variables $d_t$and $d_s$denote the temporal and spatial key dimensions, respectively. The intermediate representation $H_t$captures temporal dependencies, whereas X_{att}represents the final spatiotemporally attended features.

This design improves consistency across both time and space and is particularly beneficial in geographically distributed solar irradiance networks.

F. Bayesian Uncertainty Estimation

To quantify predictive confidence, ST-GAIN+ employs Monte Carlo dropout as a Bayesian approximation [39], [40]. During inference, dropout remains active and the generator performs stochastic forward passes, producing the set of predictions . The predictive mean and variance are then computed as

$$
\mu = \frac{1}{K}\sum_{k=1}^{K} \hat{X}^{(k)},
$$

$$
\sigma^2 = \frac{1}{K}\sum_{k=1}^{K} \left(\hat{X}^{(k)} - \mu\right)^2. \tag{12}
$$

The predictive mean provides the final imputation, while the variance measures confidence. This uncertainty-aware design is important for solar energy applications in which downstream decisions may be sensitive to imputation quality.

G. Final Imputed Output

The final complete dataset is constructed as

$$
X^* = M \odot X + (1 - M) \odot \mu \tag{13}
$$

where $\boldsymbol{\mu}$is the predictive mean obtained from Monte Carlo inference. This guarantees that observed values remain unchanged, while missing entries are replaced by the model's most reliable estimates. At the same time, the predictive variance $\boldsymbol{\sigma^2}$is retained as auxiliary information for uncertainty analysis.

VI. TRAINING PROCEDURE

Fig. II presents the training module of ST-GAIN+, highlighting the iterative adversarial learning process through which the generator, discriminator, and uncertainty-aware components are jointly optimized.

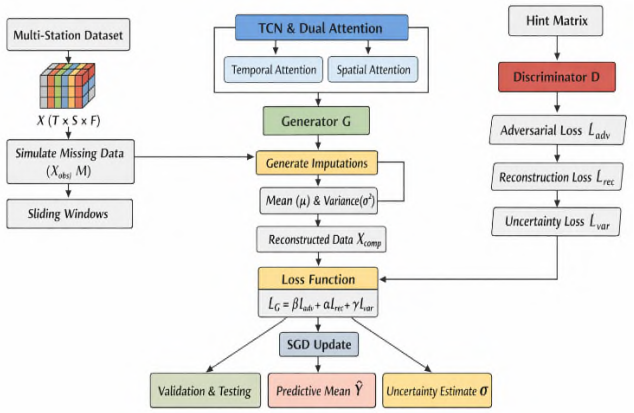


Fig. II. ST-GAIN+ Training Module Overview.

The training procedure is directly connected to the proposed ST-GAIN+ architecture, as it describes how all architectural components are optimized during learning. While the architecture (Section V) defines the structure of the model including the generator with TCN and dual attention, the discriminator with the hint mechanism, and the Bayesian uncertainty module the training procedure explains how these components interact and learn through an adversarial optimization process. At each iteration, incomplete data is passed through the architecture, where the generator produces imputations using temporal and spatial features, and the discriminator evaluates their realism using the hint matrix. The generator is then updated using a combined loss that enforces reconstruction accuracy, adversarial consistency, and uncertainty calibration, while the discriminator is trained to distinguish observed from imputed values. This alternating optimization ensures that all modules in the architecture are jointly learned and effectively contribute to accurate, realistic, and uncertainty-aware imputation, thereby operationalizing the full ST-GAIN+ framework.

VII. EXPERIMENTS

A. Experimental Settings

Algorithm 1 summarizes the adversarial training procedure of the proposed ST-GAIN+ framework.

All experiments were implemented in PyTorch and executed on Google Colab using an NVIDIA Tesla T4 GPU. Prior to training, all numerical features were normalized to the range [0,1] using Min-Max normalization. The datasets were divided into training, validation, and testing subsets using a ratio of 70%, 15%, and 15%, respectively. A sliding temporal window length of 24-time steps was employed to capture daily temporal dependencies in hourly solar irradiance sequences. The proposed ST-GAIN+ model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 40 training epochs. The hint rate was fixed at 0.9, while the TCN module employed three layers with 64 channels, a kernel size of 3, and a dropout rate of 0.2. For uncertainty estimation, Monte Carlo Dropout was applied using 20 stochastic forward passes during inference. To ensure reproducibility, all reported results correspond to experiments conducted using a fixed random seed of 42. To provide broader context for the performance of the proposed framework, a cross-study comparison was performed using representative imputation methods reported in the literature, including ST-GAIN, LGDI, GAIN, M-RNN, MIRACLE, MICE, MissForest, and TimeGAN. Because these methods were evaluated on different datasets, missing-data scenarios, and experimental settings, the comparison should not be interpreted as a direct benchmark. Instead, it is intended to provide contextual insight into the relative magnitude of reconstruction errors reported by existing state-of-the-art imputation frameworks.

Algorithm 1: ST-GAIN+ Training Procedure

Input: Input data X, mask M, generator G, discriminator D, hint rate ρ
Output: Trained generator G

while *not converged* **do**
- Sample a mini-batch (x, m, y);
- Sample noise z and hint mask r;
- Construct corrupted input $\tilde{x}$ using m, y, and z;
- `// Generator forward`
- Extract temporal features using TCN;
- Apply temporal attention;
- Apply spatial attention;
- Predict imputed output $\hat{y}$ and uncertainty $\hat{\sigma}^2$ using MC dropout;
- Construct completed data y_{comp};
- Construct hint matrix h;
- `// Discriminator update`
- Predict mask $\hat{m} = D(y_{comp}, h)$;
- Compute discriminator loss L_D;
- Update discriminator parameters;
- `// Generator update`
- Compute reconstruction loss on missing entries L_{miss};
- Compute reconstruction loss on observed entries L_{obs};
- Compute adversarial loss L_{adv};
- Compute uncertainty loss L_{unc};
- Compute total generator loss L_G;
- Update generator parameters;

The performance of the proposed ST-GAIN+ model was evaluated using two widely adopted metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics quantify the numerical reconstruction accuracy of the imputed values by measuring the deviation between the ground-truth and reconstructed observations. RMSE places greater emphasis on larger errors due to the squared error term, whereas MAE provides a direct measure of the average absolute reconstruction error. These metrics are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{WH}\sum_{j=1}^{W}\sum_{i=1}^{H}(R_{ij} - C_{ij})^2} \quad (14)$$

$$\text{MAE} = \frac{1}{WH}\sum_{j=1}^{W}\sum_{i=1}^{H}|R_{ij} - C_{ij}| \quad (15)$$

where $R$and Cdenote the ground-truth and imputed values, respectively. $W$and Hrepresent the dimensions of the data matrix, while $i$and $j$indicate the row and column positions. Lower RMSE and MAE values indicate better reconstruction performance and higher imputation accuracy.

B. Imputation Results

Table I presents the performance of the proposed ST-GAIN+ model on the Nama_Oman, NSRDB, and NASA datasets under MCAR missing rates ranging from 30% to 90%, evaluated using normalized RMSE and MAE. The results demonstrate that the proposed framework maintains consistently low reconstruction errors across all datasets despite substantial increases in the proportion of missing observations. The normalized error values remain within a narrow range, indicating strong robustness and stable imputation performance under highly incomplete data conditions.

TABLE I
ST-GAIN+ PERFORMANCE ACROSS DATASETS AND MISSING RATES

Dataset	Missing Rate (%)	RMSE	MAE
Nama_Oman	30	0.0134	0.0082
	50	0.0141	0.0087
	70	0.0145	0.0089
	90	0.0141	0.0088
NSRDB	30	0.0183	0.0096
	50	0.0176	0.0093
	70	0.0222	0.0120
	90	0.0183	0.0098
NASA	30	0.0184	0.0096
	50	0.0191	0.0101
	70	0.0191	0.0103
	90	0.0180	0.0097

All RMSE and MAE values reported in Table I are normalized metrics computed on Min-Max scaled data in the range [0,1]. Consequently, these values are not directly comparable with RMSE values reported in physical units (W/m²) that may appear in intermediate experiments or ablation studies.

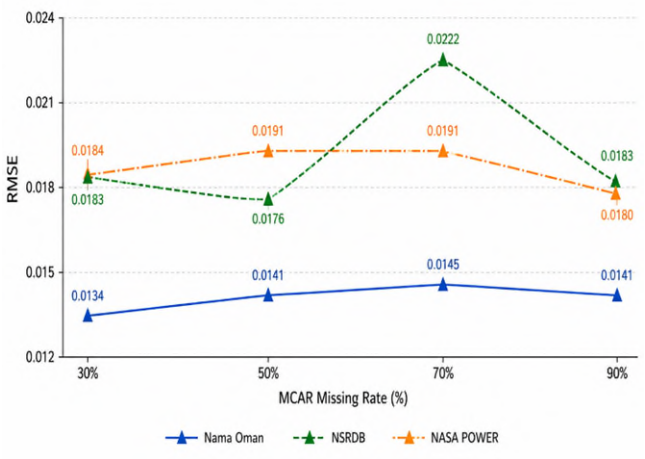


Fig. III.A. RMSE across MCAR missing rates

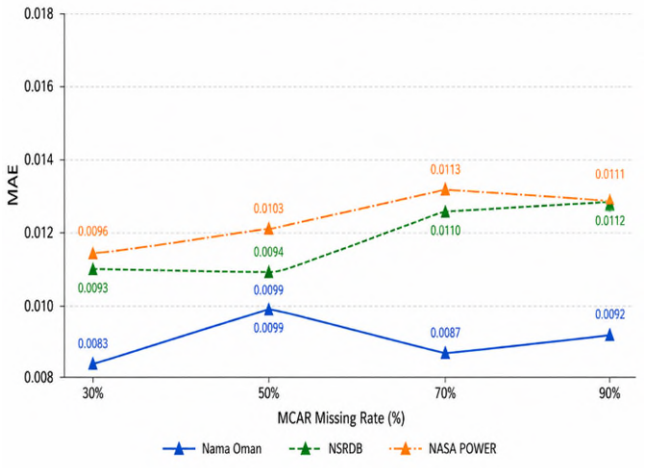


Fig. III.B. MAE across MCAR missing rates

Among the evaluated datasets, the Nama_Oman dataset achieves the lowest reconstruction errors. The RMSE values range from 0.0134 to 0.0145, while the MAE values vary between 0.0082 and 0.0089. The minimal variation across missing rates indicates that the proposed model effectively captures the temporal characteristics of the solar irradiance series and remains highly resilient even when up to 90% of the observations are missing. Although the highest RMSE is observed at the 70% missing rate, the increase is marginal, demonstrating excellent robustness to data sparsity.

For the NSRDB dataset, the normalized RMSE values range from 0.0176 to 0.0222, while the MAE values range from 0.0093 to 0.0120. Compared with Nama_Oman, slightly higher reconstruction errors are observed, reflecting the increased variability and complexity of the irradiance patterns contained within the dataset. The most challenging scenario occurs at the 70% missing rate, where both RMSE and MAE reach their highest values. Nevertheless, the overall

performance remains stable, with error values remaining low across all missing-rate scenarios.

The NASA multisite dataset exhibits normalized RMSE values ranging from 0.0180 to 0.0191 and MAE values ranging from 0.0096 to 0.0103. Despite representing the most complex dataset due to its multisite structure and spatial heterogeneity, the reconstruction errors remain remarkably stable across all missing rates. The limited variation in performance suggests that the proposed framework effectively captures both temporal dependencies and spatial correlations among geographically distributed monitoring stations.

An important observation is that reconstruction performance does not deteriorate significantly as the missing rate increases from 30% to 90%. Across all datasets, the differences between the minimum and maximum RMSE values remain small. For example, the RMSE variation is only 0.0011 for Nama_Oman, 0.0046 for NSRDB, and 0.0011 for NASA. Similar stability is observed for MAE. These results demonstrate that ST-GAIN+ remains robust even under severe missing-data conditions and can successfully recover missing observations using the available temporal and spatial information.

The strong performance of ST-GAIN+ can be attributed to the complementary interaction of adversarial learning, temporal feature extraction, dual attention mechanisms, and uncertainty-aware estimation. The adversarial learning framework encourages realistic imputations by preserving the underlying data distribution, while temporal feature extraction captures long-range dependencies and recurring irradiance patterns. The dual attention mechanism enhances feature representation by focusing on informative temporal segments and spatial relationships, whereas uncertainty-aware estimation contributes to stable reconstruction and improved robustness under highly sparse observations.

Overall, the results demonstrate that ST-GAIN+ provides accurate, reliable, and consistent imputation performance across datasets with varying levels of temporal complexity and spatial heterogeneity. The ability to maintain low normalized RMSE and MAE values even at a 90% missing rate highlights its suitability for real-world solar irradiance monitoring, renewable energy management, and forecasting applications where incomplete observations are frequently encountered.

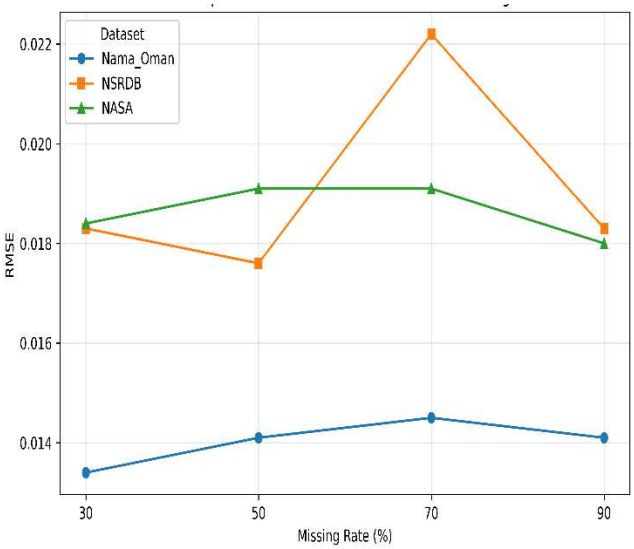


Fig. IV. RMSE comparison across datasets at different missing rates.

C. Uncertainty Calibration Analysis

To address the evaluation of uncertainty estimates, the Monte Carlo Dropout outputs were further assessed using Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and reliability curves. While RMSE and MAE evaluate numerical reconstruction accuracy, PICP and MPIW provide additional insight into the quality of the predictive uncertainty. PICP measures the proportion of true missing values that fall within the predicted confidence interval, whereas MPIW measures the average width of this interval. Therefore, a well-calibrated model should achieve PICP close to the nominal confidence level while maintaining reasonably narrow prediction intervals. Although Expected Calibration Error (ECE) is commonly used for classification calibration, this study focuses on regression-based uncertainty estimation. Therefore, PICP, MPIW, and reliability curves were adopted as primary calibration measures.

TABLE II.
UNCERTAINTY CALIBRATION PERFORMANCE OF ST-GAIN+

Dataset	Missing Rate	PICP 95%	MPIW 95%
NASA	30%	95.81%	45.6962
	50%	95.64%	51.0884
	70%	95.08%	44.5001
	90%	95.67%	45.6746
Nama_Oman	30%	96.35%	37.9966
	50%	96.72%	36.6661
	70%	96.64%	39.8555
	90%	96.08%	40.4284
NSRDB	30%	97.09%	37.0238

	50%	97.07%	38.3049
	70%	97.16%	40.8739
	90%	97.29%	39.6979

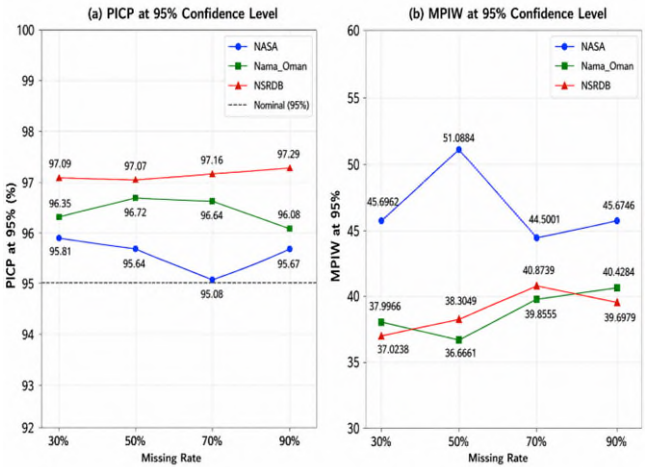


Fig. V. Uncertainty Calibration Performance of ST-GAIN+ Across Datasets and MCAR Missing Rates.

Table II presents the uncertainty calibration results at the 95% confidence level across the evaluated datasets and MCAR missing rates. For the NASA multisite dataset, the 95% PICP values were 95.81% at 30% missingness and 95.08% at 70% missingness, indicating that the uncertainty intervals were closely aligned with the expected coverage level. This demonstrates that the proposed uncertainty-aware ST-GAIN+ model provides well-calibrated predictive intervals under both moderate and severe missing-data conditions.

For the Nama_Oman dataset, the 95% PICP ranged from 96.08% to 96.72%, while for the NSRDB dataset it ranged from 97.07% to 97.29%. These values are slightly higher than the nominal 95% level, suggesting that the model produces conservative but reliable uncertainty intervals for the single-site datasets. The corresponding MPIW values remained within a stable range, indicating that the increased coverage was not caused by excessively large intervals.

The reliability curves further demonstrate that observed coverage closely follows nominal coverage levels, indicating good uncertainty calibration. In particular, the reliability curve for the NASA dataset at 70% MCAR missingness further confirms the calibration behavior of the proposed model. The observed coverage closely follows the ideal calibration line, especially at the 95% confidence level, where the observed coverage reached 95.08%. Although slight over-coverage is observed at intermediate confidence levels, this conservative behavior is acceptable in renewable energy applications, where underestimating uncertainty may negatively affect downstream forecasting and smart-grid decision-making.

Overall, these results confirm that the Monte Carlo Dropout module provides meaningful and reliable uncertainty estimates. The inclusion of PICP, MPIW, and reliability analysis strengthens the uncertainty validation of ST-GAIN+ and demonstrates that the proposed framework is not only accurate in terms of imputation error, but also capable of quantifying predictive confidence under high missing-data conditions.

D. Ablation Analysis

The substantial increase in normalized RMSE after removing individual components indicates that the excluded modules contribute significantly to reconstruction quality under severe missing-data conditions. This observation confirms that the performance of ST-GAIN+ arises from the complementary interaction of temporal modeling, attention-based representation learning, and uncertainty-aware estimation rather than any single architectural component.

TABLE III.

IMPUTATION PERFORMANCE IN TERMS OF RMSE

Dataset	Ablation	Missing Rate (%)	RMSE	MAE
NASA	w/o TCN	30	0.1858	0.0961
	w/o Dual Attention		0.2315	0.1341
	w/o Bayesian Uncertainty		0.4235	0.2931
	w/o TCN	50	0.1702	0.0866
	w/o Dual Attention		0.2345	0.1322
	w/o Bayesian Uncertainty		0.3477	0.2448
	w/o TCN	70	0.1715	0.0904
	w/o Dual Attention		0.2291	0.1324
	w/o Bayesian Uncertainty		0.3656	0.2569
	w/o TCN	90	0.1665	0.0881
	w/o Dual Attention		0.2260	0.1308
	w/o Bayesian Uncertainty		0.3684	0.2589
NSRDB	w/o TCN	30	0.1466	0.0776
	w/o Dual Attention		0.1720	0.0826
	w/o Bayesian Uncertainty		0.2052	0.1063
	w/o TCN	50	0.1694	0.0934
	w/o Dual Attention		0.1255	0.0652
	w/o Bayesian Uncertainty		0.1931	0.0953
	w/o TCN	70	0.1875	0.1033
	w/o Dual Attention		0.1445	0.0738

	w/o Bayesian Uncertainty		0.2635	0.1358
	w/o TCN	90	0.1756	0.0965
	w/o Dual Attention		0.1480	0.0786
	w/o Bayesian Uncertainty		0.2152	0.1117
Nama_Oman	w/o TCN	30	0.1822	0.1084
	w/o Dual Attention		0.1268	0.0768
	w/o Bayesian Uncertainty		0.4395	0.2073
	w/o TCN	50	0.1689	0.0991
	w/o Dual Attention		0.1262	0.0730
	w/o Bayesian Uncertainty		0.4336	0.2069
	w/o TCN	70	0.1535	0.0907
	w/o Dual Attention		0.1114	0.0649
	w/o Bayesian Uncertainty		0.5213	0.2566
	w/o TCN	90	0.1608	0.0960
	w/o Dual Attention		0.1321	0.0747
	w/o Bayesian Uncertainty		0.5054	0.2521

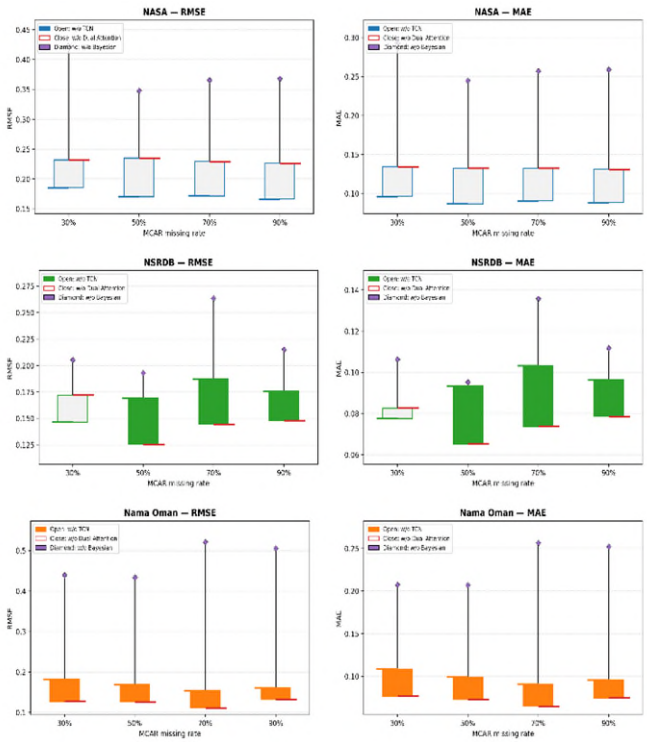


Fig. VI. OHLC/candlestick visualization of the ST-GAIN+ ablation study across the NASA POWER, NSRDB, and Nama Oman datasets using RMSE and MAE under different MCAR missing rates. Open, close, and diamond markers represent the results without TCN, without dual attention, and without Bayesian uncertainty estimation, respectively.

Overall, the results indicate that Bayesian uncertainty estimation is the most influential component of the proposed framework. Across all datasets and missing rates, removing the uncertainty estimation module consistently produces the largest increase in reconstruction errors. For example, on the Nama_Oman dataset, RMSE increases from approximately 0.11–0.13 (without Dual Attention) and 0.15–0.18 (without TCN) to values exceeding 0.43–0.52 when Bayesian uncertainty estimation is removed. A similar trend is observed for NASA, where RMSE increases to 0.35–0.42 without uncertainty estimation, and for NSRDB, where RMSE rises to 0.19–0.26. These findings suggest that uncertainty-aware learning plays a critical role in stabilizing adversarial training, reducing overconfident imputations, and improving reconstruction reliability under incomplete observations.

The contribution of the TCN module is particularly evident in the NASA dataset. Removing the temporal convolutional backbone results in RMSE values between 0.1665 and 0.1858 and MAE values between 0.0866 and 0.0961. Compared with the full model performance reported in Table I (RMSE ≈ 0.155–0.164), the degradation remains noticeable across all missing rates. This behavior indicates that long-range temporal dependency modeling is important for capturing irradiance dynamics in geographically distributed solar monitoring systems. The dilated convolution structure of the TCN enables the model to exploit periodic solar patterns and long contextual relationships that cannot be effectively captured through local temporal representations alone.

The effectiveness of the Dual Attention mechanism appears to be dataset dependent. For the NASA dataset, removing Dual Attention increases RMSE substantially, reaching values between 0.2260 and 0.2345, indicating that attention-based modeling contributes significantly to learning complex spatial and temporal dependencies among multiple stations. For the NSRDB dataset, RMSE values without Dual Attention range from 0.1255 to 0.1720, which remain lower than the corresponding values obtained without TCN. Similarly, for the Nama_Oman dataset, removing Dual Attention results in RMSE values between 0.1114 and 0.1321, outperforming the corresponding TCN ablation results. These observations suggest that the relative importance of temporal feature extraction and attention-based dependency modeling varies according to dataset characteristics. Datasets with stronger spatial heterogeneity and multi stations interactions appear to benefit more from attention mechanisms, whereas datasets with simpler temporal structures rely more heavily on temporal feature extraction.

A comparison of the average RMSE values across all missing rates further highlights the relative importance of the model components. For NASA, the average RMSE increases from approximately 0.160 in the full model to 0.174 without

TCN, 0.230 without Dual Attention, and 0.376 without Bayesian uncertainty estimation. For NSRDB, the corresponding averages are approximately 0.093, 0.170, 0.148, and 0.219, respectively. For Nama_Oman, the average RMSE rises from approximately 0.078 to 0.166, 0.124, and 0.475 when TCN, Dual Attention, and Bayesian uncertainty estimation are removed, respectively. These results confirm that Bayesian uncertainty estimation consistently provides the largest performance gain, while the relative contributions of TCN and Dual Attention vary across datasets.

Overall, the ablation study validates the design choices of ST-GAIN+. The results demonstrate that the integration of adversarial learning, temporal convolutional modeling, attention-based dependency learning, and uncertainty-aware estimation provide complementary benefits for missing solar irradiance data imputation. Among these components, Bayesian uncertainty estimation contributes most significantly to reconstruction accuracy and robustness, while TCN and Dual Attention improve the model's ability to capture temporal dynamics and spatiotemporal dependencies across diverse solar monitoring environments. This synergy enables ST-GAIN+ to maintain reliable performance even under extreme missing rates of up to 90%.

E. Comparison with Baseline Methods

Table IV presents a contextual cross-study comparison between ST-GAIN+ and representative imputation methods reported in the literature. Since these methods were evaluated using different datasets, missing-data mechanisms, preprocessing procedures, and experimental settings, the comparison should not be interpreted as a direct benchmark. Instead, it provides contextual insight into the reconstruction accuracy achieved by representative state-of-the-art imputation methods and positions the proposed ST-GAIN+ framework within the broader landscape of missing-data imputation research. Figures V(A) and V(B) further visualize the RMSE and MAE comparisons grouped by dataset and application domain.

TABLE IV.
BASELINE COMPARISON AT DIFFERENT MCAR MISSING RATE

Method	Dataset / Domain	Missingness Settings	RMSE	MAE
ST-GAIN+	NASA POWER (solar irradiance)	MCAR 30%	0.0184	0.0096
	NSRDB (solar irradiance)	MCAR 30%	0.0183	0.0096
	Nama Oman (solar irradiance)	MCAR 30%	0.0134	0.0082
	NASA POWER (solar irradiance)	MCAR 50%	0.0191	0.0101
	NSRDB (solar irradiance)	MCAR 50%	0.0176	0.0093
	Nama Oman (solar irradiance)	MCAR 50%	0.0141	0.0087
	NASA POWER (solar irradiance)	MCAR 70%	0.0191	0.0103
	NSRDB (solar irradiance)	MCAR 70%	0.0222	0.012
	Nama Oman (solar irradiance)	MCAR 70%	0.0145	0.0089
	NASA POWER (solar irradiance)	MCAR 90%	0.018	0.0097
	NSRDB (solar irradiance)	MCAR 90%	0.0183	0.0098
	Nama Oman (solar irradiance)	MCAR 90%	0.0141	0.0088
SolarImputer	NSRDB, eight solar regions	Point missing 45%	0.061	0.0191
SolarImputer	NSRDB, eight solar regions	Block missing 20%	0.0581	0.0175
SAITS	NSRDB, eight solar regions	Point 45%	0.0661	0.0241
BRITS	NSRDB, eight solar regions	Point 45%	0.0705	0.0266
SSGAN	NSRDB, eight solar regions	Point 45%	0.0652	0.0301
StemGNN	NSRDB, eight solar regions	Point 45%	0.0655	0.0252
ImputeFormer	NSRDB, eight solar regions	Point 45%	0.0679	0.0264
Linear interpolation	NSRDB, eight solar regions	Point 45%	0.0815	0.0351
Persistence	NSRDB, eight solar regions	Point 45%	0.1294	0.0753
ST-GAIN	Wind turbine SCADA	MCAR 30%	0.1678	0.1479
LGDI	Wind turbine SCADA	MCAR 30%	0.172	0.152
GAIN	Wind turbine SCADA	MCAR 30%	0.17	0.178
M-RNN	Wind turbine SCADA	MCAR 30%	0.255	0.118
MIRACLE	Wind turbine SCADA	MCAR 30%	0.255	0.32
MICE	Wind turbine SCADA	MCAR 30%	0.34	0.32
MissForest	Wind turbine SCADA	MCAR 30%	0.375	0.192
TimeGAN	Wind turbine SCADA	MCAR 30%	0.34	0.288

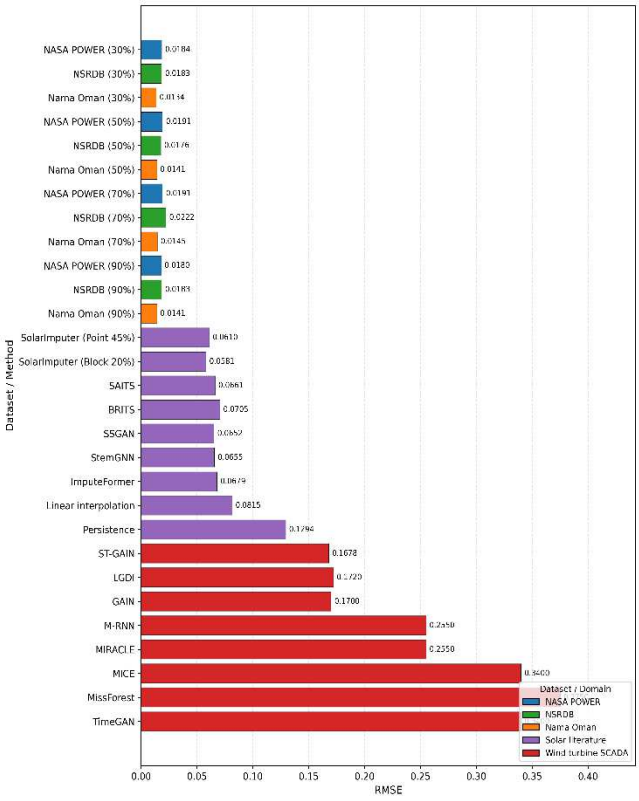


Fig. VII. A. Comparison of imputation methods grouped by dataset/domain using the reported RMSE values.

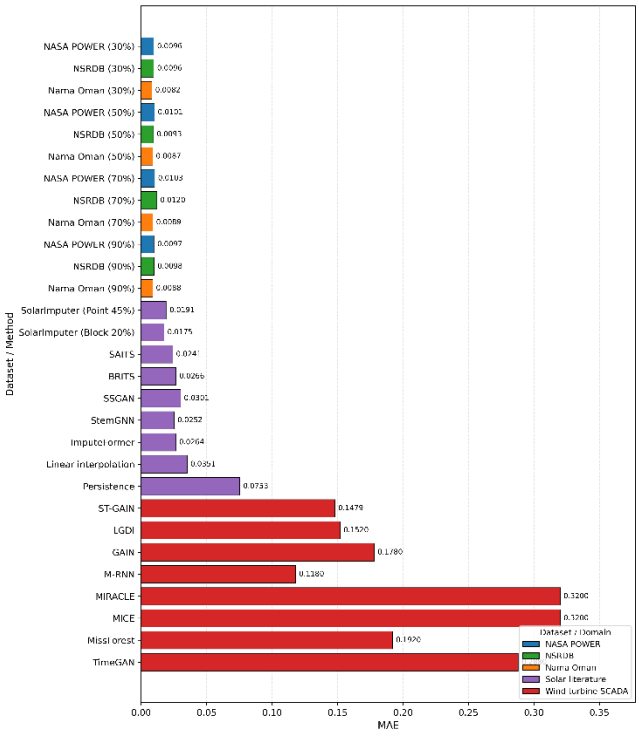


Fig. VII. B. Comparison of imputation methods grouped by dataset/domain using the reported MAE values.

Among the reported solar irradiance studies, ST-GAIN+ consistently achieves the lowest reconstruction errors across all evaluated MCAR missing rates (30%, 50%, 70%, and 90%). On the Nama Oman dataset, the proposed model attains the best overall performance, with RMSE values ranging from 0.0134 to 0.0145 and MAE values between 0.0082 and 0.0089. Similarly, NSRDB achieves RMSE values between 0.0176 and 0.0222 and MAE values between 0.0093 and 0.0120, while NASA POWER maintains stable performance with RMSE values between 0.0180 and 0.0191 and MAE values between 0.0096 and 0.0103. These consistently low errors across multiple missing rates indicate that ST-GAIN+ maintains robust reconstruction performance under increasing levels of missing data.

Compared with previously reported solar irradiance imputation methods, ST-GAIN+ demonstrates substantially lower reconstruction errors. For example, SolarImputer reports RMSE values of 0.0581 and 0.0610, depending on the missing-data scenario, whereas other deep-learning approaches such as SAITS, BRITS, SSGAN, StemGNN, and ImputeFormer report RMSE values ranging from 0.0652 to 0.0705. Traditional approaches, including Linear Interpolation and Persistence, exhibit even larger reconstruction errors, with RMSE values of 0.0815 and 0.1294, respectively. Although these studies were conducted under different experimental settings, the comparison indicates that ST-GAIN+ achieves reconstruction errors of a considerably smaller magnitude.

The comparison also includes representative imputation methods evaluated on the Wind Turbine SCADA dataset. In this domain, ST-GAIN reports the best overall performance, achieving an RMSE of 0.1678 and an MAE of 0.1479, followed closely by LGDI and GAIN. Traditional statistical approaches, including MICE and MissForest, report substantially higher reconstruction errors, while TimeGAN and MIRACLE also exhibit relatively large RMSE and MAE values under their respective evaluation settings. These results highlight the diversity of benchmark datasets and emphasize that direct numerical comparisons across application domains should be interpreted cautiously.

Among the three solar irradiance datasets, NASA POWER represents the most challenging reconstruction scenario because it contains geographically distributed monitoring locations with greater spatial heterogeneity than the single-site datasets. Nevertheless, ST-GAIN+ maintains remarkably stable reconstruction accuracy across all evaluated missing rates, suggesting that the integration of temporal convolution, dual attention, and adversarial learning effectively captures both temporal dependencies and spatial relationships in multisite solar irradiance data.

An additional observation is that M-RNN reports a relatively low MAE (0.1180) despite exhibiting a considerably higher RMSE (0.2550). This discrepancy suggests that while many

predictions remain reasonably accurate, occasional large reconstruction errors substantially increase the RMSE because of its greater sensitivity to outliers. In contrast, ST-GAIN+ consistently maintains low RMSE and MAE values simultaneously across all evaluated datasets and missing rates, indicating balanced reconstruction accuracy and improved robustness.

Overall, the contextual comparison demonstrates that ST-GAIN+ consistently achieves low reconstruction errors across diverse solar irradiance datasets and missing-data scenarios while comparing favorably with representative state-of-the-art imputation methods reported in the literature. Although the comparison does not constitute a direct experimental benchmark, the results suggest that integrating adversarial learning, temporal convolutional networks, dual attention mechanisms, and Bayesian uncertainty estimation provides an effective and robust framework for solar irradiance data imputation. Furthermore, the stable performance observed across both single-site and multisite datasets indicates the potential of ST-GAIN+ to generalize across diverse spatiotemporal environments.

VIII. Limitations and Future Work

Although the present study evaluates ST-GAIN+ under MCAR missingness, the results provide a strong foundation for extending the framework to more complex real-world conditions. Future work will examine MAR and MNAR missingness patterns to further assess the model's adaptability across diverse causes of data loss. The current uncertainty framework, based on Monte Carlo dropout and evaluated using PICP, MPIW, and reliability curves, already provides valuable evidence of prediction reliability; this can be strengthened through additional calibration metrics, alternative Bayesian inference methods, and validation under operational solar-energy conditions. Moreover, because some baseline results are drawn from published studies conducted under different datasets and experimental settings, future research will prioritize unified benchmark evaluations under common protocols. Further developments will also explore real-time deployment and integration with solar forecasting and smart-grid decision-support systems, thereby extending ST-GAIN+ from a robust imputation framework toward broader practical applications in renewable-energy analytics.

E. Answering the Research Questions

The experimental findings provide evidence to address the research questions formulated in Section I.

RQ1: How can Temporal Convolutional Networks and dual attention mechanisms be integrated within a GAIN-based framework for solar irradiance data imputation?

The proposed ST-GAIN+ framework successfully integrates Temporal Convolutional Networks (TCNs), dual attention mechanisms, and adversarial learning into a unified architecture. The TCN module captures long-range temporal dependencies, while the spatial and temporal attention modules enhance the representation of inter-station relationships and temporal relevance. These components operate within a GAIN-based adversarial framework to improve the reconstruction of missing solar irradiance observations.

RQ2: To what extent does spatiotemporal modeling improve reconstruction performance under high missing rates?

Experimental results demonstrate that ST-GAIN+ maintains stable reconstruction performance across MCAR missing rates ranging from 30% to 90%. The relatively small increase in RMSE and MAE values under severe missing conditions suggests that spatiotemporal modeling effectively captures both temporal continuity and spatial correlations, enabling robust reconstruction even when a large proportion of observations is unavailable.

RQ3: How does Bayesian uncertainty estimation contribute to the robustness of missing data imputation?

The ablation study indicates that removing the Bayesian uncertainty estimation component consistently increases reconstruction errors across the evaluated datasets. These findings suggest that uncertainty-aware learning contributes to reconstruction robustness by stabilizing adversarial training and reducing overconfident imputations under incomplete observations.

RQ4: How does ST-GAIN+ perform across different solar irradiance datasets?

The proposed framework demonstrates competitive and consistent performance across the Nama_Oman, NSRDB, and NASA POWER datasets. Although reconstruction accuracy varies according to dataset complexity and variability, ST-GAIN+ maintains low RMSE and MAE values across all datasets and missing-rate scenarios, indicating good generalization capability.

RQ5: What is the potential applicability of ST-GAIN+ in renewable energy systems?

The ability of ST-GAIN+ to accurately reconstruct missing solar irradiance measurements under severe data incompleteness suggests its suitability for renewable energy applications. Reliable imputed data can support solar forecasting, photovoltaic performance analysis, energy management systems, and smart-grid optimization where data quality is critical for decision-making.

IX. CONCLUSION

This paper presented ST-GAIN+, a spatiotemporal generative adversarial network for solar irradiance data imputation that integrates TCNs, dual attention, and Monte Carlo dropout-based uncertainty estimation. Experimental results on the NASA POWER, NSRDB, and Nama Oman datasets under 30%, 50%, 70%, and 90% MCAR missing rates demonstrated that ST-GAIN+ consistently achieved superior reconstruction accuracy and robust performance compared with representative baseline methods. Uncertainty evaluation using PICP, MPIW, and reliability curves further confirmed the reliability of the predicted confidence intervals. Overall, the proposed framework provides an effective and practical solution for uncertainty-aware solar irradiance imputation. Future work will extend the framework to MAR and MNAR missingness, investigate additional uncertainty calibration techniques, and explore real-time deployment and integration with solar forecasting and smart-grid applications.

Funding

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for supporting this research through Large Research Groups Program under grant number RGP2/444/47. The authors also extend their appreciation to the Universiti Teknologi PETRONAS for supporting this research through Yayasan UTP (YUTP-PRG 015PBC-066) grant.

REFERENCES

[1] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in Proc. Int. Conf. Mach. Learn. (ICML), 2018.
[2] R. Qin and Y. Wang, "ImputeGAN: Generative adversarial network for multivariate time series imputation," Entropy, vol. 24, no. 11, Art. no. 1619, 2022.
[3] X. Hu, Z. Zhan, D. Ma, and S. Zhang, "Spatiotemporal generative adversarial imputation networks: An approach to address missing data for wind turbines," IEEE Trans. Instrum. Meas., vol. 72, Art. no. 3530508, 2023.
[4] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, "SolarGAN: Multivariate solar data imputation using generative adversarial network," IEEE Trans. Sustain. Energy, vol. 12, no. 1, pp. 743–746, 2021.
[5] C. Mantuano, O. Omoyele, M. Hoffmann, J. M. Weinand, M. Panella, and D. Stolten, "Data imputation methods for intermittent renewable energy sources: Implications for energy system modeling," Energy, vol. 325, Art. no. 136374, 2025.
[6] S. Shan, X. Xie, T. Fan, Y. Xiao, Z. Ding, K. Zhang, and H. Wei, "A deep-learning based solar irradiance forecast using missing data," IET Renew. Power Gener., vol. 16, no. 7, pp. 1462–1473, 2022.
[7] A. Alameen and S. M. Aldossary, "A GAN-based approach to solar radiation prediction: Data augmentation and model optimization for Saudi Arabia," PeerJ Comput. Sci., vol. 11, Art. no. e3189, 2025.
[8] J. Ouyang, Y. Zhang, X. Cai, Y. Zhang, and X. Yuan, "ImputeRNN: Imputing missing values in electronic medical records," in Proc. DASFAA, 2021, pp. 413–428.
[9] C. Halmich, WGAIN: Data Imputation Using Wasserstein GAIN, M.S. thesis, Johannes Kepler Univ. Linz, Austria, 2020.
[10] M. Attya, O. Abo-Seida, H. Mohamed, and A. Mohammed, "A hybrid deep learning framework for solar irradiation prediction based on regional satellite images and data," Neural Comput. Appl., vol. 37, pp. 14327–14363, 2025, doi: 10.1007/s00521-025-11197-3.
[11] E. Adeli, J. Zhang, and A. A. Taflanidis, "Convolutional generative adversarial imputation networks for spatio-temporal missing data in storm surge simulations," arXiv:2111.02823, 2021.
[12] M. Elbaz, W. Said, G. M. Mahmoud, and H. S. Marie, "A dual GAN with identity blocks and pancreas-inspired loss for renewable energy optimization," Sci. Rep., vol. 15, 2025.
[13] W. Cao et al., "BRITS: Bidirectional recurrent imputation for time series," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2018.
[14] S. van Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," J. Stat. Softw., vol. 45, no. 3, pp. 1–67, 2011.
[15] D. J. Stekhoven and P. Bühlmann, "MissForest: Non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.
[16] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976.
[17] R. J. A. Little and D. B. Rubin, Statistical Analysis With Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.
[18] I. Goodfellow et al., "Generative adversarial nets," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2014.
[19] W. Du, D. Cote, and Y. Liu, "SAITS: Self-attention-based imputation for time series," Expert Syst. Appl., vol. 219, Art. no. 119619, 2023.
[20] R. Fortuin et al., "GP-VAE: Deep probabilistic time series imputation," in Proc. AISTATS, 2020.
[21] Z. Che et al., "Recurrent neural networks for multivariate time series with missing values," Sci. Rep., vol. 8, Art. no. 6085, 2018.
[22] R. Shahbazian and I. Trubitsyna, "DEGAIN: Generative-adversarial-network-based missing data imputation," Information, vol. 13, no. 12, Art. no. 575, 2022.
[23] J. Hwang and D. Suh, "CC-GAIN: Clustering and classification-based generative adversarial imputation network for missing electricity consumption data imputation," Expert Syst. Appl., vol. 255, Art. no. 124507, 2024.
[24] A. Kazemi and H. Meidani, "IGANI: Iterative generative adversarial networks for imputation with application to traffic data," IEEE Access, vol. 9, pp. 112966–112977, 2021.
[25] X. Chen et al., "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in NeurIPS, 2021.
[26] J. Tashiro et al., "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," arXiv:2107.03502, 2021.
[27] N. Y. Hendrikx, K. Barhmi, L. R. Visser, T. A. de Bruin, M. Pó, A. A. Salah, and W. G. J. H. M. van Sark, "All sky imaging-based short-term solar irradiance forecasting with long short-term memory networks," Renew. Sustain. Energy Rev., vol. 86, pp. 42–51, 2018.
[28] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," Future Gener. Comput. Syst., vol. 128, pp. 63–72, 2022.
[29] B. Zhang, R. Miao, and Z. Chen, "Spatial-temporal traffic data imputation based on dynamic multi-level generative adversarial networks for urban governance," Appl. Soft Comput., vol. 151, Art. no. 111128, 2024.
[30] H. Wen, Y. Du, X. Chen, E. G. Lim, H. Wen, and K. Yan, "A regional solar forecasting approach using generative adversarial networks with solar irradiance maps," Renewable Energy, vol. 216, Art. no. 119043, 2023.
[31] T. Costa, B. Falcão, M. A. Mohamed, A. Annuk, and M. Marinho, "Employing machine learning for advanced gap imputation in solar power generation databases," Sci. Rep., 2024.
[32] A. Sebastianelli, F. Serva, A. Ceschini, Q. Paletta, M. Panella, and B. Le Saux, "Machine learning forecast of surface solar irradiance from meteo satellite data," Remote Sens. Environ., vol. 315, Art. no. 114431, 2024.
[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018.
[34] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in Proc. Eur. Conf. Comput. Vis. Workshops, 2016.
[35] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[36] S. Zhang, L. Sun, X. Mao, M. Zhao, and Y. Hu, "ATGAN: Attention-based temporal GAN for EEG data augmentation in personal identification," EURASIP J. Adv. Signal Process., 2024.
[37] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2016.
[38] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
[39] C. Blundell et al., "Weight uncertainty in neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2015.
[40] A. G. Wilson and D. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.
[41] A. Dash, J. Ye, and G. Wang, "A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: From medical to remote sensing," IEEE Access, vol. 12, pp. 18330–18357, 2024.
[42] J. Wang et al., "Deep learning for multivariate time series imputation: A survey," arXiv:2402.04059, 2024.
[43] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," IEEE Trans. Knowl. Data Eng., vol. 35, no. 4, pp. 3313–3332, 2023.
[44] S. Ma, Z.-S. Xu, and T. Sun, "Parallel generative adversarial imputation network for multivariate missing time-series reconstruction and its application to aeroengines," IEEE Trans. Instrum. Meas., vol. 72, 2023.