https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

Attention-Driven and Multi-Scale Feature Integrated Approach for Earth Surface Temperature Data Reconstruction

Minghui Zhang[1], Yunjie Chen[1], Fan Yang[1], and Zhengkun Qin[2]

[1]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Atmospheric Sciences, Nanjing University of information Science and Technology, Nanjing 210044, China

Correspondence: Yunjie Chen (priestcyj@nuist.edu.cn)

Abstract. High-resolution observation is crucial for studying surface temperatures characterized by complex variations, particularly surface air temperatures in oceanic regions, which serve as significant indicators of sea-air coupling changes. Due to the scarcity of conventional observations of surface atmospheric temperatures in these areas, high-resolution surface atmospheric temperature data derived from satellite inversion has become the primary source of information. However, missing data resulting from factors such as the orbital spacing of polar satellites, cloud cover, sensor errors, and other disruptions poses substantial challenges to Earth Surface Temperature (EST) estimation. In this paper, we introduce ESTD-Net, a novel deep learning-based model designed for surface temperature data inpainting. ESTD-Net incorporates an enhanced multi-head context attention mechanism and a modified transformer block to capture long-range pixel dependencies, thereby improving the model's ability to focus on boundary regions. Additionally, we integrate a convolutional U-Net to refine high-frequency details, leveraging the texture enhancement capabilities of convolutional neural networks (CNNs) to further elevate the quality of the reconstructed images. The model is augmented by two key innovations: (1) Weighted Reconstruction Loss, which prioritizes masked regions to ensure accurate reconstruction of missing data; and (2) Gradient Consistency Regularization, which minimizes gradient discrepancies between ground truth and reconstructed images to ensure structural coherence and consistency. Evaluation results demonstrate that ESTD-Net outperforms existing methods in both pixel-level accuracy and perceptual quality. Our approach offers a robust and reliable solution for restoring earth surface temperature data.

1 Introduction

Earth Surface Temperature (EST) refers to the kinetic temperature of the Earth's surface, encompassing both land and ocean regions. In oceanic areas, EST is a crucial parameter that reflects the thermodynamic interactions between the ocean surface and the atmosphere, playing a vital role in ocean-atmosphere coupling processes. According to surface energy balance (SEB) theory, the ocean surface absorbs energy from both incoming solar radiation and atmospheric long-wave radiation. This absorbed energy is redistributed through several mechanisms: (1) outgoing thermal radiation, which directly influences EST; (2) vertical heat transport via ocean mixing and conduction; (3) turbulent heat exchanges at the air-sea interface; and (4) phase changes in surface water, including evaporation and condensation. Given its importance in climate and weather systems, EST over oceanic regions is typically estimated using various observational approaches, including in situ measurements, reanalysis datasets, ocean models, and satellite remote sensing techniques (Zhou et al., 2018). Among these methods, satellite-derived measurements offer a highly efficient and accurate means of capturing global-scale EST variations, facilitating continuous monitoring of temperature fluctuations across the ocean surface. These observations are essential for understanding large-scale climate dynamics, enhancing numerical weather prediction, and supporting oceanographic and meteorological research.

Cloud cover presents a significant challenge in obtaining accurate EST data over oceanic regions, as it consistently obscures more than 55% of the Earth's surface (King et al., 2013). Clouds obstruct satellite sensors from detecting surface thermal radiation, resulting in extensive missing data in ocean temperature observations. This issue becomes particularly problematic when cloud masking is inadequate, as thin cirrus clouds can partially obscure the ocean surface, leading to anomalously low temperature readings, especially during daytime. Additionally, sensor malfunctions and gaps in satellite coverage further exacerbate data deficiencies. These missing data points introduce considerable uncertainty in various oceanographic and atmospheric applications, including EST spatiotemporal variability analysis (Xu et al., 2023), air-sea interaction studies, ocean heat content estimation, and numerical weather prediction (Deo and Şahin, 2017). Addressing these data gaps is essential for improving climate modeling, understanding large-scale ocean-atmosphere exchange processes, and enhancing the accuracy of temperature retrievals based on satellite remote sensing.

Despite these challenges, EST remains a crucial variable in the climate and oceanic systems, with applications spanning ocean circulation studies, air-sea interactions, marine ecosystem monitoring, and climate change assessments. Recent advancements in satellite remote sensing have significantly enhanced the accessibility of global EST datasets, providing a more comprehensive alternative to traditional in situ ocean temperature measurements. Unlike geostationary satellites, which offer high temporal resolution but are limited to fixed observational coverage, polar-orbiting satellites such as FY-3D provide near-global coverage, making them essential for large-scale ocean temperature monitoring. One of the key advantages of microwave imaging instruments, such as the MWRI onboard FY-3D, is their ability to penetrate most non-precipitating clouds, thereby facilitating more comprehensive retrievals of ocean surface temperature. However, these instruments also have inherent limitations, including narrow swath widths that result in significant inter-orbital gaps, particularly in tropical regions. These extensive data gaps pose a considerable challenge, as conventional interpolation techniques often fail to deliver reliable reconstructions due to the high spatial variability of oceanic temperature patterns. In this context, deep learning-based image inpainting methods present a promising solution for reconstructing missing EST data with greater accuracy and robustness.

EST data can be effectively represented as image-like datasets, making image inpainting a relevant approach for restoring missing or degraded observations. Image inpainting has emerged as a significant research direction in computer vision, aiming to automatically complete incomplete images (Elharrouss et al., 2020). With advancements in deep learning, convolutional neural networks (CNNs), such as U-Net (Ronneberger et al., 2015), and self-attention-based architectures like the Transformer (Vaswani et al., 2017), have driven substantial progress in image inpainting, leading to their widespread application in tasks involving image reconstruction. In 2016, Pathak et al. (2016) introduced a CNN-based autoencoder for image inpainting that learned both low-level features and high-level semantics by alternately training on known and unknown regions to achieve automatic completion. Building on this foundation, Iizuka et al. (2017) proposed a GAN-based inpainting method in 2017, utilizing both global and local discriminators to generate high-quality and diverse inpainting results. In 2021, Deng et al.

(2021) developed a fully convolutional network with attention modules that improved the model's ability to capture spatial affinities between different image regions, leading to enhanced inpainting quality and consistency. In the domain of data inpainting, researchers have successfully applied convolutional neural networks—including fully convolutional networks, U-Net (Lepetit et al., 2021), GANs (Geiss and Hardin, 2021), and conditional GANs (Tan and Chen, 2023)—to address tasks such as multisource data fusion and recovery (Xie et al., 2020). By leveraging the power of deep learning, these approaches enable neural networks to learn high-level semantic features, facilitating the generation of high-quality inpainted results. Consequently, these methods achieve performance levels that significantly surpass traditional data correction techniques within the context of data inpainting applications.

Despite recent advancements, deep neural network methods still encounter specific challenges in data inpainting tasks. One significant challenge arises from the differences between conventional image data and pixel-level remote sensing data, such as satellite and radar imagery. These datasets often exhibit complex spatial features and high spatial resolution (Atlas et al., 1973; Lengfeld et al., 2020), characterized by fine-scale structures and surface roughness that complicate their analysis. To evaluate the quality of these features, metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) (Hore and Ziou, 2010) are commonly employed. Deep learning models frequently struggle to accurately capture discrete features, which can result in issues like over-smoothing in the inpainted regions, loss of critical details, and an increase in false positives. To address these limitations, recent approaches have been proposed to mitigate over-smoothing and enhance the preservation of important details in image inpainting tasks (Petrovska et al., 2020; Wang et al., 2022). A key aspect of successful inpainting is the ability to effectively capture contextual information, especially when dealing with large missing regions. To generate realistic structures and textures for these areas, it is crucial to leverage non-local priors and understand the broader context of the image. Such methods enable the model to draw relevant information from distant parts of the image, thereby ensuring more accurate and natural inpainting (Berman et al., 2016; Wang et al., 2018). To explicitly model long-range dependencies, some studies (Xie et al., 2019; Yi et al., 2020) have integrated attention modules into CNN-based generators. However, due to the quadratic computational complexity of attention mechanisms, these modules are typically limited to small-scale feature maps, restricting the full utilization of long-range context modeling. Unlike CNNs with attention modules, transformers (Vaswani et al., 2017) are inherently suited for non-local modeling since attention is a fundamental component in each block. Recent research (Wan et al., 2021; Yu et al., 2021) has explored transformer architectures for addressing inpainting tasks. However, due to computational limitations, these approaches often restrict the use of transformers to low-resolution predictions, which can lead to coarse and incomplete image structures. This limitation can significantly degrade overall inpainting quality, especially when handling large missing regions.

In this paper, we present an advanced transformer architecture specifically designed for image inpainting. In scenarios where useful information is sparse, we have observed that the standard transformer block struggles to perform effectively during adversarial training. To address this challenge, we propose modifications to the original transformer block aimed at improving both stability and performance. Specifically, we eliminate traditional layer normalization (Lei Ba et al., 2016) and transition from residual learning to fusion learning through feature concatenation. Additionally, to tackle the computational challenges posed by intensive interactions among numerous tokens in high-resolution inputs, we introduce a modified version of multi-head self-attention, termed multi-head context attention (Li et al., 2022). This variant computes non-local dependencies using only a subset of valid tokens. A dynamic mask, initialized by the input mask and updated through spatial constraints and long-range interactions, selectively chooses these tokens, thereby enhancing computational efficiency without compromising performance. Our contributions are as follows:

- We propose a novel gradient consistency regularization framework that enforces physical consistency in inpainted regions by minimizing the L1-norm of gradient discrepancies between generated and ground-truth images. This method excels in preserving critical physical properties (e.g., thermal gradients), significantly improving both visual fidelity and physical accuracy. Its robustness is particularly evident when handling images with complex gradient structures, setting a new benchmark for physics-aware inpainting.

- We design an adaptive weighted reconstruction loss that dynamically prioritizes missing regions during optimization. This mechanism forces the network to allocate higher attention to masked areas, substantially improving inpainting precision while maintaining global coherence.

- We develop a boundary-aware transformer module with reinforced attention mechanisms for edge preservation. By explicitly modeling boundary pixel relationships, it achieves subpixel-level accuracy in transition zones, yielding seamless blending between inpainted and original regions.

- We integrate a lightweight CNN-based U-Net for autoregressive refinement, capitalizing on its local texture modeling strengths. This hybrid design effectively suppresses local artifacts while maintaining computational efficiency, outperforming pure transformer architectures in fine-detail recovery.

- We curate a temporally diagnostic dataset of surface temperatures at 0600/1800 UTC (capturing thermal transition states during diurnal minima/maxima). This uniquely timed data provides critical baselines for studying climate dynamics, with direct applications in meteorology, agroecology, and environmental modeling-enabling new insights into diurnal thermal inertia and its systemic effects.

2 Related Work

Accurate reconstruction of missing values in EST data represents a critical challenge in geoscientific research. Existing methodologies for EST gap-filling can be systematically classified into three principal paradigms: Spatial reconstruction methods, temporal reconstruction methods and spatiotemporal reconstruction methods. Spatial reconstruction methods utilize surrounding valid pixels to estimate missing values, employing interpolation techniques such as inverse distance weighting (IDW) (Kilibarda et al., 2014; Fleit, 2024), cokriging interpolation (Dowd and Pardo-Igúzquiza, 2024), and spline interpolation (Li and Heap, 2014). These methods are straightforward to implement and perform effectively in homogeneous areas with limited missing data. However, their performance deteriorates as the amount of missing data increases, particularly in complex terrains where capturing spatial patterns becomes more challenging.

Temporal reconstruction methods rely on complementary images from nearby time intervals to estimate missing pixels. Common approaches include linear temporal interpolation (Zhang et al., 2015), harmonic analysis (Mohanasundaram et al., 2023), and temporal Fourier analysis (Scharlemann et al., 2008). More advanced techniques, such as LSTM neural networks (Cui et al., 2022), multi-temporal Bayesian dictionary learning (Li et al., 2014), and time-aware implicit neural representations (Wang et al., 2023), have been explored to capture the temporal variability within EST time series more effectively. While temporal reconstruction methods successfully capture time-dependent patterns, they may encounter difficulties when spatial context is not adequately integrated. To address the limitations of purely spatial or temporal approaches, spatiotemporal methods have been developed. These methods combine both spatial and temporal information to reconstruct missing EST values more comprehensively. For instance, Liu et al. (2017) introduced a spatiotemporal reconstruction technique for Feng Yun-2F satellite EST data, achieving root mean square error (RMSE) values within 2°C in most cases. Similarly, Weiss et al. (2015) developed a gap-filling method that integrates neighboring data with historical data from different time periods. While these techniques offer certain advantages, they often require substantial manual intervention and depend heavily on large datasets. Additionally, their performance can degrade in the presence of extensive missing data, as they struggle to capture the complex spatiotemporal relationships inherent in the data.

Methodologically, recovering missing data in EST is analogous to image inpainting, where the objective is to restore missing regions within an image. Traditional image inpainting techniques are generally categorized into two types: diffusion-based methods (Ballester et al., 2001) and patch-based approaches (Criminisi et al., 2004). Diffusion-based methods propagate pixel values from neighboring regions to fill in missing areas, similar to techniques such as linear interpolation or nearest neighbor. In contrast, patch-based methods copy pixel information from known regions, utilizing strategies such as mean imputation, k-nearest neighbors (KNN), or regression to restore missing values. Traditional image inpainting methods often struggle to preserve semantic coherence and texture consistency, particularly when dealing with large missing regions. This limitation parallels the challenges faced by conventional data recovery techniques when addressing extensive missing data. In contrast, recent advancements in deep learning have significantly improved image inpainting, resulting in notable enhancements in both performance and consistency. Techniques such as low-rank decomposition, generative models (e.g., VAE, GAN), and encoder-decoder CNN architectures have proven highly effective in producing high-quality inpainting results. For instance, Malek et al. (2017) employed a contextualized autoencoder CNN to reconstruct cloud-contaminated remote sensing images, addressing both pixel- and patch-level inpainting. Following this, several variants of the U-Net architecture (Liu et al., 2020; Yan et al., 2018; Zeng et al., 2019) have been introduced, further enhancing the performance of image completion tasks.

3 Methodology

3.1 Overall Architecture

In this paper, we propose the ESTD-Net method for reconstructing missing data, utilizing a two-stage architecture specifically optimized for surface temperature imagery reconstruction. This two-stage design is crucial for effectively capturing both global structures and local details, thereby ensuring more accurate and visually coherent reconstructions. In the first stage,

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

the network integrates a convolutional module with a transformer module to leverage both local spatial correlations and long-range dependencies, resulting in an initial reconstruction. The second stage employs a Conv-U-Net structure to further refine the results, enhancing fine-grained details and structural consistency. Additionally, the discriminator architecture is inspired by CoModGAN (Zhao et al., 2021), contributing to the overall enhancement of the model's performance by ensuring more realistic temperature distributions.

Given an input of size $H \times W$, the convolutional module first processes the input channels—comprising three image channels and one mask channel—transforming them into 180 feature channels through a series of convolutions. Subsequently, two strided convolutions, each with a stride of 2, are applied to downsample the feature map to a size of $\frac{H}{4} \times \frac{W}{4}$. The extracted features are then converted into tokens, which are fed into the masked transformer module. The masked transformer consists of five stages, with block configurations of $\{2,3,4,3,2\}$ corresponding to feature map sizes of $\{\frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{8} \times \frac{W}{8}, \frac{H}{4} \times \frac{W}{4}\}$. Both downsampling and upsampling operations are performed using convolutional layers. Details of the transformer block architecture are provided in Section 3.3. The output tokens from the transformer are reshaped into a 2D feature map, which is subsequently passed to the reconstruction module.

The convolutional reconstruction module upsamples the feature map from $\frac{H}{4} \times \frac{W}{4}$ to the original resolution of $H \times W$, producing a complete image. In the second stage, a Conv-U-Net is employed to refine the output by leveraging both the coarse prediction and the input mask. This network first downsamples the features to $\frac{H}{16} \times \frac{W}{16}$ and then upsamples them back to the original size of $H \times W$, enhancing high-fidelity details in the image. Shortcut connections are incorporated at each resolution level to preserve essential spatial information. The encoder begins with 64 convolutional channels, doubling the channel count after each downsampling step until reaching a maximum of 512 channels. The decoder is symmetrically structured, halving the number of channels at each upsampling step, thereby ensuring effective information flow and detail restoration throughout the network.

As illustrated in Fig. 1, the architecture of ESTD-Net seamlessly integrates these components to leverage the strengths of both convolutional and transformer-based approaches. The convolutional module efficiently extracts key tokens, while the transformer module utilizes the multi-head context attention mechanism, as outlined in the MAT framework (Li et al., 2022), to capture long-range dependencies between features. This enables more accurate and context-aware reconstructions. The output tokens are further refined through a convolution-based reconstruction module, which restores the spatial resolution to match the input dimensions. The subsequent Conv-U-Net stage enhances high-frequency details by leveraging the local texture refinement capabilities of CNNs, thereby improving the fidelity of the reconstruction.

3.2 Convolutional Module

Given a masked image $X_M = X \odot M$, where M is a binary mask that assigns a value of 1 to valid (error-free) pixels and 0 to invalid (missing or erroneous) ones, the goal of image completion is to generate visually coherent and semantically meaningful content for the missing regions. The convolutional module processes this incomplete image X_M along with the mask M, producing feature maps at a reduced spatial resolution of $\frac{1}{4}$ of the original dimensions. These feature maps are then utilized as tokens for further processing.

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

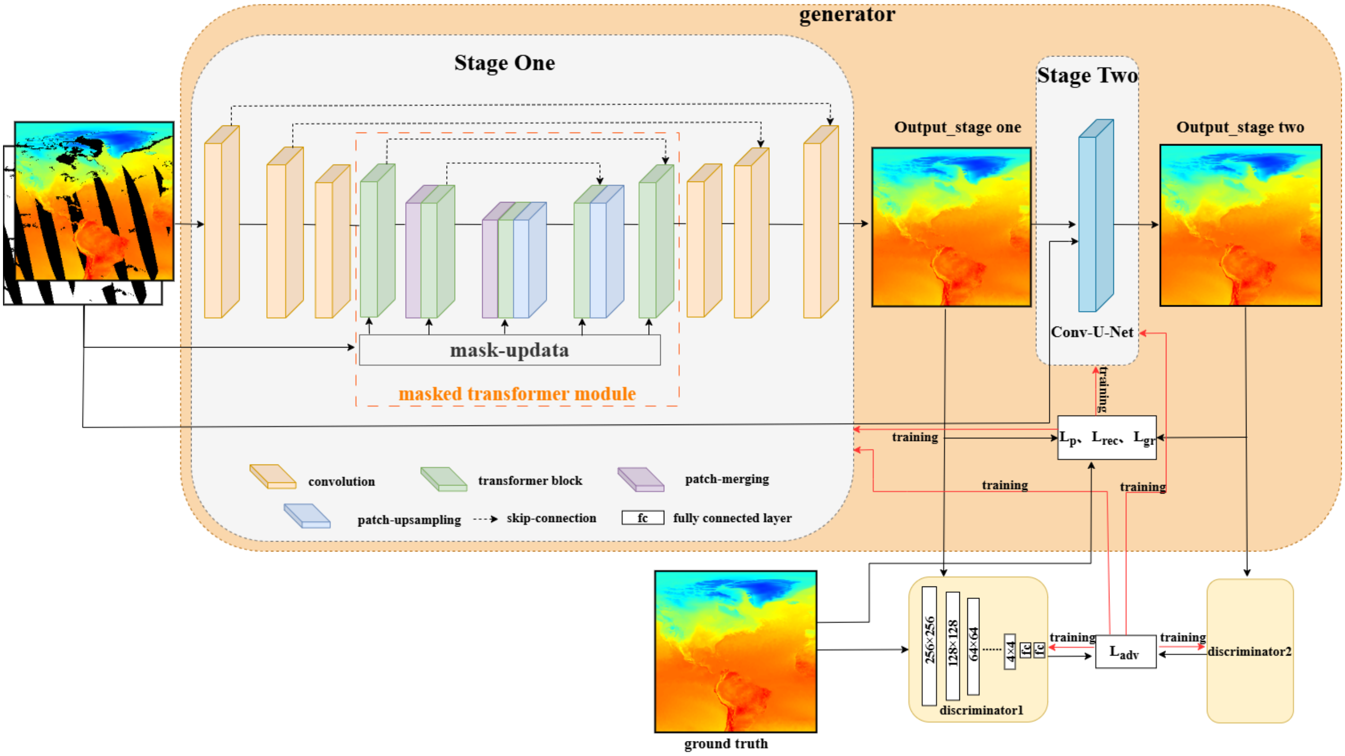


Figure 1. The proposed ESTD-Net framework. In Stage One, the convolutional module and masked transformer module are employed for feature extraction and initial reconstruction. Patch merging and patch upsampling manage downsampling and upsampling, respectively. Stage Two utilizes a convolutional U-Net structure to refine high-frequency details and ensure texture consistency.

The module consists of three convolutional layers: one for resizing the input and two for reducing the resolution. It serves two main purposes: first, to effectively capture the basic features of the masked images; second, by reducing spatial dimensions, it allows the model to focus on high-level semantic information while maintaining computational efficiency. On one hand, local induction must be integrated a priori in the initial stage of visual processing to enhance feature representation quality and model performance. On the other hand, this approach reduces resolution, significantly decreasing computational complexity and memory usage.

We incorporate a stack of convolutional layers within the convolutional module to extract tokens specifically tailored for the inpainting task. This design offers several advantages over traditional linear projection methods, such as those used in Vision Transformers (ViT) (Dosovitskiy, 2020), by effectively capturing local features and spatial relationships crucial for image reconstruction. The stacked convolutions facilitate a gradual and more effective filling of missing regions in the image, leading to the generation of more informative tokens. Furthermore, multi-scale downsampled features are efficiently passed to the decoder via shortcut connections, which enhances optimization and improves the overall reconstruction process. In contrast, models that rely solely on linear projections often introduce artifacts and struggle to effectively leverage surrounding textures for inpainting.

3.3 Masked Transformer Module

The masked transformer module processes tokens by capturing long-range dependencies between different regions of the image. It consists of five stages, each employing modified transformer blocks to effectively simulate these relationships. These blocks integrate an enhanced attention mechanism that incorporates additional dynamic masks to guide the process. This attention mechanism enables the model to focus on the most relevant regions, thereby improving its ability to address missing or incomplete areas effectively. This design is particularly well-suited for image inpainting tasks, where capturing spatial dependencies across large areas is essential. The mask further enhances the model's performance by directing its attention towards valid regions, ensuring that the reconstructed content remains consistent with the surrounding context. This approach, combined with the multi-stage architecture, significantly improves both the quality and coherence of the generated image.

3.3.1 Context Attention Module based on Mask

To efficiently manage large numbers of tokens and address the low fidelity of individual tokens, the context attention module employs dynamic masks and shifted windows. This design facilitates non-local interactions among a relevant subset of tokens only. The output from the context attention mechanism is computed as a weighted sum of the valid tokens, as follows:

$$\mathrm{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}'}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively, and $\frac{1}{\sqrt{d_k}}$ is a scale factor. The mask M' is defined as:

$$M'_{ij} = \begin{cases} 0, & \text{if token } j \text{ is valid} \\ -\tau, & \text{if token } j \text{ is invalid} \end{cases}$$

with τ set to a large integer (100 in this experiment) to suppress the impact of invalid tokens. After each attention computation, the $w \times w$ windows are shifted by $\left(\frac{w}{2}, \frac{w}{2}\right)$ pixels, enabling interactions between tokens across different windows. This mechanism facilitates better information flow and enhances the model's ability to capture long-range dependencies.

Mask Update. The mask M' evolves dynamically across layers to represent the validity of tokens, enabling the model to selectively focus on relevant regions. Initially, M' aligns with the input mask; however, it progressively adapts during each propagation step, ensuring that the model's attention remains directed toward meaningful and valid areas throughout the process. The key aspect of our approach is the adaptive propagation rule: if a window contains at least one valid token, all tokens within that window are considered valid following the attention operation. Conversely, windows devoid of valid tokens remain invalid, ensuring that attention is selectively concentrated on sparse, relevant regions where it is most needed.

As illustrated in Fig. 2, this process begins with localized validity (from (a) to (b)), progressively expanding the valid regions through successive window shifts and attention passes. This adaptive mask update scheme allows the mask to gradually cover the entire image, optimizing token propagation and enhancing the model's ability to capture contextual relationships.

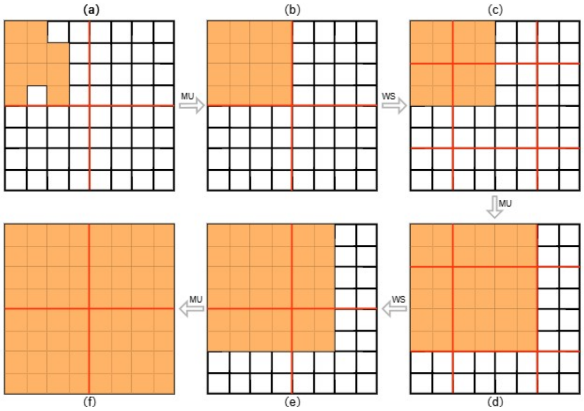


Figure 2. The mask updating process. Orange represents valid areas and white represents invalid areas. Initially, the feature map is divided into 4 × 4 regions (highlighted in red). MU represents the mask update that occurs following the attention mechanism and WS denotes the window shift operation.

235 ### 3.3.2 Modified Transformer Block

In conventional transformer architectures (Vaswani et al., 2017), each block consists of two essential components: multi-head self-attention and multi-layer perceptron. Typically, layer normalization (LN) is applied prior to each block, and a residual connection (He et al., 2016) is incorporated after each operation. However, when dealing with masks that contain large missing regions, we have observed that the standard block structure often results in unstable optimization, including instances of gradient explosion. This instability can be primarily attributed to the high proportion of invalid tokens, which are close to zero. In such scenarios, layer normalization tends to disproportionately amplify these near-zero tokens, leading to training instability. Furthermore, the residual connections in conventional transformers generally encourage the model to focus on high-frequency content, which may not be optimal for inpainting tasks that require smooth and coherent reconstructions. Given that a significant number of tokens are initially invalid, directly learning high-frequency features becomes challenging. A stable optimization process typically requires a robust low-frequency foundation, especially in GAN training, to ensure reliable convergence and avoid instability.

To address these challenges, we propose a modified transformer block specifically designed to optimize masks with missing regions. In this approach, we replace residual connections with concatenation and eliminate layer normalization altogether. As illustrated in Fig. 3, our method concatenates the output processed by context attention with the unprocessed input before passing it through a fully connected layer:

$$\mathbf{X}_{r,\ell} = \mathrm{MLP}\left[\mathrm{FC}\left(\left[\mathrm{MCA}\left(\mathbf{X}_{r,\ell-1}\right), \mathbf{X}_{r,\ell-1}\right]\right)\right] \quad (2)$$

where $X_{r,\ell}$ is the output of multi-layer perceptron at the ℓ-th block in the r-th stage. After passing through several modified transformer blocks, as shown in Fig. 3, we introduce a convolutional layer with a global residual connection. Additionally, we cancel positional embeddings in our transformer design. Previous studies (Wu et al., 2021; Xie et al., 2021) have demonstrated that 3×3 convolutions can incorporate positional information into transformers. Consequently, feature interactions are primarily driven by feature similarity, which strengthens long-range dependencies and facilitates more effective interactions within the data.

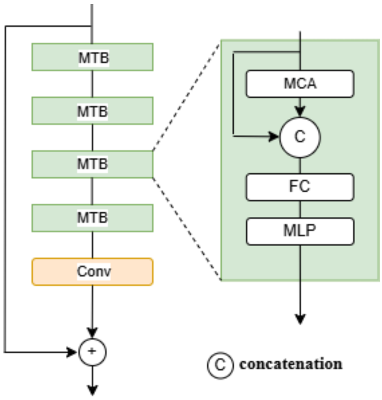


Figure 3. Illustration of a single transformer stage, where MTB denotes the modified transformer block, and MCA refers to the context attention module.

3.4 Loss Functions

To enhance the quality of the generated content, we employ adversarial loss (Mirza and Osindero, 2014) in both stages of our framework. This loss function guides the model in generating more realistic outputs by encouraging the generator to produce content that closely resembles real data, as evaluated by the discriminator. Additionally, we incorporate perceptual loss (Johnson et al., 2016) with a reduced empirical coefficient, as we have observed that this modification improves both optimization stability and effectiveness. Furthermore, perceptual loss directs the model to focus on high-level feature similarities between the generated and ground-truth images, thereby enhancing perceptual quality, particularly in the reconstruction regions. To optimize the quality of generated images, we calculate the adversarial loss as follows:

$$\mathcal{L}_G = -\mathbb{E}_{\hat{x}}[\log(D(\hat{x}))] \tag{3}$$

$$\mathcal{L}_D = -\mathbb{E}_x[\log(D(x))] - \mathbb{E}_{\hat{x}}[\log(1-D(\hat{x}))] \tag{4}$$

$$\mathcal{L}_{adv} = \mathcal{L}_D + \mathcal{L}_G + \alpha\mathcal{L}_{gp} \tag{5}$$

where x represents the real images and $\hat{x}$ (the generated images) is defined as $\hat{x} = G(x)$. The gradient penalty L_{gp} is given by $L_{\text{gp}} = \mathbb{E}_{\hat{x}} \left\|\nabla_{\hat{x}} D(\hat{x})\right\|^2$ (Gulrajani et al., 2017) , enhances the stability of the model during training and helps mitigate the risk of overfitting with $\alpha = 0.001$.

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

To reduce the difference between real images and generated images, we utilize the high-level features of the pretrained VGG-19 (Simonyan, 2014) to construct the perceptual loss:

$$\mathcal{L}_P = \sum_{i=4}^{5} \eta_i \left\| \phi_i(\hat{x}) - \phi_i(x) \right\|_1 \tag{6}$$

where $\phi_i(\cdot)$ represents the activation of layer i in a pre-trained VGG-19 network . η_i are non-negative parameters.

For the task of reconstructing global surface temperature data—which exhibits comprehensive spatial coverage but suffers from temporal sparsity and extensive missing values, we propose an improved loss function architecture based on generative adversarial networks (GANs). The primary challenge associated with this dataset stems from the coexistence of spatial continuity and temporal fragmentation, necessitating not only accurate imputation of missing temperature values but also seamless spatial and temporal transitions between observed and reconstructed regions. To address these challenges, we introduce two key modifications to the loss function, which collectively enhance the precision of temperature reconstruction while ensuring physically consistent and smooth gradients across discontinuities. In order to improve the accuracy of temperature reconstruction in missing regions, we define a weighted reconstruction loss function:

$$\mathcal{L}_{\text{rec}} = \frac{\left\| \mathbf{M} \otimes (\mathbf{x} - \hat{\mathbf{x}}) \right\|_1}{\varsigma \left\| \mathbf{M} \right\|_1} + \frac{\left\| (\mathbf{1} - \mathbf{M}) \otimes (\mathbf{x} - \hat{\mathbf{x}}) \right\|_1}{\left\| \mathbf{1} - \mathbf{M} \right\|_1} \tag{7}$$

where $\mathbf{x}$ represents ground-truth, $\hat{\mathbf{x}}$ represents generated image, and $\mathbf{M}$ represents mask. α is the weight between the known and missing data. This approach ensures that the reconstruction of missing regions closely aligns with the original data, while preserving consistency in the known areas. Similar to mask-based reconstruction losses used in image restoration tasks, this weighted loss method is particularly well-suited for surface temperature data reconstruction, effectively avoiding unnatural temperature gradients.

To improve the physical plausibility and visual coherence of inpainted data, we introduce a novel gradient regularization loss L_{gr}, designed to minimize the L1 norm of the gradient difference between the generated output $\hat{\mathbf{x}}$ and the ground truth $\mathbf{x}$. The loss is formally defined as:

$$L_{\text{gr}} = \|\nabla(x) - \nabla(\hat{x})\|_1 \tag{8}$$

By integrating gradient consistency regularization, our model learns to produce reconstructions that maintain both visual smoothness and physical fidelity with respect to the observed data. This constraint is particularly crucial in applications where gradient structures, such as temperature fronts or atmospheric transitions, play a key role in data interpretation.

The overall first stage loss functions is:

$$L = L_{adv} + \eta L_P + \lambda L_{\text{rec}} + \beta L_{\text{gr}} \tag{9}$$

where η, λ and β are non-negative parameters. The total losses in the second stage are consistent with those in the first stage.

4 Experiments

4.1 Datasets and Metrics

The reference dataset employed in this study originates from the Microwave Radiation Imager (MWRI) aboard the FengYun-3D (FY-3D) satellite. As a passive microwave sensor, MWRI offers distinct advantages for atmospheric temperature retrieval: (1) its low-frequency channels penetrate most non-precipitating clouds, and (2) the peak sensitivity of its weighting functions occurs near the surface level. These characteristics make MWRI particularly suitable for monitoring lower atmospheric temperature.

The analysis focuses on the full calendar year of 2023. FY-3D operates in an afternoon orbit with an equatorial crossing time of approximately 1400 UTC. To optimize temporal representation, we process the temperature retrievals into two global datasets centered at 0600 UTC and 1800 UTC, generated by aggregating the nearest available observations within ±3 hours of each reference time. This processing yields a twice-daily global gridded product at 0.5° × 0.5° spatial resolution.

For the purpose of method evaluation, we employed the ERA5 reanalysis surface temperature dataset as the reference baseline. This dataset was generated by first gridding the MWRI brightness temperature data from FY-3D to establish a global brightness temperature dataset with a spatial resolution of 0.5°. Subsequently, ERA5 data were interpolated to correspond spatially with the 0.5° resolution brightness temperature grid. The selection of ERA5 was motivated by two key factors: (1) its high temporal resolution, providing hourly data that allow for precise temporal alignment with MWRI observations, (2) its established reliability and widespread use in climate research and (3) interpolate the ERA5 data to the same grid points, allowing for the retrieval of true values in areas where satellite observations are missing.

The primary focus of this investigation centers on developing effective reconstruction methods for satellite orbital gaps (vacancies). Through careful spatiotemporal matching, we ensure accurate localization of these data voids for subsequent analysis and repair.

To facilitate model training and enhance intercomparability of temperature values, we applied global min-max normalization to standardize the data range. The normalization process follows:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{10}$$

Where X denotes the original temperature values, with $X_{\min}$ and $X_{\max}$ representing the global minimum and maximum temperatures across the entire dataset, respectively. This transformation maps all values to the interval [0,1], ensuring numerical stability during model optimization while preserving the relative thermal gradients.

The normalized dataset was systematically partitioned into 256×256 regions, a size selected to balance computational efficiency with sufficient spatial context for pattern recognition. From each original data file, we extracted six non-overlapping subregions through a sliding window approach, yielding a total of 4,374 analyzable units. The complete dataset was randomly partitioned into training and test sets containing 3,600 and 774 samples respectively, maintaining an approximately 4:1 ratio to ensure sufficient representation in both subsets. To comprehensively assess the reconstruction performance, we implemented four established evaluation metrics: mean absolute error (MAE), root mean square error (RMSE), peak signal-to-noise ratio

(PSNR), and structural similarity index (SSIM). Each of these metrics offers a unique perspective on the accuracy and quality of the reconstructed data.

Mean Absolute Error (MAE). MAE is a metric that quantifies the average magnitude of errors between the reconstructed data and the ground truth, without accounting for their direction. It offers a clear understanding of the average deviation of predicted values from actual values, making it particularly useful for evaluating overall reconstruction accuracy. The formula for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{11}$$

where x_i represents the ground truth values, $\hat{x}_i$ the predicted values, and n the number of samples.

Root Mean Square Error (RMSE). RMSE is similar to MAE, but it places greater emphasis on larger errors by squaring the differences before averaging. This makes RMSE more sensitive to outliers and particularly valuable when minimizing large reconstruction errors is a priority. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{12}$$

Both MAE and RMSE assess the overall reconstruction error, with RMSE placing greater emphasis on larger deviations due to its sensitivity to outliers. Together, these metrics offer complementary perspectives on the quality of the reconstruction.

Peak Signal-to-Noise Ratio (PSNR). PSNR is a widely used metric for evaluating the quality of image reconstruction. It compares the maximum possible pixel value of an image (*MAX*) to the mean squared error (MSE) between the ground truth and generated images. Higher PSNR values indicate better reconstruction quality, with less distortion. The formula for PSNR is:

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{\text{MAX}}{\text{RMSE}}\right) \tag{13}$$

Structural Similarity Index (SSIM). SSIM evaluates the perceptual quality of the reconstructed data by considering changes in structural information, luminance, and contrast. Unlike MSE and RMSE, which focus on absolute pixel differences, SSIM assesses the similarity between corresponding pixels based on their structural, brightness, and contrast characteristics. This makes SSIM a more accurate measure of image quality in terms of human visual perception. The formula for SSIM is:

$$\text{SSIM}(x,\hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{14}$$

where μ_x and $\mu_{\hat{x}}$ are the mean pixel values of the original and generated images, σ_x and $\sigma_{\hat{x}}$ are their variances, $\sigma_{x\hat{x}}$ is the covariance between the two images, and c_1 and c_2 are constants to stabilize the division when the denominator is close to zero.

These four metrics provide a comprehensive evaluation of the model's performance, capturing both pixel-level accuracy and perceptual quality. By integrating these metrics, we ensure a thorough assessment of the reconstructed images, considering both numerical precision and perceptual realism.

4.2 Implementation Details

All experiments were conducted using two NVIDIA A6000 GPUs. The model was trained on the processed ERA5 reanalysis dataset, with a batch size of 32 to optimize training efficiency. During training, we set the learning rate to 0.001. We employed the Adam optimizer, a widely used choice in deep learning, due to its effectiveness in managing sparse gradients and adapting learning rates for individual parameters. ς, η, λ and β are set as 10, 0.1,10 and 0.01, respectively.

4.3 Comparative assessment

In this section, we evaluate the effectiveness of our proposed reconstruction method, ESTD-Net, through a comparative analysis with both traditional and deep learning-based approaches. For traditional reconstruction methods, we selected a technique based on spatial information to minimize manual intervention. Specifically, we employed inverse distance-weighted interpolation(IDW)(Kilibarda et al., 2014), a simple yet effective spatial data interpolation method.

For deep learning-based reconstruction, we utilized Partial Convolutions combined with the U-Net architecture (Liu et al., 2018), which is well-suited for handling irregularly shaped missing regions in image inpainting. This combination has become a standard approach in the image processing community, particularly for inpainting tasks involving non-uniform gaps. Previous studies have demonstrated that U-Net with partial convolutions outperforms alternative methods such as PatchMatch (Barnes et al., 2009), convolutional U-Net architectures with varying null-value initializations, and extended frameworks like Content Encoders (Iizuka et al., 2017), which incorporate both global and local discriminators along with Poisson blending as a post-processing step. Additionally, Yu et al. (2018) proposed replacing post-processing with a refinement network that utilizes context attention layers. Despite these alternatives, U-Net with partial convolutions remains the preferred choice due to its superior ability to handle irregular gaps, making it particularly effective for reconstruction tasks.

The quantitative evaluation results are summarized in Table 1. From the results, we can find that our method can obtain more accurate results. To further assess the performance of our method, we conducted a qualitative analysis by visually comparing the reconstruction results of ESTD-Net with those of other reconstruction approaches. As illustrated in Figure 4, ESTD-Net demonstrates superior reconstruction capabilities, particularly in preserving the structural continuity of missing regions. Specifically, our method effectively smooths the boundaries of missing areas, mitigates artifacts commonly observed in conventional interpolation methods, and accurately reconstructs the internal spatial patterns within these regions. Furthermore, ESTD-Net exhibits a strong capacity to capture zonal gradient variations in Sea Surface Temperature (SST), ensuring consistency with large-scale oceanic temperature structures.

To further highlight the advantages of our method, we compute the absolute differences between the reconstructed results and the true values for each approach. To amplify these differences, we apply the logarithm to the absolute error plus one, where the addition of one helps avoid negative infinity values resulting from zero errors. The difference maps, presented in Figure 5, provide a detailed visualization of the reconstruction errors. Compared to traditional and deep learning-based methods, ESTD-Net significantly reduces errors along the edges of missing regions, better preserves temperature gradients, and maintains

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

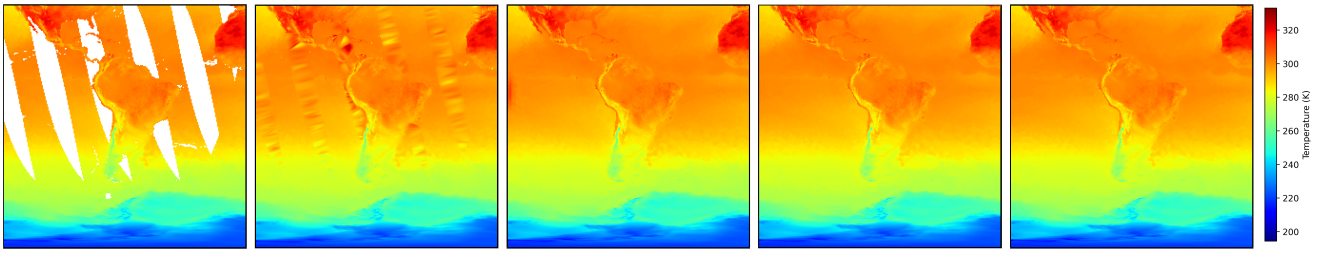


Figure 4. Reconstruction Results of Surface Temperature. From left to right, the columns display the initial incomplete data, the results of inverse distance weighting (IDW) interpolation, the results from partial convolution U-Net (Pconv U-Net), the results from our proposed method, and the ground truth. All panels utilize identical color scaling to facilitate direct visual comparison.

physically plausible spatial patterns. These improvements underscore the model's ability to leverage both spatial and temporal correlations for more accurate and reliable SST reconstructions.

Table 1. Comparison of different methods.

	MAE↓	RMSE↓	PSNR↑	SSIM↑
IDW	1.1040	2.9320	33.8333	0.9136
Pconv U-Net	0.0771	0.3223	53.4528	0.9974
ESTD-Net(ours)	**0.0522**	**0.2000**	**56.9911**	**0.9985**

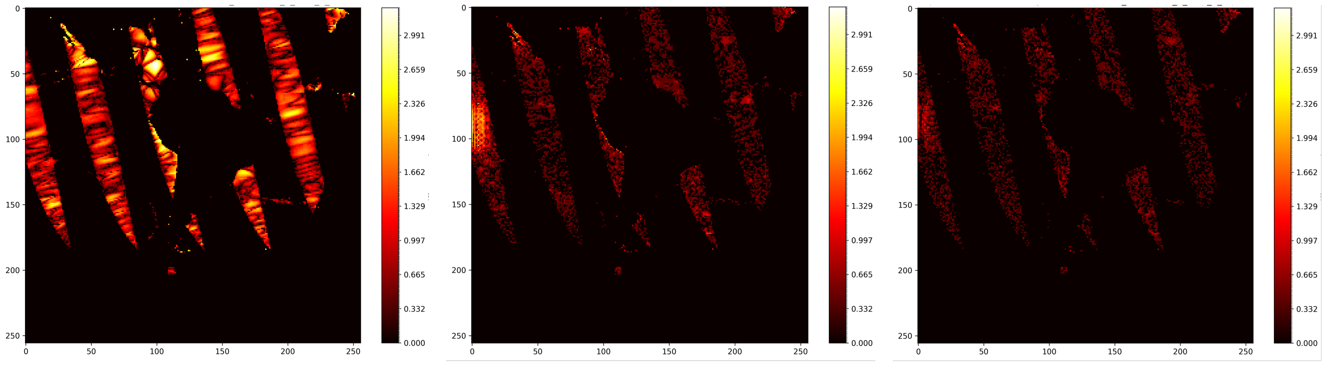


Figure 5. Comparison of Different Methods with Ground Truth. From left to right, the columns display the difference of the reconstructed data of IDW, Pconv_U-net and our method with those of the ground truth.

4.4 Ablation and analysis

In this section, we present a detailed analysis of its performance metrics, highlighting the model's superior results and elucidating the contributions of each key component. To systematically evaluate the impact of different loss function components on reconstruction performance, we conducted an ablation study focusing on the Weighted Reconstruction Loss and Gradient

Consistency Regularization. In satellite-based temperature retrieval, missing data typically arise from factors such as cloud contamination and orbital gaps. In our approach, we employ various weighting schemes for the reconstruction loss to emphasize the restoration of these missing regions. Specifically, we compare a baseline model that incorporates adversarial loss and perceptual loss against models that introduce a weighting ratio for masked (missing) to unmasked (observed) areas in the reconstruction loss. This weighting ensures a stronger emphasis on missing regions, which is crucial for effectively filling large gaps in Earth Surface Temperature data.

Additionally, we examine the effect of incorporating Gradient Consistency Regularization, which enforces smooth transitions and structural coherence in the reconstructed regions. The results, summarized in Table 2, demonstrate the effectiveness of the proposed loss terms. Compared to the baseline model, which achieves a Mean Absolute Error (MAE) of 0.0669, incorporating the Weighted Reconstruction Loss significantly reduces MAE to 0.0553, representing a 17.3% reduction. Similarly, the Root Mean Square Error (RMSE) decreases from 0.2658 to 0.2177, and Peak Signal-to-Noise Ratio (PSNR) improves from 55.1125 dB to 56.4303 dB, indicating enhanced reconstruction accuracy. The inclusion of Gradient Consistency Regularization further refines these results: the full model achieves an MAE of 0.0522 and an RMSE of 0.2000, corresponding to overall reductions of 22.0% and 24.7%, respectively, relative to the baseline model. These improvements suggest that our approach not only minimizes pixel-level errors but also enhances the physical consistency of Sea Surface Temperature (SST) patterns by better preserving zonal temperature gradients and reducing discontinuities at the boundaries of missing regions.

Table 2. Ablation study on the effect of loss functions.

	MAE↓	RMSE↓	PSNR↑	SSIM↑
$L_{\mathrm{adv}} + L_{\mathrm{P}}$	0.0669	0.2658	55.1125	0.9979
$L_{\mathrm{adv}} + L_{\mathrm{P}} + L_{\mathrm{gp}}$	0.0641	0.2473	55.4460	0.9980
$L_{\mathrm{adv}} + L_{\mathrm{P}} + L_{\mathrm{rec}}$	0.0553	0.2177	56.4303	0.9984
$L_{\mathrm{adv}} + L_{\mathrm{P}} + L_{\mathrm{rec}} + L_{\mathrm{gp}}$	0.0522	0.2000	56.9911	0.9985

As shown in Table 2, the combination of adversarial loss, perceptual loss, and weighted reconstruction loss achieves the best performance among the various configurations. This combination results in the lowest MAE and RMSE, while also yielding the highest PSNR and SSIM. These results underscore the significance of the weighted reconstruction loss in guiding the model to accurately reconstruct missing regions, thereby significantly enhancing overall performance. The introduction of Gradient Consistency Regularization, which minimizes gradient discrepancies between the ground truth and reconstructed images, further improves perceptual quality, as evidenced by the increased PSNR and SSIM scores. This regularization term helps maintain smooth transitions and structural integrity within the reconstructed data. Notably, its effect is most pronounced when combined with other loss components, highlighting the synergistic benefits of this multi-loss framework.

Overall, the integration of adversarial loss, perceptual loss, weighted reconstruction loss, and Gradient Consistency Regularization results in a well-balanced performance that enhances both pixel-level accuracy and perceptual quality in the reconstructed EST data. These findings demonstrate the robustness of our model in handling large-scale missing data while effectively capturing complex spatial and temporal dependencies within the EST data.

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

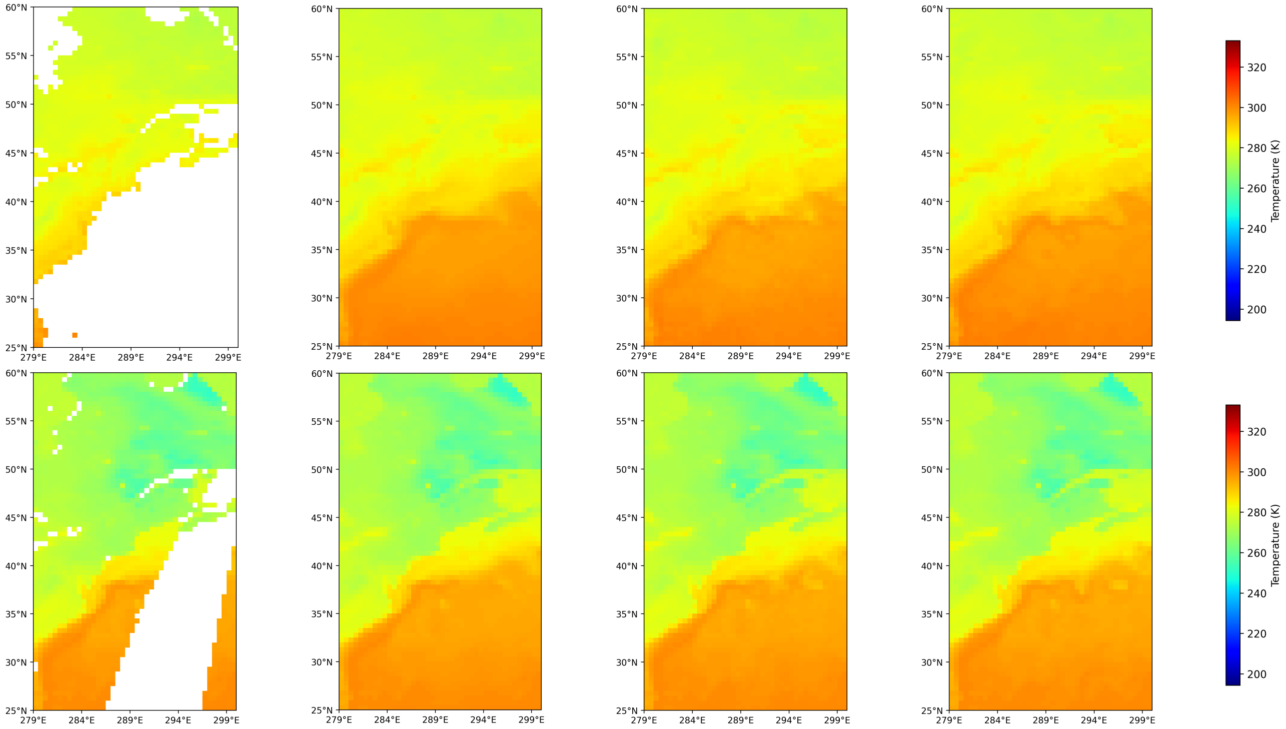


Figure 6. Comparison of Gradient Settings on different missing ratios. From left to right, the columns display the initial data, the reconstruction result without gradient settings, the reconstruction result with gradient settings applied, and the ground truth. The white regions in the original data indicate missing values.

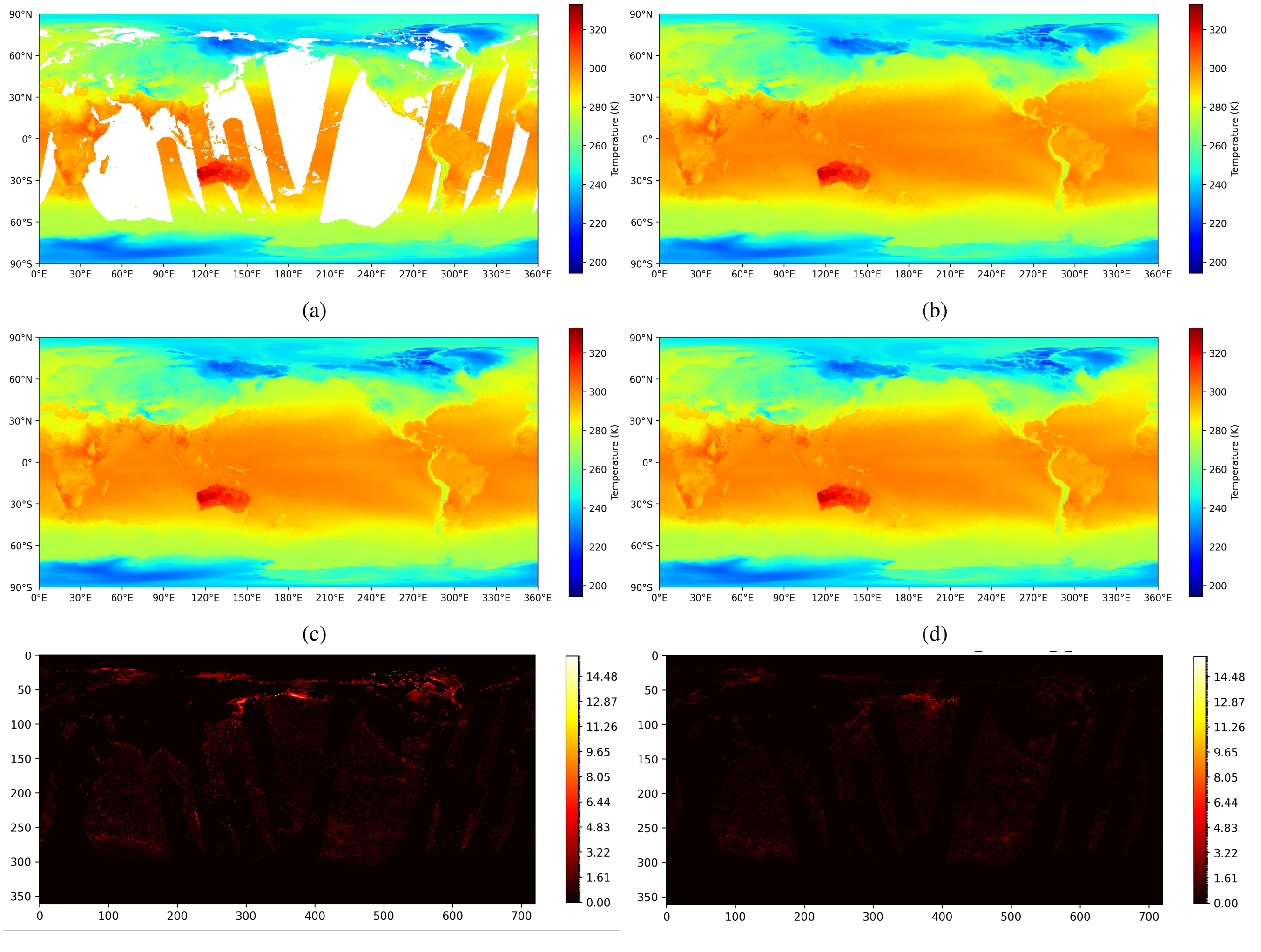
The experimental results underscore the critical roles of both weighted reconstruction loss and gradient consistency regularization in enhancing performance in the surface temperature image inpainting task. The weighted reconstruction loss, by prioritizing errors in the unmasked regions, is essential for improving the overall quality of the reconstruction. Meanwhile, gradient consistency regularization enhances the structural coherence of the generated images, thereby improving both the overall quality and structural consistency of the inpainting results.

Gradient setting comparison. To evaluate the effectiveness of gradient consistency regularization in preserving physically meaningful structures, we conducted a focused assessment over the eastern coastline of North America and the adjacent Atlantic Ocean, a region characterized by sharp land-sea thermal contrasts. This area spans latitudes from 25°N to 60°N and longitudes from 279°E to 300°E, exhibiting pronounced temperature gradients at coastal boundaries, thereby making it an ideal testbed for evaluating boundary reconstruction performance.

As illustrated in Figure 6, the inclusion of the gradient consistency regularization term significantly enhances the model's ability to capture fine-scale temperature variations. The white regions in the original data indicate missing values. Specifically, this term promotes the alignment of gradient structures between the generated output and the ground truth, resulting in sharper transitions and improved preservation of boundary features. In comparison to the reconstruction result without this setting

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

(Column 2), the model incorporating gradient consistency regularization (Column 3) demonstrates clearer and more continuous land-sea edges, more accurately reflecting the ground truth (Column 4). This enhancement is particularly evident in the recovery of temperature fronts and the retention of cross-boundary gradients, which are often smoothed out or distorted in models lacking explicit gradient guidance. Such improvements illustrate that this regularization not only enhances the visual coherence of the restored image but also contributes to the physical plausibility of the reconstructed temperature field, an essential aspect for downstream geoscientific analyses where maintaining spatial gradient integrity is critical.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7. Reconstructed data of Stage Two steps. (a) Initial data. (b) Ground truth. (c) Result of the first stage. (d) Result of the second stage. (e) Difference Between the reconstructed data of the first stage and the ground truth. (f) Difference Between the reconstructed data of the second stage and the ground truth.

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025

The role of the Second Stage. The interpolated data generated in the first stage is produced by applying adaptive weights to valid pixels within local windows. While this method can restore the overall temperature field, it often results in imprecise outputs in regions with complex spatial variations, leading to localized artifacts and inconsistent transitions. To address these issues, we introduce a convolutional U-Net in the second stage to autoregressively refine the initial results. This refinement process aims to enhance local continuity and correct structural inconsistencies.

As illustrated in Figure 7, a direct visual comparison of the global outputs from the first and second stages (Figure 7(c) and (d)) reveals only subtle differences. However, the absolute difference maps relative to the ground truth (Figure 7(e) and (f)) more clearly highlight the improvements achieved by the second stage. In Figure 7(e), the absolute error between the first-stage output and the ground truth displays numerous scattered high-magnitude differences, reflecting the limitations of local window-based interpolation in accurately capturing detailed spatial structures.

In contrast, the absolute difference shown in Figure 7(f) is notably smoother and less concentrated, indicating that the second-stage U-Net effectively reduces local anomalies and refines spatial transitions. This improvement is particularly evident in areas characterized by sharp temperature gradients or complex patterns, where the second-stage refinement yields results that align more closely with the physical characteristics of the temperature field.

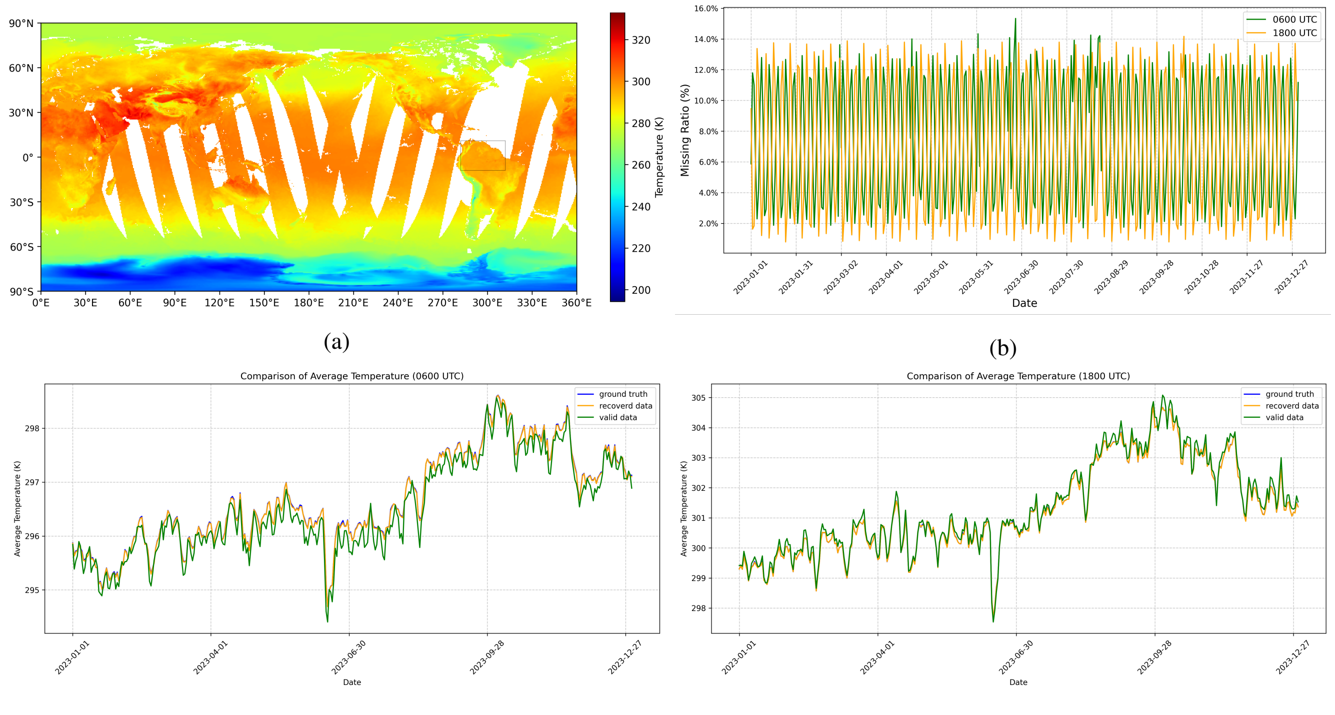
These enhancements demonstrate that the convolutional U-Net plays a crucial role in improving reconstruction quality by minimizing abrupt local deviations and producing smoother, more physically plausible outputs.

Temporal Stability Verification of the Reconstruction Method. To verify the temporal stability of our proposed reconstruction method over an extended period, we selected a representative region in the eastern Pacific, spanning longitudes 282°E to 312°E and latitudes 10°N to 10°S. This area was chosen for its meteorological significance and diverse surface types, including oceanic zones, coastal regions, and land areas influenced by large-scale climatic phenomena. As illustrated in Figure 8(a), the selected region is clearly marked on a global temperature map, highlighting its spatial context.

Given the strong El Niño event observed in 2023, which contributed to notable temperature anomalies—this region serves as a valuable case for assessing the consistency and robustness of our reconstruction method throughout the year. We evaluated the performance of our method by comparing the average temperature of valid pixels before reconstruction with the reconstructed results and the actual average temperature in the selected region across the entire year of 2023.

Figure 8(b) presents the missing data rates at 0600 UTC and 1800 UTC, indicating that the proportion of missing data varies throughout the year, with noticeable differences between these two time points. Despite these variations, our method demonstrates stable and accurate reconstruction performance. As shown in Figures 8(c) and (d), the average temperatures of the reconstructed data remain closely aligned with actual values for both time periods, significantly improving upon averages computed from valid data alone. This indicates that our method can effectively compensate for missing observations while preserving temporal consistency over extended periods and under varying conditions of missing data.

https://doi.org/10.5194/egusphere-2025-1980
Preprint. Discussion started: 27 May 2025



(a)

(b)

(c)

(d)

Figure 8. Comparison of average temperatures before and after reconstruction with actual temperatures in the selected eastern Pacific region. (a) Global spatial distribution of Earth Surface Temperature, with the chosen region (282°E–312°E, 10°N–10°S) outlined in black. (b) Missing data rates over the entire year of 2023 at 0600 UTC and 1800 UTC in the selected region. (c-d) Comparison of average temperatures before and after reconstruction with the actual temperatures at 0600 UTC and 1800 UTC across 2023, respectvely. The blue line represents the average temperature of the actual data, the orange line shows the reconstructed results, and the green line indicates the average temperature of valid (non-missing) data before reconstruction.

5 Conclusions

This paper presents ESTD-Net, a novel network architecture specifically designed for surface temperature data inpainting. ESTD-Net employs an enhanced multi-head context attention mechanism and a modified transformer block to effectively capture long-range pixel dependencies. These architectural advancements improve the model's ability to focus on boundary regions, resulting in more accurate reconstructions. Furthermore, we leverage the local texture refinement capabilities of convolutional neural networks (CNNs) by incorporating a dedicated convolutional U-Net, which enhances high-frequency details in the generated images. The integration of weighted reconstruction loss and gradient consistency regularization significantly elevates restoration quality, ensuring that the inpainted images closely align with ground truth in terms of both structural consistency and pixel-level accuracy. Collectively, these innovations lead to substantial improvements in the overall quality of surface temperature image restoration.

While the results achieved using simulated data from the ERA5 reanalysis dataset demonstrate promising outcomes, real-world data introduces additional complexities. In practical applications, satellite observations are often incomplete, with certain regions consistently missing data due to factors such as cloud cover or sensor limitations. This inherent challenge results in a scarcity of fully complete surface temperature data. To address this issue, a practical approach involves extracting data from regions where observations are intact and artificially introducing gaps to simulate missing data for testing and evaluation purposes. This simulated dataset serves as a proxy for real-world conditions, enabling us to assess the model's robustness and performance within a more practical context.

By employing this method, we can further validate and fine-tune the reconstruction technique, ensuring its effectiveness in handling incomplete surface temperature data encountered in real-world applications. This approach provides a viable pathway for bridging the gap between idealized simulation-based testing and the complexities of real-world data, ensuring that ESTD-Net remains applicable across a wide range of environmental and climate research contexts.

Code and data availability. The source codes are available at https://doi.org/10.5281/zenodo.15273464 (Zhang, 2025). All data used in this study are publicly available. The Microwave Radiation Imager (MWRI) data aboard the FengYun-3D (FY-3D) satellite are publicly accessible via the China Meteorological Administration's National Satellite Meteorological Center portal (http://data.nsmc.org.cn/). ERA5 data can be downloaded from the Copernicus Climate Data Store (https://doi.org/10.24381/cds.adbb2d47 (Hersbach et al., 2020)).

Author contributions. MHZ and YJC designed the study. MHZ performed the analyses and wrote the paper, with contributions from all co-authors.

Competing interests. The Authors declare no competing interest.

Financial support. This work was supported by the National Social Science Foundation of China (Grant No. 42375004).

References

Atlas, D., Srivastava, R., and Sekhon, R. S.: Doppler radar characteristics of precipitation at vertical incidence, Reviews of Geophysics, 11, 1–35, https://doi.org/10.1029/RG011i001p00001, 1973.

Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., and Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels, IEEE transactions on image processing, 10, 1200–1211, https://doi.org/10.1109/83.935036, 2001.

Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B.: PatchMatch: A randomized correspondence algorithm for structural image editing, ACM Trans. Graph., 28, 24, https://doi.org/10.1145/1531326.1531330, 2009.

Berman, D., Avidan, S., et al.: Non-local image dehazing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1674–1682, https://doi.org/10.1109/CVPR.2016.185, 2016.

Criminisi, A., Pérez, P., and Toyama, K.: Region filling and object removal by exemplar-based image inpainting, IEEE Transactions on image processing, 13, 1200–1212, https://doi.org/10.1109/TIP.2004.833105, 2004.

Cui, J., Zhang, M., Song, D., Shan, X., and Wang, B.: MODIS land surface temperature product reconstruction based on the SSA-BiLSTM model, Remote Sensing, 14, 958, https://doi.org/10.3390/rs14040958, 2022.

Deng, Y., Hui, S., Zhou, S., Meng, D., and Wang, J.: Learning contextual transformer network for image inpainting, in: Proceedings of the 29th ACM international conference on multimedia, pp. 2529–2538, https://doi.org/10.1145/3474085.3475426, 2021.

Deo, R. C. and Şahin, M.: Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland, Renewable and Sustainable Energy Reviews, 72, 828–848, https://doi.org/10.1016/j.rser.2017.01.114, 2017.

Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

Dowd, P. A. and Pardo-Igúzquiza, E.: The many forms of co-kriging: A diversity of multivariate spatial estimators, Mathematical Geosciences, 56, 387–413, https://doi.org/10.1007/s11004-023-10104-7, 2024.

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., and Akbari, Y.: Image inpainting: A review, Neural Processing Letters, 51, 2007–2028, https://doi.org/10.1007/s11063-019-10163-0, 2020.

Fleit, G.: Windowed anisotropic local inverse distance-weighted (WALID) interpolation method for riverbed mapping, Acta Geophysica, pp. 1–15, https://doi.org/10.1007/s11600-024-01510-4, 2024.

Geiss, A. and Hardin, J. C.: Inpainting radar missing data regions with deep learning, Atmospheric Measurement Techniques, 14, 7729–7747, https://doi.org/10.5194/amt-14-7729-2021, 2021.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C.: Improved training of wasserstein gans, arXiv preprint arXiv:1704.00028, 2017.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999–2049, https://doi.org/10.24381/cds.adbb2d47, 2020.

Hore, A. and Ziou, D.: Image quality metrics: PSNR vs. SSIM, in: 2010 20th international conference on pattern recognition, pp. 2366–2369, IEEE, https://doi.org/10.1109/ICPR.2010.579, 2010.

Iizuka, S., Simo-Serra, E., and Ishikawa, H.: Globally and locally consistent image completion, ACM Transactions on Graphics (ToG), 36, 1–14, https://doi.org/10.1145/3072959.3073659, 2017.

Johnson, J., Alahi, A., and Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 694–711, Springer, https://doi.org/10.1007/978-3-319-46475-6_43, 2016.

Kilibarda, M., Hengl, T., Heuvelink, G. B., Gräler, B., Pebesma, E., Perčec Tadić, M., and Bajat, B.: Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution, Journal of Geophysical Research: Atmospheres, 119, 2294–2313, https://doi.org/10.1002/2013JD020803, 2014.

King, M. D., Platnick, S., Menzel, W. P., Ackerman, S. A., and Hubanks, P. A.: Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites, IEEE transactions on geoscience and remote sensing, 51, 3826–3852, https://doi.org/10.1109/TGRS.2012.2227333, 2013.

Lei Ba, J., Kiros, J. R., and Hinton, G. E.: Layer normalization, arXiv preprint arXiv:1607.06450, pp. arXiv–1607, 2016.

Lengfeld, K., Kirstetter, P.-E., Fowler, H. J., Yu, J., Becker, A., Flamig, Z., and Gourley, J.: Use of radar data for characterizing extreme precipitation at fine scales and short durations, Environmental Research Letters, 15, 085 003, https://doi.org/10.1088/1748-9326/ab98b4, 2020.

Lepetit, P., Ly, C., Barthès, L., Mallet, C., Viltard, N., Lemaitre, Y., and Rottner, L.: Using deep learning for restoration of precipitation echoes in radar data, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14, https://doi.org/10.1109/TGRS.2021.3052582, 2021.

Li, J. and Heap, A. D.: Spatial interpolation methods applied in the environmental sciences: A review, Environmental Modelling & Software, 53, 173–189, https://doi.org/10.1016/j.envsoft.2013.12.008, 2014.

Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., and Jia, J.: Mat: Mask-aware transformer for large hole image inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10 758–10 768, https://doi.org/10.1109/CVPR52688.2022.01049, 2022.

Li, X., Shen, H., Zhang, L., Zhang, H., Yuan, Q., and Yang, G.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning, IEEE Transactions on Geoscience and Remote Sensing, 52, 7086–7098, https://doi.org/10.1109/TGRS.2014.2307354, 2014.

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B.: Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European conference on computer vision (ECCV), pp. 85–100, https://doi.org/10.1007/978-3-030-01252-6_6, 2018.

Liu, H., Jiang, B., Song, Y., Huang, W., and Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 725–741, Springer, https://doi.org/10.1007/978-3-030-58536-5_43, 2020.

Liu, Z., Wu, P., Duan, S., Zhan, W., Ma, X., and Wu, Y.: Spatiotemporal reconstruction of land surface temperature derived from fengyun geostationary satellite data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 4531–4543, https://doi.org/10.1109/JSTARS.2017.2716376, 2017.

Malek, S., Melgani, F., Bazi, Y., and Alajlan, N.: Reconstructing cloud-contaminated multispectral images with contextualized autoencoder neural networks, IEEE Transactions on Geoscience and Remote Sensing, 56, 2270–2282, https://doi.org/10.1109/TGRS.2017.2777886, 2017.

Mirza, M. and Osindero, S.: Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784, 2014.

Mohanasundaram, S., Baghel, T., Thakur, V., Udmale, P., and Shrestha, S.: Reconstructing NDVI and land surface temperature for cloud cover pixels of Landsat-8 images for assessing vegetation health index in the Northeast region of Thailand, Environmental monitoring and assessment, 195, 211, https://doi.org/10.1007/s10661-022-10802-5, 2023.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A.: Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544, https://doi.org/10.1109/CVPR.2016.278, 2016.

Petrovska, B., Zdravevski, E., Lameski, P., Corizzo, R., Štajduhar, I., and Lerga, J.: Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification, Sensors, 20, 3906, https://doi.org/10.3390/s20143906, 2020.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pp. 234–241, Springer International Publishing, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Scharlemann, J. P., Benz, D., Hay, S. I., Purse, B. V., Tatem, A. J., Wint, G. W., and Rogers, D. J.: Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data, PloS one, 3, e1408, https://doi.org/10.1371/journal.pone.0001408, 2008.

Simonyan, K.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

Tan, S. and Chen, H.: A conditional generative adversarial network for weather radar beam blockage correction, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–14, https://doi.org/10.1109/TGRS.2023.3286181, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I.: Attention is all you need, arXiv preprint arXiv:1706.03762, 2017.

Wan, Z., Zhang, J., Chen, D., and Liao, J.: High-fidelity pluralistic image completion with transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4692–4701, https://doi.org/10.1109/ICCV48922.2021.00465, 2021.

Wang, P., Zheng, W., Chen, T., and Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice, arXiv preprint arXiv:2203.05962, 2022.

Wang, X., Girshick, R., Gupta, A., and He, K.: Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803, https://doi.org/10.1109/CVPR.2018.00813, 2018.

Wang, Y., Karimi, H. A., and Jia, X.: Reconstruction of Continuous High-Resolution Sea Surface Temperature Data Using Time-Aware Implicit Neural Representation, Remote Sensing, 15, 5646, https://doi.org/10.3390/rs15245646, 2023.

Weiss, D. J., Mappin, B., Dalrymple, U., Bhatt, S., Cameron, E., Hay, S. I., and Gething, P. W.: Re-examining environmental correlates of Plasmodium falciparum malaria endemicity: a data-intensive variable selection approach, Malaria journal, 14, 1–18, https://doi.org/10.1186/s12936-015-0574-x, 2015.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L.: Cvt: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 22–31, https://doi.org/10.1109/ICCV48922.2021.00009, 2021.

Xie, C., Liu, S., Li, C., Cheng, M.-M., Zuo, W., Liu, X., Wen, S., and Ding, E.: Image inpainting with learnable bidirectional attention maps, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8858–8867, https://doi.org/10.1109/ICCV.2019.00895, 2019.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers, arXiv preprint arXiv:2105.15203, 2021.

Xie, L., Zhao, Q., Huo, J., and Cheng, G.: A ground penetrating radar data reconstruction method based on generation networks, in: 2020 IEEE Radar Conference (RadarConf20), pp. 1–4, IEEE, https://doi.org/10.1109/RadarConf2043947.2020.9266648, 2020.

Xu, S., Wang, D., Liang, S., Liu, Y., and Jia, A.: Assessment of gridded datasets of various near surface temperature variables over Heihe River Basin: Uncertainties, spatial heterogeneity and clear-sky bias, International Journal of Applied Earth Observation and Geoinformation, 120, 103 347, https://doi.org/10.1016/j.jag.2023.103347, 2023.

Yan, Z., Li, X., Li, M., Zuo, W., and Shan, S.: Shift-net: Image inpainting via deep feature rearrangement, in: Proceedings of the European conference on computer vision (ECCV), pp. 1–17, https://doi.org/10.1007/978-3-030-01264-9_1, 2018.

Yi, Z., Tang, Q., Azizi, S., Jang, D., and Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7508–7517, https://doi.org/10.1109/CVPR42600.2020.00753, 2020.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S.: Generative image inpainting with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5505–5514, https://doi.org/10.1109/CVPR.2018.00577, 2018.

Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., and Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 69–78, https://doi.org/10.1145/3474085.3475436, 2021.

Zeng, Y., Fu, J., Chao, H., and Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1486–1494, https://doi.org/10.1109/CVPR.2019.00158, 2019.

Zhang, G., Xiao, X., Dong, J., Kou, W., Jin, C., Qin, Y., Zhou, Y., Wang, J., Menarguez, M. A., and Biradar, C.: Mapping paddy rice planting areas through time series analysis of MODIS land surface temperature and vegetation index data, ISPRS Journal of Photogrammetry and Remote Sensing, 106, 157–171, https://doi.org/10.1016/j.isprsjprs.2015.05.011, 2015.

Zhang, M.: Attention-Driven and Multi-Scale Feature Integrated Approach for Earth Surface Temperature Data Reconstruction, Zenodo [code], https://doi.org/10.5281/zenodo.15273464, 2025.

Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I., and Xu, Y.: Large scale image completion via co-modulated generative adversarial networks, arXiv preprint arXiv:2103.10428, 2021.

Zhou, J., Liang, S., Cheng, J., Wang, Y., and Ma, J.: The GLASS land surface temperature product, IEEE journal of selected topics in applied earth observations and remote sensing, 12, 493–507, https://doi.org/10.1109/JSTARS.2018.2870130, 2018.