

Response to Reviewers

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

Authors: Minghui Zhang, Yunjie Chen, Fan Yang, Zhengkun Qin

The authors sincerely thank the editors and reviewers for their valuable comments and suggestions. Therefore, we have provided supplementary material to respond to the editors' and reviewers' opinions.

Detailed Response to Reviewers

Reviewer #1:

This manuscript introduces a deep learning-based data restoration method, ESTD-Net, aimed at recovering surface temperature from high-resolution observations. This method is built upon advanced models and restores temperature data with pixel-level accuracy, enhancing the quality of the reconstructed images. The manuscript writing motivation is clear, and the experimental results show improvements. However, there are still many issues that need to be further revised.

Major Comments:

1. The ESTD-Net proposed in the manuscript effectively solves the time complexity of traditional Transformers in image quality restoration tasks, but there is no experimental comparison with the training time of Transformer models presented in the article.

--- Thanks for this suggestion. We agree that empirical evidence is essential to support our claim regarding training efficiency. In response, we have added a direct comparison of the total training time and model parameters between ESTD-Net and the ViT-baseline in the Supplementary Material section 1 "Training Time Comparison". Specifically, the results show that:

1. ESTD-Net requires 4 days and 14 hours of training with 95.8M parameters.

2. ViT-baseline requires 5 days and 1 hour with 102M parameters.

These experiments were conducted under identical training settings (same dataset, GPU type, and optimizer configuration). This supports our claim that ESTD-Net is more efficient than conventional Transformer models in terms of both training time and model complexity. We hope this addition clarifies our contribution regarding time complexity reduction.

2. The manuscript presents the superiority of ESTD-Net in the task of recovering high-resolution temperature data, but lacks experimental comparisons with more recent extreme models for data recovery tasks, which cannot fully demonstrate the effectiveness of the proposed method. Additional evidence can be provided by comparing experimental results with more advanced models.

--- Thanks for this suggestion. In the Supplementary Material section 2 "Model Comparison", we added a new evaluation of ESTD-Net against two recent data recovery models: MAT^[1] and Palette^[2]. We hope these experimental additions address your concerns.

3. The overall model diagram and description are not clear enough. It is unclear how the

Conv-U-Net in the second stage further improves the reconstruction accuracy. Additionally, the structure of the discriminator is unclear. Is it composed only of fully connected layers?

--- Thanks for this suggestion.

1.Stage-II Conv-U-Net: The Conv-U-Net in Stage II is a standard U-Net architecture without modification. Its purpose is to refine the coarse output from Stage I in an autoregressive manner, enhancing local texture continuity and smoothing boundaries, which is critical for temperature field restoration. We have also added an ablation study in the Supplementary Material section 5" Ablation and analysis " to demonstrate its impact. Removing this stage causes a significant increase in MAE and RMSE, confirming its effectiveness.

2.Discriminator Architecture: The discriminator is not composed of fully connected layers only. Instead, it adopts a PatchGAN-like architecture consisting of stacked convolutional blocks, followed by two fully connected layers for final prediction. This structure enables local realism evaluation and helps the generator produce sharper outputs.

We hope these clarifications and experimental additions address your concerns.

4.The overall workflow of training and inference is not very intuitive.

--- Thanks for this suggestion. To improve clarity, we have added a concise step-wise description of our workflow (see Supplementary Material section 3" The overall workflow of training and inference "), explicitly outlining the inputs, two-stage reconstruction process, and the difference between training and inference. Notably, since our task is spatiotemporal data imputation (rather than image inpainting), we have avoided visual/image terminology and clarified that our inputs are gridded temperature fields.

5.The manuscript directly reconstructs the processed brightness temperature data. Is the brightness temperature first processed into surface temperature and then verified with the surface temperature of ERA5? The description and rationality of the dataset need to be further explained.

--- Thanks for this suggestion. FY-3D's MWRI provides near-real-time surface temperature retrievals, yet the sensor's narrow swath leaves large gaps between successive orbits and frequent data voids under cloudy conditions. To generate a gap-free surface temperature dataset, we have developed a reconstruction method that infills these missing areas. Because satellite retrievals lack in-situ truth, we construct synthetic MWRI data for evaluation: starting from ERA5 surface temperature, we retain only those grid points that coincide with valid MWRI retrievals and mask the remainder, thereby yielding a benchmark dataset where ERA5 values serve as the "truth" in the artificially created gaps. This setup enables rigorous assessment of the reconstruction algorithm's accuracy.

Minor Comments:

1.Regarding the mask based contextual attention module proposed in section 3.3.1 of

this manuscript, although a detailed calculation process is introduced, no corresponding figure is attached for explanation, which will reduce the readability of the paper.

--- Thank you for the suggestion. To improve clarity, we have added a step-by-step pseudocode description (see Supplementary Material section 4" The contextual attention module based on mask") to illustrate how the contextual attention mechanism operates in our spatial data imputation setting. The pseudocode clearly describes how missing values are imputed by attending to observed regions within a local spatial window. This representation serves the purpose of a figure while maintaining technical precision. We hope this addition improves the interpretability of our method.

2.This manuscript validated the impact of various loss factors in the loss function on model performance in the ablation experiment section, but it seems that there were no confirmatory experiments on the impact of contextual attention and other modules on the model, which may lead to ambiguity in the study of model architecture.

--- Thank you for the suggestion. We have now added an ablation study specifically evaluating the mask-based contextual attention module and the Stage-II Conv-U-Net. The new results are presented in the Supplementary Material section 5" Ablation and analysis ". Removing contextual attention increases MAE from 0.0522 to 0.0717 and reduces PSNR from 56.9911 dB to 54.2463 dB. Omitting the second-stage Conv-U-Net further degrades performance (MAE 0.1014, PSNR 51.2518 dB). These confirmatory experiments substantiate the critical contributions of both modules to ESTD-Net's reconstruction quality.

Reviewer #2:

This manuscript presents ESTD-Net, a two-stage hybrid architecture for inpainting missing Earth Surface Temperature (EST) fields derived from MWRI/FY-3D satellite data. The first stage employs a modified Transformer ("boundary-aware" multi-head context attention, dynamic masks, no LayerNorm, concatenated residuals) to perform global reconstruction. The second stage refines outputs via a convolutional U-Net. Losses include a Weighted Reconstruction Loss, Gradient Consistency Regularization, and an adversarial GAN loss. Experiments on ERA5-simulated gaps compare ESTD-Net against inverse-distance weighting and a partial-conv U-Net, reporting improvements in MAE, RMSE, PSNR, and SSIM.

General Comments:

1. High-Frequency Refinement by U-Net

Why do you claim that the convolutional U-Net in ESTD-Net can "refine" high-frequency details? What theoretical basis supports this? Could other network architectures achieve similar refinement? Which component specifically drives the refinement? Moreover, the U-Net design here is very crude—how should it be improved?

--- Thank you for the suggestion. The Stage II's Conv-U-Net as a self-regressive refinement module whose primary function is to extract common residual patterns across the dataset—smoothing out minor, localized artifacts from the coarse Transformer output and preventing over-fitting, rather than generating new

high-frequency textures. Its encoder–decoder structure with skip connections ensures that global structures are preserved while correcting small localization errors introduced in Stage I. Although alternative lightweight architectures (e.g., simple residual CNN blocks) could perform similar smoothing, our shallow U-Net was chosen for its simplicity and minimal overhead. As demonstrated by our added ablation study (see Supplementary Material section 5 "Ablation and analysis"), omitting this stage raises MAE from 0.0522 to 0.1014 and lowers PSNR from 56.9911 dB to 51.2518 dB, confirming its modest but valuable contribution.

2. Missing Diffusion-Model Comparison

The absence of any diffusion-based baseline is a serious gap. By 2025, diffusion models aren't just "nice to have," they've become the gold standard for high-fidelity inpainting. Showing that your hybrid Transformer-GAN outperforms a 2018 CNN is only a sanity check; it tells us nothing about where ESTD-Net sits relative to the true state of the art. I would strongly recommend benchmarking against at least one modern diffusion inpainting model

e.g., RePaint in CVPR 2022 (see <https://arxiv.org/abs/2201.09865>) or Palette in SIGGRAPH 2022 (see <https://arxiv.org/abs/2111.05826>).

You could also consult the state-of-the-art leaderboard for image inpainting (see <https://paperswithcode.com/task/image-inpainting>).

--- Thank you for the suggestion. In the Supplementary Material section 2 "Model Comparison", we have included experimental comparisons with Palette^[2], a state-of-the-art diffusion-based inpainting model, as well as MAT^[1], which similarly focuses on masked reconstruction.

Notably, our task is not traditional image inpainting, but rather the imputation of gridded Earth surface temperature fields. These fields differ from natural images in that they lack complex texture or semantic edges. This makes diffusion-based priors (which rely on natural image statistics) less effective.

Our results confirm this: ESTD-Net outperforms both Palette and MAT across all major metrics (MAE, RMSE, PSNR, SSIM). We believe this supports the suitability of our hybrid Transformer-U-Net-GAN architecture for meteorological data imputation tasks.

3. Limited Baselines and Ablation

The set of baselines is very limited, which makes it hard to highlight the proposed method's advantages. Additionally, there is no comprehensive ablation study or stability analysis, and the choice of model hyperparameters is not discussed.

--- Thank you for the suggestion. To address this issue, we have provided detailed experimental results in the Supplementary Material, including:

1. Baseline comparison: We compared the proposed ESTD-Net with two additional representative methods, namely MAT and Palette. The results (Table 2 and Figure 1) show that our model consistently outperforms these baseline models in terms of mean absolute error (MAE), root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM).
2. Ablation study: We conducted ablation experiments to evaluate the role of key

modules (context attention mechanism and the second-stage U-network). The experimental results (Table 3) indicate that these components contribute to improving the reconstruction accuracy.

3. Hyperparameter sensitivity: We analyzed the impact of different hyperparameter values (α in the gradient penalty term and β in the gradient consistency regularization) on the model performance. As shown in Tables 4 and 5, the settings we chose ($\alpha = 0.001$, $\beta = 0.01$) provide the best performance and maintain stable results over a wide range.

4. Extreme temperature scenarios: To further verify its stability, we manually masked the regions with drastic temperature gradient changes and compared the performance of ESTD-Net with the benchmark model (see Figures 5 and 6). Our method demonstrated superior continuity and accuracy in these challenging scenarios.

We hope this comprehensive set of experiments can address your concerns.

4. Lack of Domain-Specific Adaptation

The proposed method appears entirely generic—usable for standard image inpainting or sea-surface-temperature fields alike—without any targeted adaptation. It seems transplanted wholesale from computer vision, with no domain-specific modifications. Are surface-temperature gaps really analogous to arbitrary image holes? Authors should justify their design choices from a physical/meteorological standpoint.

--- Thank you for the suggestion. We agree that a direct transplant of image inpainting techniques to temperature field imputation without adaptation would be insufficient due to fundamental differences in data characteristics.

However, our method is not a generic transplant. Atmospheric temperature variability arises from the interplay of planetary-scale circulation patterns and localized convective systems. To disentangle and correct the temperature signals generated by these two intimately linked yet markedly distinct weather regimes, we devise a two-stage restoration framework. While we draw inspiration from vision-based architectures, we explicitly adapt them to the specific nature of geophysical temperature data, which lacks rich texture information but exhibits strong spatial smoothness and structural continuity, especially across large gaps.

To ensure domain-specific suitability, we have made the following targeted adaptations:

1. Gradient Consistency Regularization: Unlike texture-rich images, surface temperature fields are governed by physical gradients. We introduce a gradient consistency loss to enforce smooth transitions and avoid unrealistic discontinuities in restored areas.

2. Weighted Reconstruction Loss: Missing regions in temperature maps are often large and irregular. We assign higher weights to regions near boundaries where accurate interpolation is most critical, aligning better with physical expectations of smooth boundary propagation.

3. Two-stage Design Philosophy: The first stage (Transformer) captures broad, coarse-scale spatial structures; the second stage (U-Net) applies autoregressive refinement to reduce localized errors without introducing overfitting, which is critical given the low-texture nature of temperature fields.

4. Validation with Real Geophysical Data: As demonstrated in our Supplementary Material (e.g., edge-case temperature variation tests and additional comparisons with generic vision models like MAT and Palette), methods designed purely for visual images perform poorly on surface temperature data. These results highlight the limitations of direct application and the value of our proposed adaptations.

Minor Comments:

1. “Key Innovations” should be stated more objectively

The paper claims, "The model is augmented by two key innovations," yet both the "Weighted Reconstruction Loss" and the "Gradient Consistency Regularization" are commonplace in the image-inpainting literature. It is unclear whether these truly qualify as "innovations."

--- We thank the reviewers for their comments and agree that the initial statement might have exaggerated the novelty of the weighted reconstruction loss and the gradient consistency regularization method. We clarify that these components, when considered individually, are not entirely novel. However, specifically, the change in atmospheric temperature is caused by the interaction between the planetary-scale circulation patterns and the local convective systems. To separate and correct the temperature signals generated by these two closely related but distinct weather patterns, we designed a two-stage restoration framework. These two parts were adapted and integrated into our two-stage hybrid framework, which is specifically designed for the interpolation of Earth's surface temperature data for the Earth science task. Compared to traditional image restoration tasks, this field has unique challenges, such as textureless spatial areas and the need for physically meaningful gradients. Our ablation study shows that these loss terms significantly improve the spatial consistency and stability of the interpolation results in this situation. We will modify the wording in the final version to describe these components more objectively and refer to them as customized improvement measures.

2. Definition of PSNR's MAX Value

Why is the PSNR's MAX defined as it is? How does this differ from standard computer-vision images? Does MAX relate to temperature units? Is it a constant or variable? Please explain the phrase on line 350: "the maximum possible pixel value of an image."

--- Thank you for raising this important question. We would like to clarify that our task is not image inpainting, but rather data imputation for Earth surface temperature fields. We adopt image-derived metrics such as PSNR purely as a quantitative means to evaluate reconstruction quality, due to the 2D spatial structure of our data and the advantages of such metrics in reflecting pixel-level differences.

In our data preprocessing pipeline, we globally normalize the temperature data and multiply the normalized values by 255 in order to facilitate visualization and enable fair comparison with other models that operate on image-like data. Thus, in the PSNR computation, MAX is set to 255, following standard practice for 8-bit intensity values. However, we emphasize that this is a representational choice for evaluation only—it does not change the fact that the underlying task remains scientific data reconstruction.

In contrast, metrics such as MAE and RMSE are computed in the original temperature units. We have updated the manuscript to clearly state this distinction and avoid any potential confusion between our task and image restoration.

3. Edge-Case Temperature Variations

How does the proposed method perform in scenarios of rapid or high-amplitude temperature variation? These edge cases may reveal significant weaknesses.

--- We appreciate this insightful question. To verify the data recovery performance of our model in regions with extreme temperature variations, we conducted a new set of experiments. These regions were selected manually and randomly masked in the temperature map. These regions were chosen based on obvious high spatial gradients (for example, the boundary areas between land and sea). As shown in Section 7 of the Supplementary Material "Comparison of Edge-Case Temperature Variations", we compared the performance of ESTD-Net with three powerful baseline models (Palette, PConv U-Net, and MAT). Additionally, we also compared the absolute error graphs of the reconstructed outputs with the true values. Our model significantly outperforms the baseline models in terms of continuity and accuracy in high-gradient regions. Moreover, the absolute error graphs indicate that ESTD-Net can always generate lower reconstruction errors in these challenging situations, demonstrating its robustness and generalization ability, even under extreme spatial variation conditions.

4. Gradient-Consistency Implementation and Hyperparameter

Eq. (5)'s gradient-consistency term L_{gp} may require second-order derivatives—how is this implemented in code and what is the computational cost? Why is α set to 0.001? Similar hyperparameter questions apply to the Gradient Consistency Regularization.

--- Thank you for raising this point. First, we clarify that the gradient consistency regularization L_{gp} is implemented using first-order derivatives only, and does not require second-order gradients. Specifically, we compute the spatial gradients of both the real and generated data using PyTorch's `torch.autograd.grad` function with `create_graph=True`, which supports backpropagation through gradient operations without invoking higher-order differentiation. This method ensures stable training and low computational overhead. Second, to address the stability and effectiveness of the hyperparameters α (for adversarial gradient penalty) and β (for gradient consistency regularization), we conducted sensitivity experiments over a range of values: For $\alpha \in \{0.0005, 0.001, 0.005\}$, and For $\beta \in \{0.005, 0.01, 0.05\}$. As shown in our Supplementary Material (Table 4 and Table 5), the results remain stable across a reasonable range, with only slight variations. The default settings $\alpha = 0.001$ and $\beta = 0.01$ were selected as they yielded slightly better balance between reconstruction accuracy and smoothness. This confirms that the proposed gradient-consistency mechanism is not only computationally tractable but also robust to reasonable hyperparameter choices.

5. Equation Numbering and Punctuation

The definition of M_{ij}' is missing an equation number—it should be Eq.

(2). Furthermore, none of the equations ends with a comma or period, which is non-standard.

--- We thank the reviewer for pointing this out. The missing equation number for the definition of M_{ij}^{\prime} (currently unnumbered) should indeed be Eq. (2), and we acknowledge that all equations should follow standard punctuation conventions. These formatting issues will be carefully corrected in the next revision stage, as we understand they are important for clarity and consistency.

6. MTB Module Design Motivation

The motivation for the MTB module in Figure 3 is unclear. Why is the concatenation placed as shown? Why the specific sequence MCA - C - FC - MLP? Why eliminate Layer Normalization? It is not explained how these choices realize the authors' stated goal at line 245: "To address these challenges..."

--- We thank the reviewer for this insightful comment. The design of the MTB (Masked Transformer Block) is inspired by the MAT framework [1], which has shown that removing LayerNorm and replacing residual connections with feature concatenation can improve optimization stability and performance, particularly in high-mask-ratio scenarios. In our Earth surface temperature reconstruction task, many regions may lack observations, leading to sparse input features. Traditional LayerNorm tends to amplify invalid tokens in such cases, causing gradient instability during training. Therefore, we eliminate LayerNorm and adopt a feature concatenation strategy to preserve both high-level contextual and low-level spatial information. The order of MCA - Concatenation - FC - MLP allows us to first focus attention on valid regions (via MCA), then integrate the attended output with original features to retain detailed information before transformation. This design directly supports our goal of robust reconstruction under large-scale missing data.

7. Definition of FC vs. MLP

In Eq. (2), what exactly is "FC"? A fully connected layer (e.g. PyTorch's `nn.Linear`)? How does FC differ from MLP? If FC plus an activation is an MLP, why distinguish them? The authors provide no discussion.

--- Thank you for pointing this out. In a fact, "FC" refers to a single fully connected layer (i.e., a linear projection implemented via `nn.Linear` in PyTorch), while "MLP" refers to a sequence of two or more FC layers with non-linear activations, often used as a feed-forward network. Although technically an FC layer plus activation could be considered a minimal MLP, we maintain this distinction to clearly differentiate between shallow and deep transformations.

8. Adversarial Loss Stability

Eqs. (3) and (4) define the adversarial loss—yet is this formulation stable? Did the authors assess training convergence and robustness? In image tasks, these losses are notoriously unstable; more analysis is needed beyond Table 2 to support claims of "robust and reliable" performance.

--- We appreciate the reviewer's concern regarding the stability of adversarial training,

which is a well-known challenge in generative models. Several key strategies were employed in our framework to ensure stable and robust training:

R1 Regularization: We apply the R1 gradient penalty to the discriminator, which is widely recognized for stabilizing GAN training by preventing overfitting and ensuring smoother gradient flows. This technique effectively addresses the issue of discriminator dominance and encourages more reliable generator updates.

Exponential Moving Average (EMA): An EMA of the generator weights is maintained throughout training. This helps suppress training noise, smooths the weight trajectory over time, and improves the consistency and stability of the generated outputs — particularly beneficial in later training stages.

Adaptive Data Augmentation (ADA): We adopt ADA with automatic strength adjustment ($\text{aug}=\text{ada}$), which dynamically calibrates augmentation intensity based on discriminator feedback. This mechanism prevents the discriminator from overfitting to limited modes or learning shortcut solutions, thereby maintaining healthy competition between the generator and discriminator.

References

- [1] Li W, Lin Z, Zhou K, et al. Mat: Mask-aware transformer for large hole image inpainting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10758-10768.
- [2] Saharia C, Chan W, Chang H, et al. Palette: Image-to-image diffusion models[C]//ACM SIGGRAPH 2022 conference proceedings. 2022: 1-10.