

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

Minghui Zhang¹, Yunjie Chen¹, Fan Yang¹, and Zhengkun Qin²

¹School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China

²School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China

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

1 Training Time Comparison

We have compared our model with the traditional transformer model in terms of training time. The specific results are shown in the table 1. 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.

Table 1. Training Time Comparison.

Method	Time	Params
ViT-baseline	5d01h	102M
ESTD-Net(ours)	4d14h	95.8M

2 Model Comparison

To demonstrate the superiority of ESTD-Net in the task of restoring high-resolution temperature data, we conducted an experimental comparison with two recent models (MAT(Li et al., 2022) and Palette(Saharia et al., 2022)) used for data restoration tasks. The specific results are shown in Table 2 and Figure 1. 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 2, provide a detailed visualization of the reconstruction errors.

Table 2. Comparison of different methods.

Method	MAE↓	RMSE↓	PSNR↑	SSIM↑
MAT	0.0619	0.2717	54.4745	0.9977
Palette	0.1669	0.5558	49.1252	0.9953
ESTD-Net(ours)	0.0522	0.2000	56.9911	0.9985

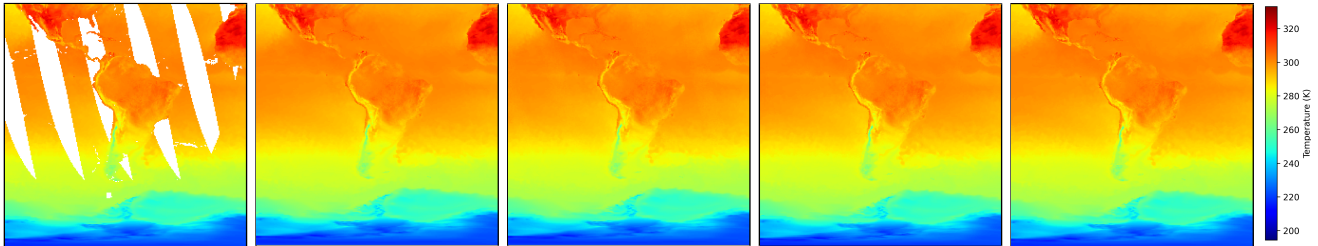


Figure 1. Reconstruction Results of Surface Temperature. From left to right, the columns display the initial incomplete data, the results of MAT, the results from Palette, the results from our proposed method, and the ground truth. All panels utilize identical color scaling to facilitate direct visual comparison.

3 The overall workflow of training and inference

To enhance clarity, we have included a concise step-by-step process description, as shown in Figure 3, which details our workflow, including the input, the two-stage reconstruction process, and the differences between training and inference.

4 The contextual attention module based on mask

To enhance readability, we have added a step-by-step pseudo-code description, as shown in Figure 4, to illustrate how the context attention mechanism operates in our spatial data imputation setting. This pseudo-code clearly explains how to fill in the missing values by focusing on the observed areas within the local spatial window.

5 Ablation and analysis

To verify the effects of the dedicated mask-based context attention module and the Stage II Conv-U-Net, we conducted relevant ablation experiments. The specific results are shown in Table 3. Removing the context attention would increase the mean absolute error from 0.0522 to 0.0717, and reduce the peak signal-to-noise ratio from 56.9911 decibels to 54.2463 decibels. Omitting the second-stage convolutional U-network would further degrade the performance (mean absolute error 0.1014, peak signal-to-noise ratio 51.2518 decibels). These verification experiments confirmed the crucial contributions of these two modules to the reconstruction quality of the ESTD network.

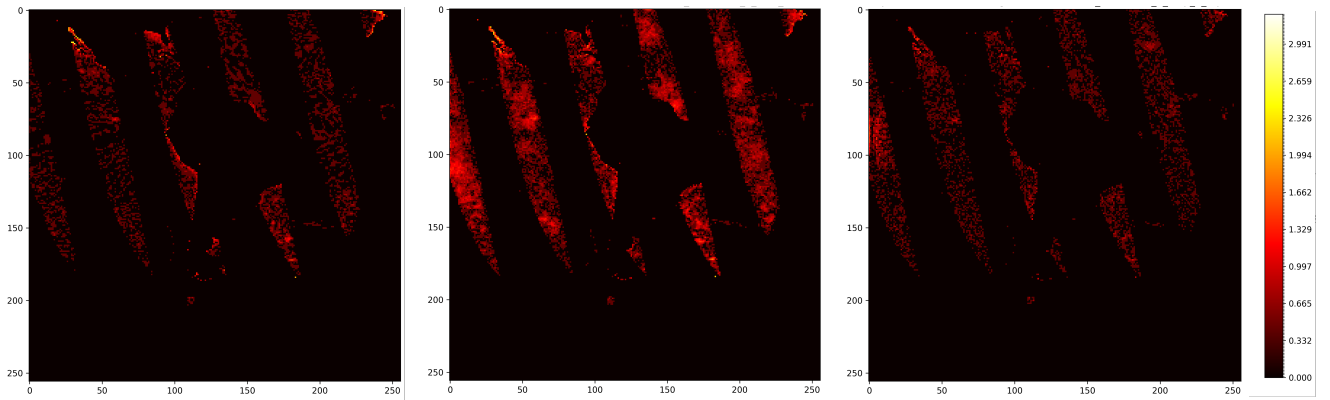


Figure 2. Comparison of Different Methods with Ground Truth. From left to right, the columns display the difference of the reconstructed data of MAT, Palette and our method with those of the ground truth.

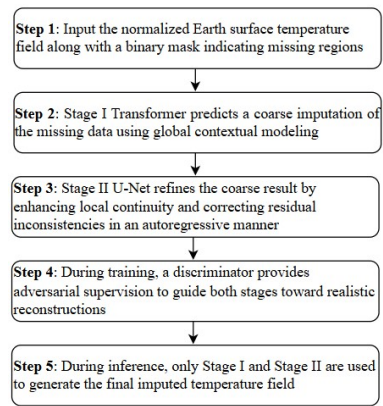


Figure 3. The overall workflow of training and inference.

Input:

F : Input data feature map, shape $[C, H, W]$

M : Binary mask indicating missing values (0 = missing, 1 = observed), shape $[1, H, W]$

win_size : Size of local attention window

Output:

F_out : Output feature map after applying contextual attention

- 1: Initialize $F_out \leftarrow F$
- 2: for each spatial location (i, j) where $M[i, j] == 0$ do
- 3: Extract a local window W centered at (i, j)
- 4: Define $Q \leftarrow F[:, i, j]$ // Query: feature at missing location
- 5: Collect $K, V \leftarrow$ features at positions $(m, n) \in W$ where $M[m, n] == 1$
- 6: For each (m, n) in K :
- 7: Compute attention score $A[m, n] \leftarrow \text{sim}(Q, K[m, n])$
- 8: Normalize A using softmax over all valid (m, n)
- 9: $F_out[:, i, j] \leftarrow \sum A[m, n] \times V[m, n]$
- 10: return F_out

Figure 4. The contextual attention module based on mask.

Table 3. Comparison of model variant.

Model Variant	MAE↓	RMSE↓	PSNR↑	SSIM↑
w/o Contextual Attention	0.0717	0.2770	54.2463	0.9975
w/o Stage-II Conv-U-Net	0.1014	0.3902	51.2518	0.9953
Full ESTD-Net	0.0522	0.2000	56.9911	0.9985

6 Hyperparameter analysis

Regarding the selection of the hyperparameters α in the gradient penalty term and β in the gradient consistency regularization term, we conducted sensitivity experiments by choosing different values for these hyperparameters. The results are shown in Table 4 (analysis of hyperparameter α in the gradient penalty term) and Table 5 (analysis of hyperparameter β in the gradient consistency regularization term). The analysis results indicate that the performance is relatively stable within a wide range of values, but our chosen hyperparameters yield the best results.

Table 4. Analysis of hyperparameter α in the gradient penalty term.

α Value	MAE↓	RMSE↓	PSNR↑	SSIM↑
0.005	0.0524	0.2021	56.9023	0.9985
0.001	0.0522	0.2000	56.9911	0.9985
0.0005	0.0533	0.2060	56.7626	0.9985

Table 5. Analysis of hyperparameter β in the gradient consistency regularization term.

β Value	MAE↓	RMSE↓	PSNR↑	SSIM↑
0.05	0.0585	0.2336	56.4033	0.9984
0.01	0.0522	0.2000	56.9911	0.9985
0.005	0.0559	0.2143	56.4088	0.9984

7 Comparison of Edge-Case Temperature Variations

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). We compared the performance of ESTD-Net with three powerful baseline models (Palette, PConv U-Net, and MAT), and the results are shown in Figure 5. Additionally, we also compared the absolute error graphs of the reconstructed outputs with the true values, and the results are shown in Figure 6. Our model significantly outperforms the baseline models in terms of continuity and accuracy in high-

40 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.

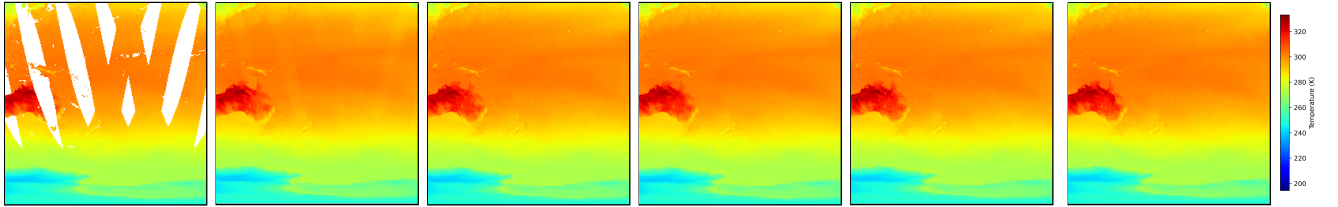


Figure 5. Reconstruction Results of Surface Temperature. From left to right, the columns display the initial incomplete data, the results from Palette, the results from PConv U-Net, the results of MAT, the results from our proposed method, and the ground truth. All panels utilize identical color scaling to facilitate direct visual comparison.

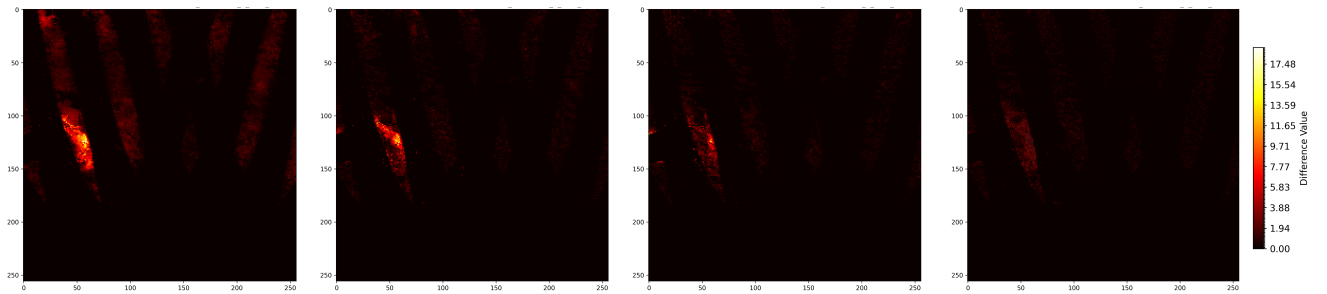


Figure 6. Comparison of Different Methods with Ground Truth. From left to right, the columns display the difference of the reconstructed data of Palette, PConv U-Net, MAT and our method with those of the ground truth.

References

- 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
45 IEEE/CVF conference on computer vision and pattern recognition, pp. 10 758–10 768, <https://doi.org/10.1109/CVPR52688.2022.01049>,
2022.
- Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M.: Palette: Image-to-image diffusion models, in:
ACM SIGGRAPH 2022 conference proceedings, pp. 1–10, 2022.