SVIFNN: Robust Inpainting Fourier Neural Network for SST Scientific Visualization Image: Leveraging Significant Stability and Non-Significant Anomalies

Zijie Zuo, *Student Member, IEEE*, Jie Nie, *Member, IEEE*, Xin Wang, *Member, IEEE*, Junyu Dong, *Member, IEEE*

***Abstract*—The Scientific Visualization Images (SVI) of Sea Surface Temperature (SST) play a pivotal role as visual resources for investigating oceanographic processes. However, they are often plagued by extensive data gaps due to objective factors like cloud cover. Additionally, their content deviates significantly from ordinary images, posing challenges for conventional completion techniques. Given the intricate nature of marine systems, completing the visualization of sea surface temperature presents several challenges. Firstly, predicting missing segments relies not only on prominent patterns but also on subtle anomalies, which are often overlooked by methods focused on extracting prominent features. Secondly, these images exhibit chaos and lack clear semantics, making it difficult for methods primarily focused on semantic extraction to effectively complete them. To address these challenges, this study presents a novel method named the Inpainting Fourier Neural Network for SVI (SVIFNN). This approach employs a twin-stream architecture to highlight both significant stability and non-significant anomalies. Notably, it incorporates a "reverse attention mechanism" in the non-significant anomalies extraction stream to preserve unconventional information. Furthermore, by cascading Fourier neural operator (FNO), it leverages frequency domain characteristics to mitigate spatial chaos. Through a frequency domain feature extraction module, it achieves an adaptive fusion of significant stability and non-significant anomalies. Experiment results demonstrate SVIFNN's superiority over State-Of-The-Art (SOTA) methods, particularly under a 68% missing rate condition. Significant improvements are observed in R^2 (18.1%, 19.3%, and 21.8%) and reductions in $RMSE$ (22.6%, 28.8%, and 23.8%) across different Noise-to-Signal (N/S) ratios of 0.1, 0.2, and 0.3, respectively, underscoring SVIFNN's robustness in handling SVIs extensive data gaps. Adequate ablation experiments further validate the effectiveness of the proposed non-significant anomalies extraction stream and frequency domain operators, with the latter demonstrating superior performance for scientific visualization images compared to traditional spatial domain CNN and ViT operators.**

***Index Terms*—Sea surface temperature, Deep learning, Scientific visualization image, Inpainting, Gap-free.**

I. Introduction

This work was supported in part by the National Key R&D Program of China (2023YFC3108700) and the National Natural Science Foundation of China (U23A20320). *(corresponding author: Jie Nie)*

Zijie Zuo, Jie Nie, and Junyu Dong are with the Faculty of Information Science and Engineering, the Ocean University of China, Qingdao 266100, China. (e-mail: zuozijie@stu.ouc.edu.cn; niejie@ouc.edu.cn; dongjunyu@ouc.edu.cn)

Xin Wang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. (e-mail: xin_wang@tsinghua.edu.cn)

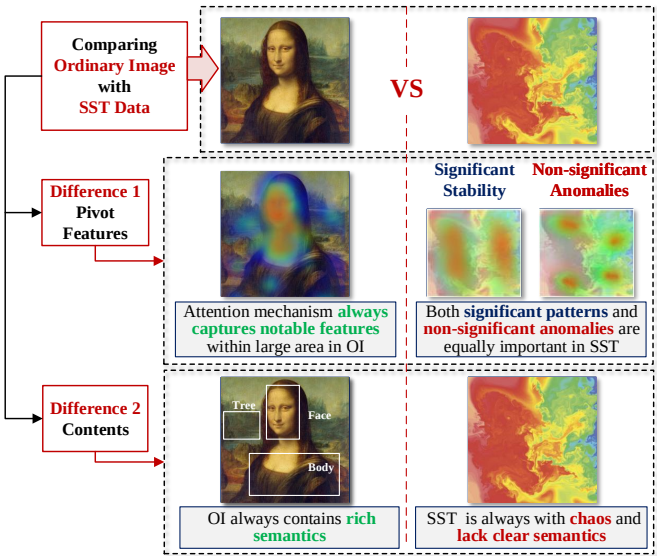


Fig. 1. The comparison between Ordinary Image (OI) and Scientific Visualization Image (SVI) underlines key differences. Difference 1: the pivot features. Compared to OI methods generally capturing notable features, the prediction of missing segments in SVI relies not just on prominent patterns but also on subtle anomalies. Difference 2: The contents. Compared to OI contains rich semantics, SVI is always with chaos and lack clear semantics. These issues pose significant challenges for conventional methods of handling SVIs.

SCIENTIFIC Visualization Images (SVI) of Sea Surface Temperature (SST) are critically important visual resources for exploring oceanographic phenomena [1], [2]. Nevertheless, these resources frequently encounter significant data voids, which are attributable to unavoidable elements such as cloud coverage. Thus, inpainting technologies in SST SVI are crucial for studying phenomena like ocean currents and marine heatwaves, enhancing our grasp of marine ecosystems, climate change effects, and maritime sustainability.

Traditional completion techniques are inspired by image completion [3]–[8], utilizing Deep Neural Network (DNN) architecture to fully explore the relationship between missing and known data through extensive historical datasets, thereby completing image completion tasks. Among these, [9]–[12] applied the Convolutional Neural Network (CNN) to capture significant patterns [9], [10] and semantics [11], [12] within historical data effectively, aiding in restoring detailed information in missing areas. However, constrained by the scope of known historical data, CNNs struggle to depict unknown

patterns. To address this issue, [11], [13]–[16] utilized the Generative Adversarial Network (GAN) to ensure that the completion results not only adhere to historical patterns but also introduce a certain degree of diversity. However, these methods are primarily developed for ordinary images and often overlook the differences and specific characteristics within SST SVI, as depicted in Fig. 1. Firstly, most conventional methods, particularly those employing attention mechanisms, concentrate on capturing significant representations by detecting similarities and adhering to notable patterns across spatial or temporal dimensions to achieve the inpainting task. However, the data patterns in SVI do not consistently exhibit such positive correlations, for example, certain local patterns in SVI, which significantly deviate from the notable patterns (manifested as anomalous changes in sea temperature, i.e., non-significant anomalies in SST SVI), are crucial indicators for predicting missing regions. Yet, these anomalous features are often overlooked or smoothed out by such methods. Secondly, these techniques rely on architectures like CNN or Transformer to capture rich semantic information [17] in the spatial domain (e.g., tree, face, river) for inpainting tasks. However, the complexity of marine systems results in chaotic characteristics for SST SVI, which lack distinct semantic information, thus hampering effective representation.

To tackle these challenges, this study introduces a groundbreaking Inpainting Fourier Neural Network specifically designed for SST SVI, referred to as SVIFNN. SVIFNN employs a twin-stream structure to capture both significant stability and non-significant anomalies simultaneously. Both streams share a similar architecture, but a unique "reverse attention mechanism" is integrated within the non-significant anomalies extraction stream to target non-significant anomalies. Additionally, drawing inspiration from the Fourier Neural Operator (FNO) [18], known for its proficiency in identifying effective patterns in complex Partial Differential Equation (PDE) solutions through frequency domain analysis, we design a frequency domain feature extraction module to facilitate an adaptive fusion of significant stability and non-significant anomalies to a unified frequency-domain feature set, and then achieve the completion of SST SVI.

Here, we make contributions as follows:

- We first propose an Inpainting Fourier Neural Network, namely SVIFNN, tailored for SST SVI, employing a twin-stream structure to separately capture significant stability and non-significant anomalies. Particularly, the non-significant anomalies extraction stream uses a unique "reverse attention mechanism" to fully preserve the unconventional anomalies features in daily SST images.
- We extend FNO to the SST SVI completion task by introducing a novel frequency domain feature extraction module to adaptively and effectively merge both types of information into a comprehensive feature set in the frequency domain, leveraging frequency domain characteristics to mitigate spatial chaos and improving the robustness in handling SVI's extensive data gaps.
- Our experimental evaluations reveal SVIFNN's superiority over state-of-the-art methods, particularly under a 68% missing rate condition. Notable improvements in R^2 (18.1%, 19.3%, and 21.8%) and reductions in $RMSE$ (22.6%, 28.8%, and 23.8%) across different conditions of Noise-to-Signal (N/S) ratios of 0.1, 0.2, and 0.3, which underscores SVIFNN's robustness in handling extensive data gaps. Adequate ablation experiments further validate the effectiveness of the proposed non-significant anomalies extraction stream and frequency domain operators.

II. Related Work

A. Traditional Completion Methods

Traditional completion methods range from deterministic and statistical interpolations to data assimilation techniques. Deterministic interpolation [19]–[22] relies on data patterns for quick estimates, with polynomial methods [19]–[21] enhancing interpolation accuracy. Regression interpolation [23], [24] and Kriging [25]–[28] focus on data's statistical properties and spatial autocorrelation, respectively. DINEOF-based works [29]–[33] tackle both spatial and temporal aspects to capture main patterns out from seen data. However, the effectiveness of these methods gradually diminishes as the data missing rate increases. Data assimilation methods [34]–[38], which integrate observations with ocean physical models through ensemble Kalman filtering [34]–[36], can effectively fill large data gaps. However, the search for optimal parameters demands significant computational resources, limiting their border application. Additionally, traditional methods often overlook the potential of historical data.

B. DNN-based Completion Methods

With the training on vast amounts of historical data, DNN-based methods like CNN and GAN have advanced significantly in SST image completion by leveraging extensive historical data. CNN-based approaches, exemplified by works such as [9]–[12], highlight CNN's utility in capturing significant patterns, with [12] introducing DINCAE 1.0 for SST reconstruction and reliability assessments. GANs, noted for their adeptness in image generation via adversarial training, have shown promise in biased data generation tasks in studies like [11], [13]–[16]. Among them, the works [10], [11], [13], [14] leverage the capture of significant patterns to enhance the quality of SST image inpainting. However, these methods often overlook non-significant anomalies in SST SVI. Recently, some studies [9], [12], [15], [16] have shifted focus towards capturing subtle SST anomalies to improve inpainting accuracy. For instance, [15] identifies anomalies by directly subtracting from the weekly average and uses convolution to capture the significant patterns of these anomalies. The learned patterns are then used to correct deviations in the missing areas by simply adding them back to complete the SST. [16] had decoupled anomaly features at multiple scales and improved the final fusion strategy. Nevertheless, these methods merely utilize the difference between two images to describe anomalies without reinforcing them, thereby failing to ensure that anomaly patterns play an effective role in prediction. Moreover, the aforementioned works solely concentrate on representation learning in the spatial domain and face challenges in capturing the chaotic features of SST SVI.

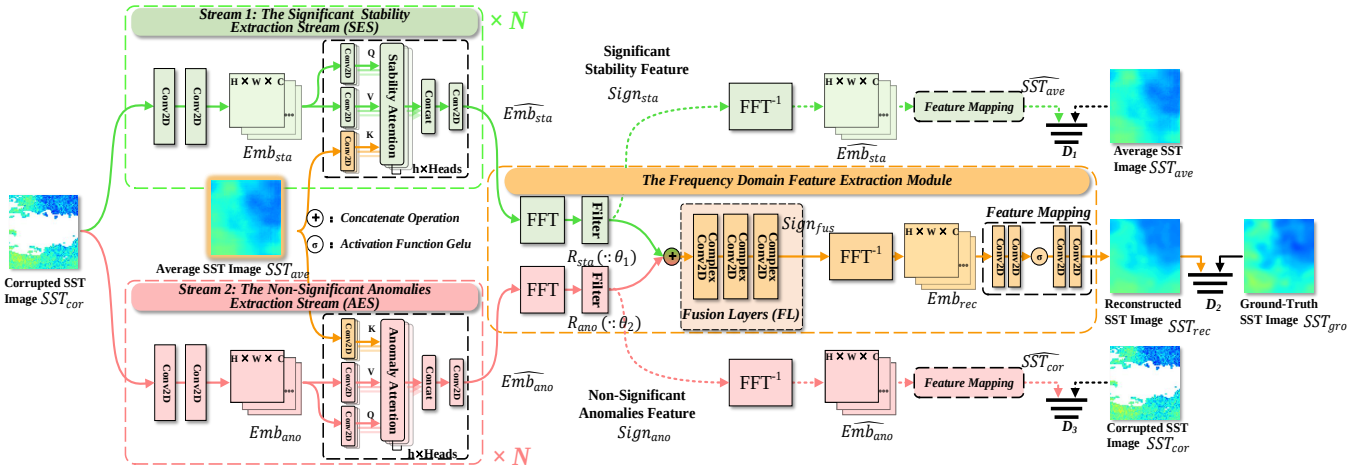


Fig. 2. The main framework of the proposed SVIFNN. SVIFNN employs a twin-stream structure to learn spatial domain representations for significant stability and non-significant anomalies in SST SVI. The significant stability extraction stream (SES) employs a multi-head attention mechanism to extract the positively correlated significant stability $\hat{Emb}_{sta}$ between SST_{cor} and SST_{ave}. The non-significant anomalies extraction stream (AES), while structurally similar to the SES, distinctively features a "reverse attention mechanism". This mechanism is intricately designed to thoroughly capture and preserve the negatively correlated non-significant anomalies $\hat{Emb}_{ano}$ between SST_{cor} and SST_{ave}. These representations are then fused in the frequency domain feature extraction module into $Sign_{fus}$ via the Fusion Layers (FL) for SST image completion, which is subsequently transformed back to the spatial domain and compiled into the final SST image (SST_{rec}). Ground Truth SST images (SST_{gro}) and discriminators assess the integrity of SST_{rec}, $\hat{SST}_{ave}$, and $\hat{SST}_{cor}$ to ensure completeness.

C. Chaotic Systems Representation Methods

Recent advancements in chaotic system modeling have been greatly influenced by the Fourier Neural Operator (FNO) [18], which captures structural patterns and dependencies in complex Partial Differential Equation (PDE) solutions within the frequency domain. Building on this foundation, recent research [39]–[41] has extended FNO applications to marine sciences and oceanography, aiming to develop effective representation methods for chaotic systems. Over the past few years, FNO and its variants have exhibited remarkable proficiency in predictive tasks [41]–[43]. For instance, the Adaptive Fourier Neural Operator (AFNO) was implemented in FourCastNet [42] for high-resolution weather forecasting, while Li et al. [43] utilized an implicit U-Net enhanced FNO (IU-FNO) to achieve long-term turbulence predictions. Furthermore, Surapaneni [41] successfully applied FNO for near-real-time predictions of ocean wave behavior. These efforts highlight the robustness of FNO in modeling complex, spatiotemporal dynamics, demonstrating its capability to represent intricate chaotic systems.

The outstanding performance of FNO in predictive tasks has prompted researchers to explore its potential in data reconstruction [44], [45]. For instance, Chen et al. [44] introduced the Fourier Imager Network (FIN) for end-to-end holographic image reconstruction, and Ehlers et al. [45] combined U-Net with FNO to reconstruct ocean wave data from spatiotemporal radar signals with high accuracy. Despite these promising advances, the reconstruction of SST SVI presents additional challenges. Compared to the holographic image data in [44], SST SVI exhibits more intricate and strongly coupled spatiotemporal patterns. Moreover, while SST and PDEs share certain structural similarities, the complexity of the real marine environment introduces substantial difficulties when attempting to apply FNO to real-world SST data. Although Ehlers et al. [45] made strides in reconstructing ocean wave data, extending FNO to the reconstruction of SST data remains an unexplored area of research.

In summary, although FNO has shown significant potential in representing chaotic systems for predictive tasks, its application to SST SVI completion is still in its infancy. This underscores the need for innovative approaches to represent and complete SST SVI data, particularly within the chaotic and highly dynamic nature of marine environments.

III. Method

A. Framewarok

Fig. 2 introduces our innovative SVIFNN framework, specifically designed for the reconstruction of SST images that are obscured by cloud coverage. We designate the cloud-covered SST image captured on a specific date T within the week W as SST_{cor}, alongside its corresponding weekly average SST image, referred to as SST_{ave}.

Utilizing SST_{cor} and SST_{ave} as inputs, SVIFNN precisely inpaints missing regions in SST_{cor} with the goal of achieving an accurate reconstruction. The process involves a twin-stream approach consisting of the significant stability extraction stream (SES) and the non-significant anomalies extraction stream (AES), which learn spatial representations of significant stability ($\hat{Emb}_{sta}$) and non-significant anomalies ($\hat{Emb}_{ano}$), respectively. The SES employs a multi-head attention mechanism to extract positively correlated stability information between SST_{cor} and SST_{ave}, while AES uses a "reverse attention mechanism" to focus on negatively correlated non-significant anomalies. The integration of these features is handled by the frequency domain feature extraction module, which uses Fast Fourier Transform (FFT) to shift these embeddings into the frequency domain, retains essential frequency signals as $Sign_{sta}$ and $Sign_{ano}$ after the learnable

TABLE I
DETAIL ARCHITECTURE OF OUR DISCRIMINATOR

Type	Kernel	Stride	Outputs
Convolution	2×2	2×2	64
Convolution	2×2	2×2	128
Convolution	2×2	2×2	256
Convolution	2×2	1×1	256
Convolution	2×2	1×1	256
Convolution	2×2	1×1	512
Convolution	5×5	1×1	1

filters, and then merges them into $Sign_{fus}$ through the Fusion Layers (FL) for effective SST image completion. This composite frequency domain feature is then transformed back to the spatial domain using an inverse FFT and enhanced through feature mapping to produce the final completed SST image, denoted as SST_{rec}. This image, along with $S\hat{S}T_{ave}$ and $S\hat{S}T_{cor}$, is evaluated against the ground truth SST image to ensure authenticity and completeness.

B. The Significant Stability Extraction Stream (SES)

The significant stability extraction stream is designed to identify and encode the significant stability traits of SST images within the spatial domain, as outlined in Fig. 2. Through initial processing of SST_{cor} and SST_{ave}, SES applies two convolutional layers to capture the spatial features of SST_{cor} (Emb_{sta}) and SST_{ave} (Emb_{ave}), both in $\mathbb{R}^{H\times W\times C}$. Utilizing a multi-head attention mechanism, it then explores the significant stability across these spatial representations, resulting in $\hat{Emb}_{sta}$, through the formula:

$$\hat{Emb}_{sta} = concat\left(head_1, head_2, \ldots, head_h\right) W_{sta}, \quad (1)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (2)$$

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \quad (3)$$

where $head_h$ signifies the multiple heads in the attention mechanism, where h represents the number of heads, and W_{sta} denotes the weight matrix for the concatenation operation. Q, K, and V respectively represent the query, key, and value, with W_i^Q, W_i^K, and W_i^V being their corresponding weight matrices. d_k is the dimension of the vectors in K. The $softmax(\cdot)$ function is utilized to compute the relevance between Q and V. Note that here, both Q and V are derived from Emb_{sta}, while K originates from Emb_{ave}. Therefore, $\hat{Emb}_{sta}$ enhances features in Emb_{sta} that bear similarity to the mean Emb_{ave}, encapsulating the significant stability features in the spatial domain.

C. The Non-Significant Anomalies Extraction Stream (AES)

The non-significant anomalies extraction stream, depicted at the bottom of Fig. 2, inputs SST_{cor} and SST_{ave} to pinpoint and encapsulate non-significant anomaly traits in the spatial domain. Similar to the SES, AES initially employs two convolutional layers to extract SST_{cor}'s spatial features as Emb_{ano}, also in $\mathbb{R}^{H\times W\times C}$. A modified multi-head attention mechanism is then used to reveal the spatial domain's non-significant anomalies between Emb_{ano} and Emb_{ave}, expressed as $\hat{Emb}_{ano}$:

$$\hat{Emb}_{ano} = concat\left(head_1, head_2, \ldots, head_h\right) W_{ano}, \quad (4)$$

$$head_i = Attention_{reverse}\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (5)$$

where $Attention_{reverse}$ denotes the "reverse attention mechanism", it is specifically designed to preserve the features of non-significant anomalies in daily SST images. This contrasts with the attention mechanism used in SES, which focuses on calculating significant stability. The formula for $Attention_{reverse}$ is as follows:

$$Attention_{reverse}\left(Q, K, V\right) = \left(1 - softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)\right) V, \quad (6)$$

where both Q and V are derived from Emb_{ano}, while K originates from Emb_{ave}, which allows AES to conserve detailed information on daily SST image inconsistencies against the average, capturing essential non-significant anomalies features.

D. The Frequency Domain Feature Extraction Module

The frequency domain feature extraction module, shown in the center of Fig. 2, integrates $\hat{Emb}_{sta}$ and $\hat{Emb}_{ano}$ from the SES and AES, respectively, merging them into a unified frequency-domain feature set through the Fusion Layers (FL). This fusion not only combines the significant features of stability with the subtle features of anomalies but also enhances the accuracy of inpainting.

Specifically, $\hat{Emb}_{sta}$ and $\hat{Emb}_{ano}$ are converted into the frequency domain using fast Fourier transform and filtered to isolate essential frequency domain signals, through:

$$Sign_{sta} = R_{sta}\left(FFT(\hat{Emb}_{sta}); \theta_1\right), \quad (7)$$

$$Sign_{ano} = R_{ano}\left(FFT(\hat{Emb}_{ano}); \theta_2\right), \quad (8)$$

where $FFT(\cdot)$ denotes the fast Fourier transform, and $R_{sta}(\cdot;\theta_1)$ and $R_{ano}(\cdot;\theta_2)$ represent the adaptive filtering operation with θ_1 and θ_2 representing these filters' learnable parameters respectively, distinct from conventional low-pass filters in FNO, to retain crucial high-frequency components of significant stability signals. Here, $R_{ano}(\cdot;\theta_2)$ effectively highlights and retains critical discrepancies for oceanographic and climatic analysis.

Subsequently, $Sign_{ave}$ and $Sign_{ano}$ are first concatenated and then passed through the Fusion Layers (FL), which consist of three consecutive convolutional layers, to achieve a deep fusion of frequency signals, denoted as $Sign_{fus}$. The process is expressed by the following formula:

$$Sign_{fus} = l\left(l\left(l\left(Sign_{ave}, Sign_{ano}\right)\right)\right), \quad (9)$$

note here that $l(\cdot)$ represents a layer of complex convolution operation, complex convolution is an improved version of the standard convolution tailored for data in complex form [47]. Note that before performing complex convolution, the input will first undergo a concatenate operation (as shown in Fig. 2).

TABLE II

THE COMPARISON RESULTS FOR SST RECONSTRUCTION USING SIX METHODS ON THE NSOAS DATASET. WE USE THE ROOT MEAN SQUARE ERROR ($RMSE$) AND R-SQUARED (R^2) AS METRICS TO EVALUATE THE RECONSTRUCTION RESULTS. THE BEST AND SECOND-BEST MEAN SQUARED ERROR ($RMSE$) AND R-SQUARED (R^2) ARE IN **BOLD** AND UNDERLINE UNDER DIFFERENT MISSING RATIOS WITH DIFFERENT NOISE-TO-SIGNAL (N/S) RATIOS

N/S	Missing Ratio	AIN [15]		CF_DGM [4]		DINEOF [46]		DINCAE [12]		Phy_INN [16]		Ours	
		$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2
0.1	8%	0.1426	0.9335	0.2383	0.7981	0.1687	0.9441	0.1612	0.8865	0.0819	0.9832	**0.0490**	**0.9967**
	25%	0.1622	0.8454	0.2415	0.5057	0.1713	0.9070	0.1705	0.7763	0.0912	0.9429	**0.0520**	**0.9889**
	46%	0.1726	0.7356	0.2431	0.2109	0.1727	0.8334	0.1779	0.5887	0.0941	0.8915	**0.0707**	**0.9536**
	68%	0.2041	0.5632	0.2676	0.1919	0.1911	0.2929	0.1921	0.3994	0.1032	0.7827	**0.0799**	**0.9246**
0.2	8%	0.1622	0.9279	0.2494	0.7521	0.1784	0.9336	0.1844	0.8731	0.0822	0.9824	**0.0527**	**0.9962**
	25%	0.1714	0.8261	0.2505	0.4541	0.1806	0.8910	0.1846	0.7676	0.0944	0.9413	**0.0621**	**0.9850**
	46%	0.1843	0.6637	0.2611	0.1857	0.1814	0.8305	0.1878	0.5402	0.1014	0.8859	**0.0758**	**0.9464**
	68%	0.2142	0.4831	0.2703	0.1834	0.1967	0.2733	0.1978	0.3579	0.1133	0.7690	**0.0807**	**0.9178**
0.3	8%	0.1766	0.9131	0.2505	0.7308	0.1835	0.9281	0.1924	0.8611	0.0825	0.9790	**0.0583**	**0.9954**
	25%	0.1779	0.7629	0.2523	0.3034	0.1860	0.8816	0.1942	0.7447	0.0964	0.9334	**0.0647**	**0.9826**
	46%	0.1941	0.6546	0.2699	0.1701	0.1899	0.8244	0.2011	0.4538	0.1065	0.8789	**0.0802**	**0.9382**
	68%	0.2231	0.4633	0.2736	0.1692	0.2014	0.2651	0.2035	0.3435	0.1141	0.7461	**0.0870**	**0.9084**

Then, $Sign_{fus}$ is transformed back to the spatial domain using an inverse FFT as follows:

$$\hat{Emb}_{rec} = FFT^{-1}\left(Sign_{fus}\right), \tag{10}$$

where $\hat{Emb}_{rec}$ represents the spatial domain representation of $Sign_{fus}$, with dimensions $\in \mathbb{R}^{H\times W\times C}$, and $FFT^{-1}(\cdot)$ represents the inverse FFT operation. Subsequently, $\hat{Emb}_{rec}$ undergoes a nonlinear enhancement through a feature mapping operation, culminating in the complete SST image, denoted as SST_{rec}. This feature mapping operation consists of four convolutional layers, interspersed with GELU activation functions.

Finally, a discriminator D_2 is used, along with the referencing ground truth SST image (SST_{gro}), to verify the integrity of SST_{rec}. The discriminator is composed of 2-D convolutional layers, as detailed in Table I, which accepts the SST field, an $H\times W\times 1$ tensor, as input and produces a scalar to distinguish whether the input is real (the ground truth SST_{gro}) or fake (the reconstructed SST_{rec} generated by SVIFNN). For activation functions, a sigmoid function is applied after the final convolutional layer, and a leaky-ReLU follows each preceding layer. Moreover, to guarantee the effectiveness of the significant stability and non-significant anomalies features captured by SES and AES, the inverse FFT and feature mapping operations are employed to convert $Sign_{sta}$ and $Sign_{ano}$ back to spatial domain images, forming $\hat{SST}_{ave}$ and $\hat{SST}_{cor}$, overseen by discriminators D_1 and D_3 to ensure comprehensive feature representation.

E. The Loss Function

The SVIFNN's loss function is bifurcated into reconstruction loss L_{rec} and adversarial loss L_{adv}, which are defined as follows:

$$L_{rec} = \lambda_{sta}^{rec}L_{sta}^{rec} + \lambda_{cor}^{rec}L_{cor}^{rec} + \lambda_{ano}^{rec}L_{ano}^{rec}, \tag{11}$$

where λ_{sta}^{rec}, λ_{cor}^{rec}, and λ_{ano}^{rec} are hyper-parameters to balance the reconstruction loss components. L_{sta}^{rec}, L_{cor}^{rec}, and L_{ano}^{rec} are defined as:

$$L_{sta}^{rec} = f\left(\hat{SST}_{ave}, SST_{ave}\right), \tag{12}$$

$$L_{cor}^{rec} = f\left(SST_{rec}, SST_{gro}\right), \tag{13}$$

$$L_{ano}^{rec} = f\left(\hat{SST}_{cor}, SST_{cor}\right) \odot Mask_{cor}, \tag{14}$$

where SST_{gro} is the ground truth, and $\odot$ is Hadamard product. The function $f(\cdot)$ is defined as a combined L2 and L1 norm over N pixels:

$$f\left(\hat{y},\ y\right) = \frac{1}{N}\sum_{1}^{N}\left(\|\hat{y} - \ y\|_2 + \|\ \hat{y} - \ y\|_1\right), \tag{15}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ is the L2 and L1 norm, respectively. N represents the number of pixels in $\hat{y}$. And the $Mask_{cor}$ is defined as:

$$Mask_{cor}^{(i)} = \begin{cases} 0, \left(Mask_{cor}^{(i)}\ is\ occluded\right) \\ 1, \left(Mask_{cor}^{(i)}\ is\ not\ occluded\right) \end{cases}. \tag{16}$$

The adversarial loss L_{adv} is formulated to measure the discriminator's ability to distinguish between generated and real SST images, which is defined as follows:

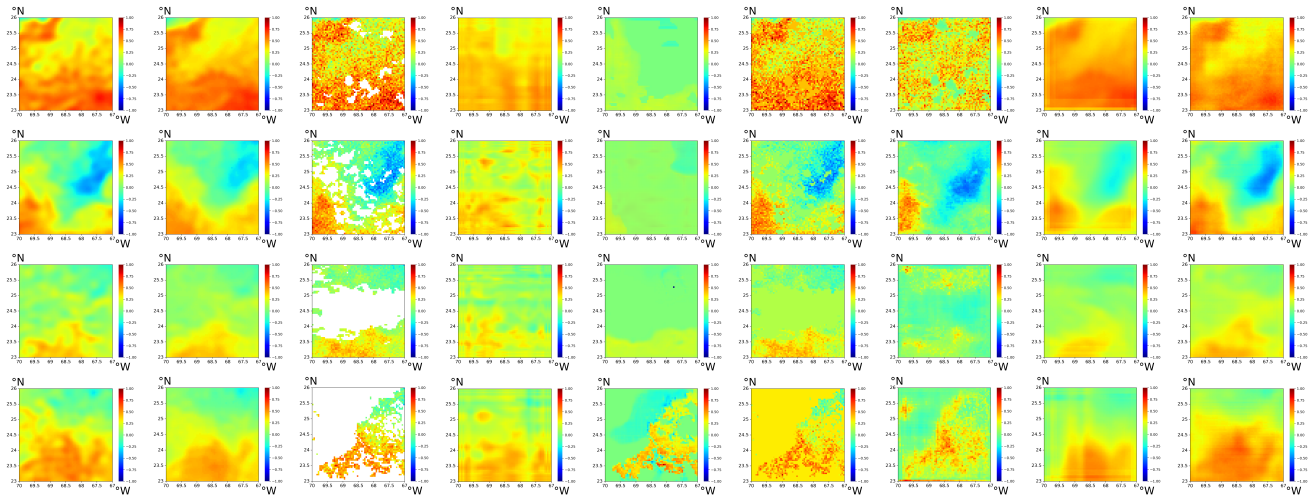
$$L_{adv} = \lambda_{sta}^{adv}L_{sta}^{adv} + \lambda_{cor}^{adv}L_{cor}^{adv} + \lambda_{ano}^{adv}L_{ano}^{adv}, \tag{17}$$

where λ_{sta}^{adv}, λ_{cor}^{adv}, and λ_{ano}^{adv} are hyper-parameters to balance each component. L_{sta}^{adv}, L_{cor}^{adv}, and L_{ano}^{adv} are defined as:

$$L_{sta}^{adv} = -(log(1 - D_1(\hat{SST}_{ave})) + logD_1(SST_{ave})), \tag{18}$$

$$L_{cor}^{adv} = -\left(log\left(1 - D_2\left(SST_{rec}\right)\right) + logD_2(SST_{gro})\right), \tag{19}$$

$$L_{ano}^{adv} = -(log(1 - D_3(\hat{SST}_{cor} \odot Mask_{cor})) + logD_3(SST_{cor} \odot Mask_{cor})), \tag{20}$$



(a) Ground Truth (b) Average SST (c) Corrupted SST (d) AIN [15] (e) CF_DGM [4] (f) DINEOF [46] (g) DINCAE [12] (h) Phy_INN [16] (i) Ours

Fig. 3. Comparative analysis of the visualization results from six methods on the NSOAS dataset. The x- and y-axes represents longitude (°W) and latitude (°N), respectively. The rows are missing ratios of 8%, 25%, 46%, and 68% from top to bottom, while N/S is all 0.1. (a) Ground truth SST images; (b) Average SST images; (c) Cloud-obscured images; (d)-(i) Results from **AIN**, **CF_DGM**, **DINEOF**, **DINCAE**, **Phy_INN**, **SVIFNN**.

where D_1, D_2, and D_3 represent discriminators that are respectively responsible for supervising $S\hat{S}T_{ave}$, SST_{rec}, and $S\hat{S}T_{cor}$. Then, we define the total loss function as follows:

$$L_{total} = L_{rec} + L_{adv}. \tag{21}$$

IV. Experiments

A. Datasets

In this study, we used visualization images of the NSOAS's publicly available SST level-4 products from January 2022 to April 2023, which are located in the western Atlantic region between 23°N to 26°N and 70°W to 67°W, with each pixel representing 5-km^2. We generated SST datasets with various cloud cover levels by integrating cloud masks from the WHU dataset [48] and added noise to simulate different N/S ratios as per DINEOF's method [46]. The dataset from January to December 2022 served as the training set, and the one from January to April 2023 was used for testing.

B. Experimental Details

Before training our model, we normalize the SST data to [-1, 1]. Then we set the learning rate as 0.0001 and use the Adam optimizer for training our model, and the parameters of the Adam optimizer, i.e., β_1 and β_2, are set to 0.5 and 0.99. The N in Fig. 2 is set as 4. The size of each SST field is set to W=64, H=64. We use grid search [49] to find the best setting of the hyper-parameters C=64, λ_{sta}^{rec}=9, λ_{cor}^{rec}=20, λ_{ano}^{rec}=9, λ_{sta}^{adv}=1, λ_{cor}^{adv}=1, λ_{ano}^{adv}=1. For experiments, the cloud cover ratio is set to [8%, 25%, 46%, 68%] and the N/S ratio is set to [0.1, 0.2, 0.3]. To ensure sufficient training and testing samples for fully training the model and enhancing its generalization capability, we generated 10 cloud templates for each missing rate category: 10%, 30%, 50%, and 70%, corresponding to average missing rates of 8%, 25%, 46%, and 68%, respectively. For the training set, all 10 cloud templates were applied to the original SST field, significantly expanding the dataset. For the testing set, 8 cloud templates were randomly selected for each missing rate. To further augment the dataset, techniques such as rotation and flipping were employed. As a result, the training set consisted of 2,880 samples per missing rate, while the testing set contained 672 samples, maintaining a training-to-testing ratio of approximately 80%:20%. During the experiments, the training and testing sets for each missing rate were randomly shuffled, and the final results were reported as the average of multiple runs.

C. Testing Metrics

To ensure consistent evaluation with traditional methods that operate in the data domain, we map the completed visualization images back to the original data domain for comparison. The reconstruction performance is assessed using $RMSE$ and R^2 as the evaluation metrics:

Root Mean Squared Errors ($RMSE$): A smaller $RMSE$ value signifies less deviation from the ground truth and higher accuracy. The formula is given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{22}$$

where N represents the number of samples, y_i denotes the actual observed values, and $\hat{y}_i$ represents the predicted values by the model.

R-squared (R^2): This metric ranges from 0 to 1, with values closer to 1 indicating greater alignment between the reconstructed data's spatiotemporal patterns and the ground truth, signifying better reconstruction quality. The formula is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{23}$$

TABLE III
THE ABLATION RESULTS FROM ***w/o* FNO**, ***w/o* AES**, ***w/o* FUSION LAYERS (FL)**, AND **SVIFNN**. WE USE THE ROOT MEAN SQUARE ERROR ($RMSE$), R-SQUARED (R^2), STRUCTURAL SIMILARITY INDEX MEASURE ($SSIM$), AND PEAK SIGNAL-TO-NOISE RATIO ($PSNR$) AS METRICS TO EVALUATE THE RECONSTRUCTION. THE BEST RESULTS ARE IN **BOLD**, AND THE SECOND-BEST RESULTS ARE IN UNDERLINED

Missing Ratio	***w/o* FNO**				***w/o* AES**				***w/o* FL**				**SVIFNN**			
	$RMSE$	R^2	$SSIM$	$PSNR$	$RMSE$	R^2	$SSIM$	$PSNR$	$RMSE$	R^2	$SSIM$	$PSNR$	$RMSE$	R^2	$SSIM$	$PSNR$
8%	0.0545	0.9855	0.6364	71.3664	0.0636	0.9764	0.6376	70.2423	0.0521	0.9874	0.6411	71.8555	**0.0490**	**0.9967**	**0.7780**	**74.5381**
25%	0.0644	0.9335	0.6156	70.2038	0.0654	0.9265	0.5998	69.8369	0.0631	0.9487	0.6217	71.1026	**0.0520**	**0.9889**	**0.7439**	**74.0515**
46%	0.0761	0.9053	0.5912	69.7893	0.0792	0.9021	0.5842	68.6067	0.0737	0.9127	0.6157	71.0629	**0.0707**	**0.9536**	**0.6710**	**71.4941**
68%	0.0817	0.8701	0.5690	69.3328	0.0836	0.8691	0.5537	67.1964	0.0804	0.8711	0.6032	70.3852	**0.0799**	**0.9246**	**0.6418**	**70.5067**

TABLE IV
THE ABLATION RESULTS FROM ***w/* CNN**, ***w/* VIT** AND **SVIFNN**. WE USE THE ROOT MEAN SQUARE ERROR ($RMSE$), R-SQUARED (R^2), STRUCTURAL SIMILARITY INDEX MEASURE ($SSIM$), AND PEAK SIGNAL-TO-NOISE RATIO ($PSNR$) AS METRICS TO EVALUATE THE RECONSTRUCTION. THE BEST AND SECOND-BEST ARE IN **BOLD** AND UNDERLINE UNDER DIFFERENT MISSING RATIOS WITH DIFFERENT NOISE-TO-SIGNAL (N/S) RATIOS

N/S	Missing Ratio	***w/* CNN**				***w/* ViT**				**SVIFNN**			
		$RMSE$	R^2	$SSIM$	$PSNR$	$RMSE$	R^2	$SSIM$	$PSNR$	$RMSE$	R^2	$SSIM$	$PSNR$
0.1	8%	0.0618	0.9876	0.7159	72.8731	0.0577	0.9891	0.6430	69.2587	**0.0490**	**0.9967**	**0.7780**	**74.5381**
	25%	0.0701	0.9742	0.6533	71.1361	0.0603	0.9759	0.6394	68.3230	**0.0520**	**0.9889**	**0.7439**	**74.0515**
	46%	0.0833	0.9242	0.6022	70.4779	0.0725	0.9446	0.5359	68.1853	**0.0707**	**0.9536**	**0.6710**	**71.4941**
	68%	0.0868	0.9038	0.5855	69.1433	0.0885	0.9089	0.5279	67.6257	**0.0799**	**0.9246**	**0.6418**	**70.5067**
0.2	8%	0.0690	0.9865	0.6781	72.6241	0.0647	0.9878	0.6245	68.1283	**0.0527**	**0.9962**	**0.7006**	**73.9045**
	25%	0.0715	0.9685	0.6422	70.2910	0.0700	0.9731	0.5944	67.4686	**0.0621**	**0.9850**	**0.6736**	**72.6034**
	46%	0.0846	0.9162	0.5998	69.4853	0.0801	0.9348	0.5249	67.4522	**0.0758**	**0.9464**	**0.6248**	**70.8031**
	68%	0.0878	0.9025	0.5707	68.4329	0.0891	0.9076	0.5147	66.7130	**0.0807**	**0.9178**	**0.6042**	**70.1523**
0.3	8%	0.0743	0.9869	0.6560	70.3468	0.0715	0.9839	0.6164	67.5602	**0.0583**	**0.9954**	**0.6587**	**72.9888**
	25%	0.0765	0.9679	0.6161	69.0778	0.0740	0.9743	0.5846	67.0155	**0.0647**	**0.9826**	**0.6358**	**72.1702**
	46%	0.0878	0.9135	0.5738	68.0030	0.0870	0.9257	0.5126	66.4418	**0.0802**	**0.9382**	**0.6181**	**70.6139**
	68%	0.0907	0.9017	0.5606	67.6504	0.0912	0.8975	0.5054	65.1297	**0.0870**	**0.9084**	**0.6034**	**70.0900**

where N represents the number of samples, y_i denotes the actual observed values, and $\hat{y}_i$ represents the predicted values by the model, and $\bar{y}$ representing the mean of the observed values.

D. Comparison to State-Of-The-Art (SOTA) Methods

In this section, we evaluate the performance of our proposed method against five SOTA methods. These methods were selected because they encompass a diverse range of approaches, including traditional methods (e.g., DINEOF [46]) and DNN-based methods specifically designed for SST inpainting (e.g., AIN [15], DINCAE [12], Phy_INN [16]), as well as DNN-based methods for general image inpainting (e.g., CF_DGM [4]). This diversity makes them highly suitable for a comprehensive comparison in the context of SST inpainting. The details of the selected methods are as follows:

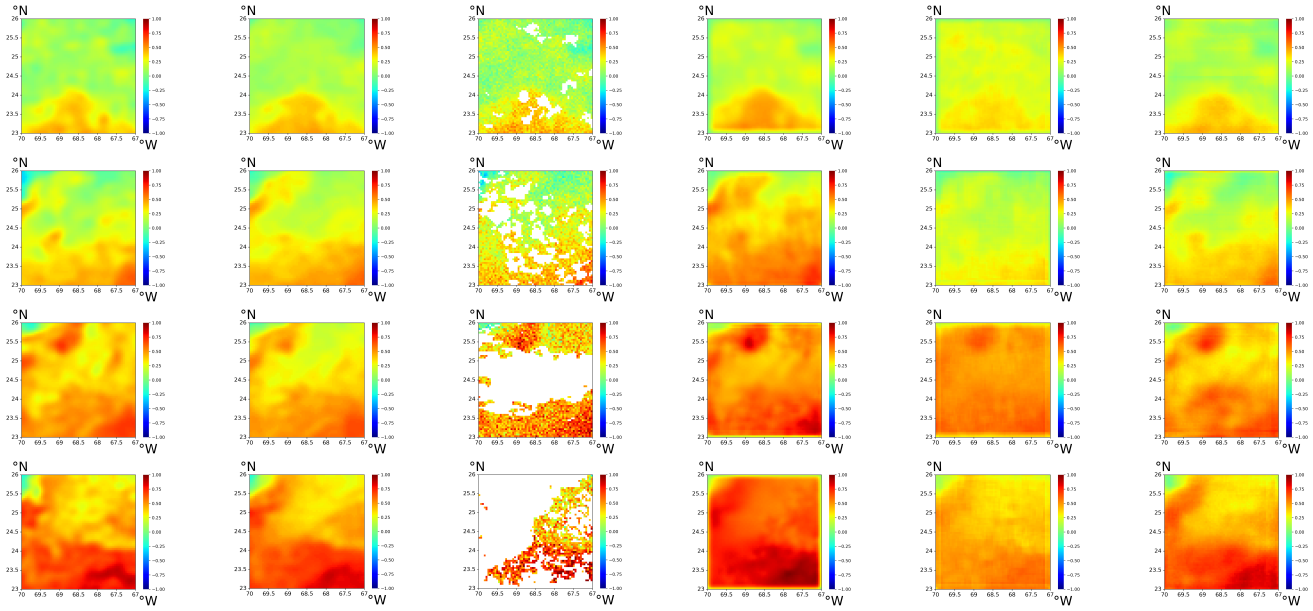
1) AIN [15]: A GAN-based framework specifically designed for SST inpainting, AIN adopts a "coarse-to-fine" strategy within the image domain. The method first predicts the weekly mean from the monthly mean, and then estimates the daily anomalies based on the predicted weekly mean. The final reconstructed SST is obtained by summing the weekly mean and daily anomalies.

2) CF_DGM [4]: Similar to AIN, CF_DGM is a GAN-based inpainting method designed for remote sensing images and also follows a "coarse-to-fine" approach. Its generator utilizes a U-net architecture and incorporates spatial semantic attention mechanisms. These mechanisms first capture global semantic correlations for an initial estimate, followed by local feature refinement to improve the inpainting results.

3) DINEOF [46]: As a classic method based on Empirical Orthogonal Functions (EOF), DINEOF is widely used in geophysical studies, including SST data, to fill in missing values. By identifying dominant spatial patterns that capture the primary variations within the dataset, this method utilizes these patterns to interpolate and reconstruct the missing data.

4) DINCAE [12]: A novel CNN-based inpainting method tailored for SST, DINCAE learns spatial and temporal features by integrating an innovative error estimation strategy. This enables the method to reconstruct missing data with high precision and robustness.

5) Phy_INN [16]: A novel GAN-based SST inpainting method, Phy_INN incorporates Atrous Spatial Pyramid Pooling (ASPP), a technique commonly used in semantic segmentation, to learn multi-scale representations. These representations are used to predict both the weekly averages



(a) Ground Truth (b) Average SST (c) Corrupted SST (d) *w/* CNN (e) *w/* ViT (f) SVIFNN

Fig. 4. Comparative analysis of the visualization results from CNN, ViT, and SVIFNN on the NSOAS dataset. The rows are missing ratios of 8%, 25%, 46%, and 68% from top to bottom. (a) Ground truth; (b) Average SST; (c) Cloud-obscured images; (d) Results from ***w/* CNN**; (e) Results from ***w/* ViT**; (f) Results from **SVIFNN**.

and daily details, which are then deeply fused to achieve high-quality SST inpainting.

The comparison results in Table II clearly demonstrate the superior performance of our proposed method across all experimental scenarios. Among the competing methods, Phy_INN, as an innovative DNN-based approach utilizing a two-stage completion strategy, achieved the second-best performance. However, SVIFNN outperformed Phy_INN significantly, with a maximum reduction in $RMSE$ of 42.9% under a 25% missing ratio and N/S ratio of 0.1, and a maximum improvement in R^2 of 21.8% under a 68% missing ratio and N/S ratio of 0.3, further validating the effectiveness of SVIFNN. Moreover, the experimental results highlight that SVIFNN exhibits better performance than other SOTA methods as data missing rates increase. Specifically, under a 68% data coverage rate, SVIFNN significantly outperformed widely used methods like DINEOF and DINCAE. In the same condition, compared to Phy_INN, the method achieving the second-best results, SVIFNN improved R^2 by 18.1%, 19.3%, and 21.8%, and reduced $RMSE$ by 22.6%, 28.8%, and 23.8% for N/S ratios of 0.1, 0.2, and 0.3, respectively. These findings underscore the robustness of SVIFNN in handling large-scale data gaps.

Fig. 3 presents the visual comparison of our method and baseline methods under an N/S ratio of 0.1, which confirms that SVIFNN achieves superior completion results compared to other SOTA methods, particularly in scenarios involving extensive data gaps. This advantage is especially evident in the visualized images, further demonstrating the effectiveness of SVIFNN.

E. Ablation Study

In this section, to more comprehensively validate the effectiveness of each component and demonstrate that SVIFNN is capable of fully preserving the unconventional anomaly features, we have additionally introduced $SSIM$ and $PSNR$ as evaluation metrics for the ablation study. These metrics aim to assess the preservation of spatial details and the visual consistency of the reconstructed images:

Structural Similarity Index Measure ($SSIM$): This metric ranges from 0 to 1, where higher values indicate greater structural similarity between images.

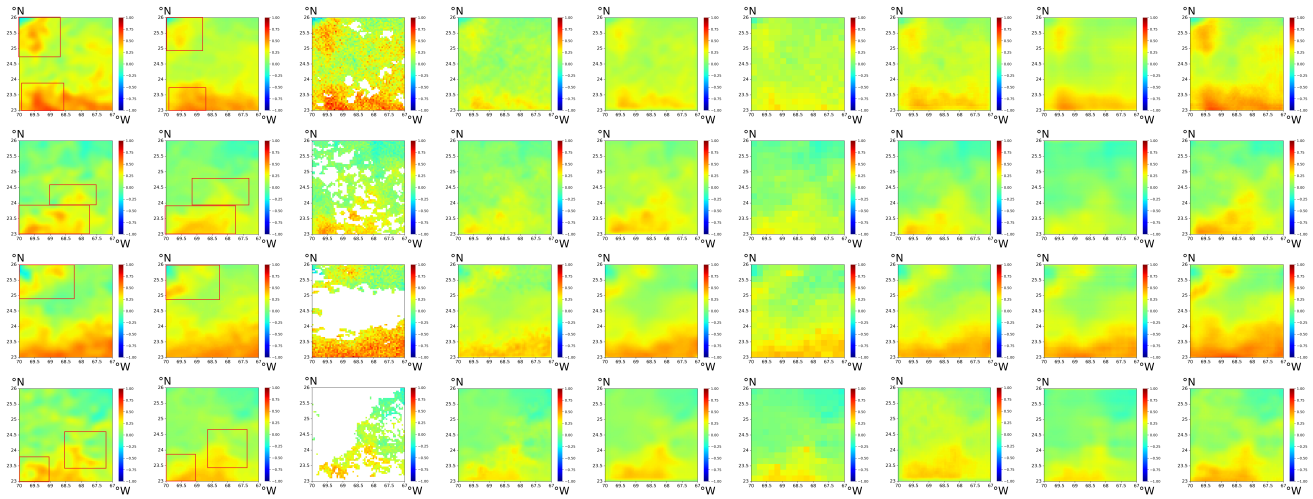
$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{24}$$

here, μ_x and μ_y denote the mean intensities of x and y, while σ_x^2 and σ_y^2 represent their variances. The covariance between x and y is given by σ_{xy}. The constants C_1 and C_2 are included to avoid instability when the denominator is close to zero, where $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$. L denotes the dynamic range of pixel values (commonly $2^{bits\ per\ pixel} - 1$), with $k_1 = 0.01$ and $k_2 = 0.03$ as default values.

Peak Signal-to-Noise Ratio ($PSNR$): This is another widely used metric for evaluating image quality, where higher values reflect better reconstruction accuracy.

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right), \tag{25}$$

where MAX_I is the maximum possible pixel value of the image, and MSE represents the Mean Squared Error between the original image and its reconstruction.



(a) Ground Truth (b) Average SST (c) Corrupted SST (d) CNN *w/o* AES (e) CNN *w/* AES (f) ViT *w/o* AES (g) ViT *w/* AES (h) Ours *w/o* AES (i) Ours

Fig. 5. Comparative analysis of the visualization results to ***w/o*** or ***w/*** **AES** in ***w/*** **CNN**, ***w/*** **ViT**, and SVIFNN on the NSOAS dataset. The rows are missing ratios of 8%, 25%, 46%, and 68% from top to bottom, while N/S is all 0.1. (a) Ground truth; (b) Average SST; (c) Corrupted images; (d)-(i) Results from ***w/*** **CNN** ***w/o*** **AES**, ***w/*** **CNN** ***w/*** **AES**, ***w/*** **ViT** ***w/o*** **AES**, ***w/*** **ViT** ***w/*** **AES**, **SVIFNN** ***w/o*** **AES**, **SVIFNN**.

1) Ablation experiment on each component of SVIFNN:

To validate the effectiveness of the FNO, the AES, and the fusion mechanism in the frequency domain (i.e., FL) within SVIFNN, we have designed the following three variants:

w/o **FNO**: We removed the FNO component from SVIFNN to validate its effectiveness. Specifically, we eliminated the $FFT(\cdot)$ operations, the $R_*(\cdot;\theta_*)$ operation in Equations (7) and (8), as well as the $FFT^{-1}(\cdot)$ operation in Equation (10) (i.e., removing the FNO-related components). Consequently, the inputs $Sign_{sta}$ and $Sign_{ano}$ in Equation (9) were replaced with $\hat{Emb}_{sta}$ and $\hat{Emb}_{ano}$.

w/o **AES**: To validate the effectiveness of the AES component, we removed it from the SVIFNN. Note that, to adhere to the principle of controlled variables and accurately assess its effectiveness, Equation (6) was replaced with Equation (3) in this experiment to achieve the removal of the AES component.

w/o **FL**: The Fusion Layers (FL) is a key component of the Frequency Domain Feature Extraction Module, which aims to fuse $Sign_{ave}$ and $Sign_{ano}$ in the frequency domain, as shown in Equation (9). Therefore, we removed the operations related to Equation (9) as a variation to validate the effectiveness of FL in this method.

Table III presents the experimental comparison results of ***w/o*** **FNO**, ***w/o*** **AES**, ***w/o*** **FL**, and SVIFNN under four missing ratio conditions with an N/S ratio of 0.1, effectively validating the importance of these three components. Notably, the results indicate that AES provides the most significant performance improvement to SVIFNN, followed by FNO and FL. Specifically, in the comparison between SVIFNN and ***w/o*** **AES**, SVIFNN achieves average improvements in $RMSE$, R^2, $SSIM$, and $PSNR$ of 14.7%, 5.2%, 19.2%, and 5.3%, respectively. Similarly, in the comparison between SVIFNN and ***w/o*** **FNO**, SVIFNN demonstrates average improvements of 9.7%, 4.7%, 17.3%, and 3.5% in the same metrics. In the comparison between SVIFNN and ***w/o*** **FL**, the average improvements in $RMSE$, R^2, $SSIM$, and $PSNR$ are 7.1%, 3.9%, 14.1%, and 2.2%, respectively.

2) Ablation experiment on the representation operators:

To verify that the frequency domain representation operator (i.e., FNO) used in SVIFNN is more effective than traditional spatial domain operators (i.e., CNN and ViT) in capturing key feature information in the frequency domain and improving the accuracy and quality of SST SVI completion, we designed the following variants:

w/ **CNN**: In our modification, we substituted the feature extraction component in the Fourier frequency domain of the SVIFNN model with a CNN as described in [50]. The rest of the model's configuration remained unchanged. The CNN, a widely recognized feature extraction tool, primarily extracts spatial local correlations through convolution operations in the spatial domain. For the purposes of our study, we simplified the CNN to consist of a single convolutional layer and excluded the final fully connected layer to better align with the specific requirements of our task.

w/ **ViT**: Similarly, we replaced the feature extraction component in the Fourier frequency domain of the SVIFNN model with the Vision Transformer (ViT) as introduced in [51]. The ViT, a visual adaptation of the Transformer architecture, is designed to capture spatial global correlations using an attention mechanism in the spatial domain. In our adaptation, we also removed the final MLP head from the ViT [51] to better tailor it to our specific task requirements.

Table IV presents the experimental comparison results between ***w/*** **CNN**, ***w/*** **ViT**, and SVIFNN. It can be observed that SVIFNN outperforms both variations in all scenarios, validating the effectiveness of using FNO compared to traditional spatial domain operators (i.e., CNN and ViT). Specifically, in the comparison between ***w/*** **CNN** and SVIFNN, the latter achieves average improvements in $RMSE$, R^2, $SSIM$, and $PSNR$ by 14.5%, 1.8%, 6.7%, and 2.9%, respectively, while in the comparison with ***w/*** **ViT**, the average improvements across these metrics are 10.8%, 1.6%, 16.8%, and 6.7%, respectively. It is noteworthy that ***w/*** **CNN** achieves better

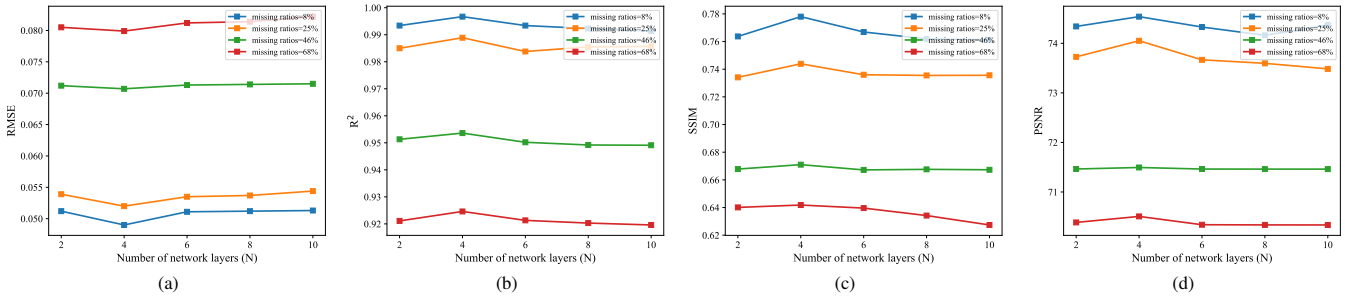


Fig. 6. The ablation analysis of network layers N with a fixed channel number C=64 and N/S=0.1. (a)-(d) represent the impact of different values of N on the metrics $RMSE$, R^2, $SSIM$, and $PSNR$, respectively, under missing rates of 8%, 25%, 46%, and 68%.

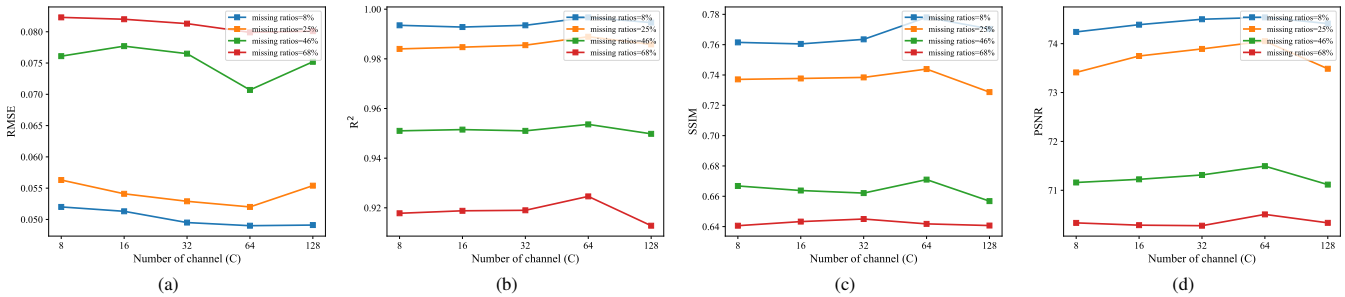


Fig. 7. The ablation analysis of channel number C with a fixed network layer N=4 and N/S=0.1. (a)-(d) represent the impact of different values of c on the metrics $RMSE$, R^2, $SSIM$, and $PSNR$, respectively, under missing rates of 8%, 25%, 46%, and 68%.

$SSIM$ and $PSNR$ scores compared to ***w/* ViT**, while the latter outperforms ***w/* CNN** in $RMSE$ and R^2. This phenomenon can be explained by the inherent strengths of each architecture: CNN, with its strong inductive bias for local feature extraction, excels at capturing fine structural details, which directly benefits metrics like $SSIM$ and $PSNR$ that emphasize structural similarity and signal quality. In contrast, ViTs self-attention mechanism is designed to process global information, making it better suited for capturing long-range dependencies and overall data consistency, which leads to better performance in $RMSE$ and R^2, metrics that assess overall reconstruction accuracy. This observation is further validated by the visual comparison in Fig. 4, which also highlights that SVIFNN achieves superior completion performance compared to both ***w/* CNN** and ***w/* ViT**.

3) Further ablation experiment on the AES module:

To further validate that the AES module can effectively preserve the unconventional anomaly features in daily SST images, we conducted an additional ablation study by further removing the AES module from the previously designed ***w/* CNN** and ***w/* ViT** models. The specific configurations are as follows: ***w/* CNN *w/o* AES** indicates that the AES module is removed from the ***w/* CNN** configuration, while ***w/* CNN *w/* AES** retains the AES module. Similarly, ***w/* ViT *w/o* AES** and ***w/* ViT *w/* AES** follow the same pattern. Here, the removal operation is consistent with the procedure described earlier for ***w/o* AES**.

The comparative visualization results in Fig. 5 clearly demonstrate that: 1) In all cases, these three models achieved better repair results using the AES module (i.e., ***w/* AES**) (as shown in columns e, g, and i), effectively preserving anomalous features in daily SST images (as indicated by the red box in the first column), while the images completed by the three models (as shown in columns d, f, and h) without the AES module (i.e., ***w/o* AES**) can only satisfy the mean distribution (as indicated by the red box in the second column); 2) The results of this ablation study further confirm that the inpainting outcomes of SVIFNN surpass those of ***w/* CNN** and ***w/* ViT**; 3) Additionally, the results of this comparative experiment underline the effectiveness of our approach, showing that even when employing traditional spatial domain representation operators, the inpainting results are notably commendable.

4) Ablation experiment on hyperparameters:

To investigate the impact of the number of layers N (highlighted in green and red font in Fig. 2) and the number of feature channels C (i.e., $H \times W \times C$) on model performance, we conducted ablation experiments. Specifically, we employed $RMSE$, R^2, $SSIM$, and $PSNR$ as evaluation metrics, analyzing the effects of different values of N and C at missing rates of 8%, 25%, 46%, and 68%. The values for N were set to 2, 4, 6, 8, and 10, while C took values of 8, 16, 32, 64, and 128. The final results are presented in Fig. 6 and 7. Notably, to control variables effectively, we fixed C=64 when examining the effect of N on model performance, and fixed N=4 when assessing the impact of C. The results in Fig. 6 indicate that: 1) the model with N=4 achieved relatively better performance; 2) at higher missing rates (such as 46% and 68%), the influence of the number of layers on model performance is not particularly pronounced, likely due to the

limited amount of learnable information. Furthermore, Fig. 7 shows that the best performance of the model occurs when the number of feature channels C is set to 64.

V. Conclution

Scientific visualization images (SVI) represent a critical resource for scientific analysis. However, due to their distinct visual properties compared to conventional images, traditional semantic modeling-based methods struggle to effectively understand and analyze them. In this study, we address the completion task of SST SVI by designing a novel inpainting method tailored for SST SVI, named SVIFNN. Firstly, the model employs a twin-stream mechanism to simultaneously capture stability and anomaly information. Particularly, it enhances the saliency of anomaly information through a designed reverse attention mechanism. Secondly, by introducing frequency neural operators (FNO), the model enhances the effectiveness of feature representation for SVI of complex marine systems, achieving an effective integration of frequency domain and spatial domain representations. Comparative experiments demonstrate significant improvements in accuracy and robustness, for example in R^2 (18.1%, 19.3%, and 21.8%) and reductions in $RMSE$ (22.6%, 28.8%, and 23.8%) when addressing extensive data gaps (68% missing rate). Furthermore, comprehensive ablation studies confirm the critical role of incorporating the FNO, the AES, and FL. These studies also validate the superiority of FNO over traditional spatial domain operators (i.e., CNN and ViT), in this specific task. Future research will continue to explore the unique characteristics of oceanographic data, advancing its perception, understanding, and predictive capabilities from a scientific perspective.

References

[1] M. Head, P. Luong, J. Costolo, K. Countryman, and C. Szczechowski, "Applications of 3-d visualizations of oceanographic data bases," in *Oceans '97. MTS/IEEE Conference Proceedings*, vol. 2, 1997, pp. 1210–1215 vol.2.

[2] F. Nyffeler, J. L. Casamor, and L. Tacher, "Three dimensional visualization of large oceanic data sets: An application of the earthvision (r) package to the data gathered during the mast ii/flubal cruise." *MTP News*, pp. 21–23, 12 1995.

[3] H. Sun, J. Ma, Q. Guo, Q. Zou, S. Song, Y. Lin, and H. Yu, "Coarse-to-fine task-driven inpainting for geoscience images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7170–7182, 2023.

[4] Y. Du, J. He, Q. Huang, Q. Sheng, and G. Tian, "A coarse-to-fine deep generative model with spatial semantic attention for high-resolution remote sensing image inpainting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[5] J. Liu, M. Gong, Z. Tang, A. K. Qin, H. Li, and F. Jiang, "Deep image inpainting with enhanced normalization and contextual attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6599–6614, 2022.

[6] D. Zhao, L. Xu, L. Ma, J. Li, and Y. Yan, "Pyramid global context network for image dehazing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3037–3050, 2021.

[7] Y. Ding, M. Li, T. Yan, F. Zhang, Y. Liu, and R. W. H. Lau, "Rain streak removal from light field images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 467–482, 2022.

[8] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7757–7765.

[9] S. Ouala, C. Herzet, and R. Fablet, "Sea surface temperature prediction and reconstruction using patch-level neural network representations," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 5628–5631.

[10] S. Zhang, Y. Yang, K. Xie, J. Gao, Z. Zhang, Q. Niu, G. Wang, Z. Che, L. Mu, and S. Jia, "Spatial-temporal siamese convolutional neural network for subsurface temperature reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2023.

[11] Z. Chen, Y. Luo, Y. Chen, J. Wang, D. Li, K. Gao, C. Wang, and J. Li, "Brgan: Blur resist generative adversarial network with multiple joint dilated residual convolutions for chlorophyll color image restoration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[12] A. Barth, A. Alvera-Azcrate, M. Licer, and J. M. Beckers, "Dincae 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations," *Geoscientific Model Development*, vol. 13, no. 3, pp. 1609–1622, 2020.

[13] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, "Inpainting of remote sensing sst images with deep convolutional generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 173–177, 2019.

[14] S. Shibata, M. Iiyama, A. Hashimoto, and M. Minoh, "Restoration of sea surface temperature satellite images using a partially occluded training set," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2771–2776.

[15] N. Hirahara, M. Sonogashira, and M. Iiyama, "Cloud-free sea-surface-temperature image reconstruction from anomaly inpainting network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[16] Q. Wei, Z. Zuo, J. Nie, J. Du, Y. Diao, M. Ye, and X. Liang, "Inpainting of remote sensing sea surface temperature image with multi-scale physical constraints," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 492–497.

[17] S. Chen, L. Zhang, and L. Zhang, "Cross-scope spatial-spectral information aggregation for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 33, pp. 5878–5891, 2024.

[18] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier Neural Operator for Parametric Partial Differential Equations," *arXiv e-prints*, p. arXiv:2010.08895, Oct. 2020.

[19] Y. Zhang and D. Pan, "Improving remote sensing ocean color using a polynomial-based interpolation method," *Journal of Applied Remote Sensing*, vol. 11, no. 1, p. 015034, 2017.

[20] J. Wang and H. Shi, "Reconstruction of missing remote sensing ocean color data using polynomial interpolation," *Journal of Coastal Research*, vol. 85, no. sp1, pp. 1206–1210, 2018.

[21] R. Ma and D. Wang, "A polynomial-based interpolation method for reconstructing missing satellite-derived ocean color data," *Remote Sensing*, vol. 11, no. 18, p. 2141, 2019.

[22] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM National Conference*, ser. ACM '68. New York, NY, USA: Association for Computing Machinery, 1968, p. 517524. [Online]. Available: https://doi.org/10.1145/800186.810616

[23] Z. Sun, C. Mao, and Y. Liu, "Statistical interpolation of sea surface temperature fields using a nonhomogeneous regression model," *Remote Sensing*, vol. 10, no. 12, p. 1919, 2018.

[24] Y. Chen and S. Zhang, "Statistical interpolation of sea surface temperature using multiple regression models," *Remote Sensing*, vol. 11, no. 2, p. 169, 2019.

[25] Y. Zhu, E. L. Kang, Y. Bo, Q. Tang, J. Cheng, and Y. He, "A robust fixed rank kriging method for improving the spatial completeness and accuracy of satellite sst products," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 9, pp. 5021–5035, 2015.

[26] H.-K. Jeon and H. Y. Cho, "Comparative study on gap-filling of goci-i chlorophyll-a product using kriging and random forest," in *ISRS 2022 (International Symposium on Remote Sensing 2022)*, 2022, Conference.

[27] N. Cressie and G. Johannesson, "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, no. 1, pp. 209–226, 01 2008. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2007.00633.x

[28] J. Ahn and Y. Lee, "Spatial gap-filling of gk2a daily sea surface temperature (sst) around the korean peninsula using meteorological data and regression residual kriging (rrk)," *Remote Sensing*, vol. 14, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/20/5265

[29] Y. Zhou, Y. Gai, and J. Li, "Research on sea surface temperature reconstruction from long-term modis data," *IOP Conference Series: Earth and Environmental Science*, vol. 631, no. 1, p. 012032, jan 2021. [Online]. Available: https://dx.doi.org/10.1088/1755-1315/631/1/012032

[30] X. Liu and M. Wang, "Gap filling of missing data for viirs global ocean color products using the dineof method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4464–4476, 2018.

[31] J. Li, W. Sun, and J. Zhang, "Infrared sea surface temperature data reconstruction using dineof method," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 7107–7110.

[32] B. Ping, F. Su, and Y. Meng, "Reconstruction of satellite-derived sea surface temperature data based on an improved dineof algorithm," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 8, pp. 4181–4188, 2015.

[33] Y. Wang, Z. Gao, and D. Liu, "Multivariate dineof reconstruction for creating long-term cloud-free chlorophyll-a data records from seawifs and modis: A case study in bohai and yellow seas, china," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 5, pp. 1383–1395, 2019.

[34] Y. Miyazawa, H. Murakami, T. Miyama, S. M. Varlamov, X. Guo, T. Waseda, and S. Sil, "Data assimilation of the high-resolution sea surface temperature obtained from the aqua-terra satellites (modis-sst) using an ensemble kalman filter," *Remote Sensing*, vol. 5, no. 6, pp. 3123–3139, 2013. [Online]. Available: https://www.mdpi.com/2072-4292/5/6/3123

[35] R. Lguensat, P. Tandeo, R. Fablet, and R. Garello, "Spatio-temporal interpolation of sea surface temperature using high resolution remote sensing data," in *2014 Oceans - St. John's*, 2014, pp. 1–4.

[36] X. Liang, M. Losch, L. Nerger, L. Mu, Q. Yang, and C. Liu, "Using sea surface temperature observations to constrain upper ocean properties in an arctic sea ice-ocean data assimilation system," *Journal of Geophysical Research: Oceans*, vol. 124, no. 7, pp. 4727–4743, 2019. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015073

[37] G. Korotaev, E. Huot, F. L. Dimet, I. Herlin, S. Stanichny, D. M. Solovyev, and L. Wu, "Retrieving ocean surface current by 4-d variational assimilation of sea surface temperature images," *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1464–1475, 2008. [Online]. Available: https://dx.doi.org/10.1016/J.RSE.2007.04.020

[38] A. Pasula and D. N. Subramani, "4d-var data assimilation of sea surface temperature in a regional model of the andaman sea," in *OCEANS 2022, Hampton Roads*, 2022, pp. 1–6.

[39] Z. Li, W. Peng, Z. Yuan, and J. Wang, "Long-term predictions of turbulence by implicit u-net enhanced fourier neural operator," *Physics of Fluids*, vol. 35, no. 7, p. 075145, 05 2023. [Online]. Available: https://dx.doi.org/10.1063/5.0158830

[40] W. . Peng, Z. . Yuan, and J. . Wang, "Attention-enhanced neural network models for turbulence simulation," *Physics of Fluids*, vol. 34, no. 2, p. 025111, 02 2022. [Online]. Available: https://doi.org/10.1063/5.0079302

[41] P. Surapaneni, "Neural network prediction of ocean wave behavior using frequency domain mapping," in *OCEANS 2023 - MTS/IEEE U.S. Gulf Coast*, 2023, pp. 1–6.

[42] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar, "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators," *arXiv e-prints*, p. arXiv:2202.11214, Feb. 2022.

[43] Z. . Li, W. . Peng, Z. . Yuan, and J. . Wang, "Long-term predictions of turbulence by implicit U-Net enhanced Fourier neural operator," *Physics of Fluids*, vol. 35, no. 7, p. 075145, 07 2023. [Online]. Available: https://doi.org/10.1063/5.0158830

[44] H. Chen, L. Huang, T. Liu, and A. Ozcan, "Fourier imager network (fin): A deep neural network for hologram reconstruction with superior external generalization," *Light: Science & Applications*, vol. 11, no. 1, p. 254, 2022.

[45] S. Ehlers, M. Klein, A. Heinlein, M. Wedler, N. Desmars, N. Hoffmann, and M. Stender, "Machine learning for phase-resolved reconstruction of nonlinear ocean wave surface elevations from sparse remote sensing data," *Ocean Engineering*, vol. 288, p. 116059, 2023.

[46] J.-M. Beckers and M. Rixen, "Eof calculations and data filling from incomplete oceanographic datasets," *Journal of Atmospheric and Oceanic Technology*, vol. 20, pp. 1839–1856, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:124505891

[47] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep Complex Networks," *arXiv e-prints*, p. arXiv:1705.09792, May 2017.

[48] S. Ji, P. Dai, M. Lu, and Y. Zhang, "Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 732–748, 2021.

[49] M. Ogunsanya, J. Isichei, and S. Desai, "Grid search hyperparameter tuning in additive manufacturing processes," *Manufacturing Letters*, vol. 35, pp. 1031–1042, 2023.

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv e-prints*, p. arXiv:2010.11929, Oct. 2020.

Zijie Zuo (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China.

His research interests are in the fields of artificial intelligence and visual analysis of marine big data.

Jie Nie (Member, IEEE) received the B.S. and Ph.D. degrees from the Ocean University of China, Qingdao, China, in 2002 and 2011, respectively, all in computer science. She was a Visiting Scholar at the University of Pittsburgh, Pittsburgh, PA, USA, from 2009 to 2010. After that, she was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2015 to 2017. She is currently a Professor at the Ocean University of China.

Her research interests are in the fields of artificial intelligence and visual analysis of marine big data.

Xin Wang (Member, IEEE) received the B.E. and first Ph.D. degrees in computer science and technology from Zhejiang University, China, and the second Ph.D. degree in computing science from Simon Fraser University, Canada. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. He has published over 150 high-quality research papers in ICML, NeurIPS, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Knowledge and Data Engineering, ACM KDD, WWW, ACM SIGIR, and ACM Multimedia. His research interests include multimedia intelligence and machine learning and its applications. He won three best paper awards, including ACM Multimedia Asia. He was a recipient of the ACM China Rising Star Award, the IEEE TCMC Rising Star Award, and a DAMO Academy Young Fellow.

Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, in 1993 and 1999, respectively, and Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in November 2003. He joined the Ocean University of China in 2004, where he is currently a Professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision, and underwater image processing.