Journal of
Marine Science and Engineering

MDPI

Article

SICFormer: A 3D-Swin Transformer for Sea Ice Concentration Prediction

Zhuoqing Jiang [1,2], Bing Guo [3], Huihui Zhao [1,*], Yangming Jiang [1] and Yi Sun [2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] College of Mathematics and System Science, Xinjiang University, Urumqi 830017, China
[3] School of Civil Architectural Engineering, Shandong University of Technology, Zibo 255000, China
* Correspondence: zhh@aircas.ac.cn; Tel.: +86-13718105758

Abstract: Sea ice concentration (SIC) is an important dimension for characterising the geographical features of the pan-Arctic region. Trends in SIC bring new opportunities for human activities in the Arctic region. In this paper, we propose a deep learning technology-based sea ice concentration prediction model, SICFormer, which can realise end-to-end daily sea ice concentration prediction. Specifically, the model uses a 3D-Swin Transformer as an encoder and designs a decoder to reconstruct the predicted image based on PixelShuffle. This is a new model architecture that we have proposed. Single-day SIC data from the National Snow and Ice Data Center (NSIDC) for the years 2006 to 2022 are utilised. The results of 8-day short-term prediction experiments show that the average Mean Absolute Error (MAE) of the SICFormer model on the test set over the 5 years is 1.89%, the Root Mean Squared Error (RMSE) is 5.99%, the Mean Absolute Percentage Error (MAPE) is 4.32%, and the Nash–Sutcliffe Efficiency (NSE) is 0.98. Furthermore, the current popular deep learning models for spatio-temporal prediction are employed as a point of comparison given their proven efficacy on numerous public datasets. The comparison experiments show that the SICFormer model achieves the best overall performance.

Keywords: spatiotemporal prediction; 3D-Swin Transformer; sea ice concentration; attention mechanisms

Citation: Jiang, Z.; Guo, B.; Zhao, H.; Jiang, Y.; Sun, Y. SICFormer: A 3D-Swin Transformer for Sea Ice Concentration Prediction. *J. Mar. Sci. Eng.* **2024**, *12*, 1424. https://doi.org/10.3390/jmse12081424

Academic Editor: Anatoly Gusev

Received: 15 June 2024
Revised: 13 August 2024
Accepted: 14 August 2024
Published: 17 August 2024

1. Introduction

In recent years, the geography of the pan-Arctic region has undergone profound changes as global temperatures have risen, one manifestation of which has been the melting of sea ice [1,2], which has created new opportunities and challenges for humankind, a process that will continue for the foreseeable future. A new study suggests that the Arctic may experience an "ice-free" summer by 2030 [3]. Sea ice concentration is important for understanding the geography of the Arctic. It affects the global climate and creates new opportunities for human activities in the Arctic, such as shipping, resource exploration, and scientific expeditions [3,4]. Arctic shipping lanes can greatly shorten the distance of international shipping lanes, reduce trade costs, and further promote the prosperous development of international trade. Some studies have shown that by 2030, the opening cycle of Arctic shipping lanes may be raised from the current four months to more than half a year [4]. In general, SIC is an important factor influencing the availability of suitable operational conditions in shipping lanes. Consequently, the short-term prediction of sea ice concentration can provide valuable support for the dynamic planning of short-term navigation in Arctic shipping lanes. The utilisation rate of the waterway can be enhanced while maintaining the safety of ship navigation.

Currently, studies related to predicting sea ice concentration can be broadly classified into three categories: numerical simulation methods, statistical modelling-based methods, and deep learning methods. Numerical simulation methods usually use differential techniques to solve equations for sea ice system dynamics and thermodynamics, such as the

MIT General Circulation Model (MITgcm), the Hybrid Coordinate Ocean Model (HYCOM), and the Modular Ocean Model (MOM6) [5–7]. Satellite observation assimilation techniques can greatly improve the prediction accuracy of numerical simulation methods, such as the Global Ice Ocean Prediction System (GIOPS), the Arctic Cap Nowcast/Forecast System (ACNFS), the Arctic Ice Ocean Prediction System (ArcIOPS), and the Sea Ice Seasonal Prediction System (SISPS) [8–12]. However, numerical simulation-based methods have to take into account various types of environmental information, such as that related to land, ocean, and climate, at the same time, and obtaining high-quality information is also a challenging task; as more information is used, the errors within the collected information itself will also be larger. More importantly, numerical simulation-based methods require the computational resources of large-scale central processing unit (CPU) computing clusters, which cannot make predictions quickly. In terms of statistical models, a vector autoregressive model (VAR) [13], a linear Markov model [14], and a vector Markov model [15] are statistical models that establish a statistical relationship between sea ice concentration and the atmospheric environment, ocean environment, and other characteristic variables of sea ice in predicting the dependent variable of sea ice concentration. Statistical models do not require a large number of computational resources compared to numerical simulation methods, but it is more difficult for them to capture nonlinear relationships among variables, and they rely on existing experience. Purely data-driven deep learning techniques come without the need for data other than historical sea ice concentration information, and prediction using trained parameters often only takes a few seconds. In recent years, deep learning methods have gradually been applied to sea ice concentration prediction.

Deep learning techniques, which use artificial neural networks as the basic unit and can easily learn nonlinear relationships from massive data, have been successfully applied in the field of geosciences to help researchers understand scientific problems from a new perspective [16–23]. The existing deep learning based spatio-temporal prediction techniques generally contain one or more of the three structures of convolutional neural network (CNNs) [24], Recurrent Neural Network (RNNs) [25], and Attention Mechanisms [26]. They can be roughly categorised into four groups in terms of the types of networks used and their constituent structures [27]. The first is the RNN-RNN-RNN structure, which generally has better flexibility and accuracy and thus is often used as a baseline model for spatio-temporal prediction tasks. The main models of this kind are MIM-LSTM and PredRNN [28,29]. The second is the CNN-RNN-CNN structure, which relies on an RNN to capture temporal features and a CNN to capture spatial features, combining the advantages of both. The representative models of this structure are ConvLSTM, VRNN, PhyDNet, and so on [30–32]. The third kind is CNN-ViTs-CNN ViTs, referring to various types of attention mechanisms that have evolved on the basis of Vision Transformers [33], including the work of ViViT, TimesFormer, and MViT [34–36]. The fourth one is CNN-CNN-CNN structures, such as PredCNN, DPG, and so on [37,38]. In conclusion, spatio-temporal prediction based on deep learning techniques can satisfy the end-to-end prediction needs, can quickly realise the prediction work, and can also capture the nonlinear relationships in the time series of the prediction target, without requiring additional variable inputs.

Sea ice concentration prediction can be regarded a spatio-temporal prediction problem, and researchers have also tried to apply various types of neural network models to sea ice concentration prediction tasks. Chi and other researchers performed work related to sea ice concentration prediction using a simple stacked deep neural network (DNN) in 2017 [39]. A two-stream ConvLSTM (TS-ConvLSTM) model with a new perceptual loss function was proposed in 2021. The model combines two different scales of ConvLSTM to capture the sea ice features at multiple scales of sea ice concentrations to predict the monthly sequence of sea ice concentration with good results [40]. In 2020, Kim used a convolutional neural network for monthly sea ice concentration predictions [41]. Andersson proposed probabilistic deep learning (PDL) in 2021 to predict the seasonal mean data on sea ice concentration using 50 variables, including oceanic factors, climatic factors, and sea ice factors [42]. The impact of using different input data on the prediction results was

explored. Most of the above research works use a more basic network structure and a single prediction structure, failing to fully consider the trend in spatio-temporal series data and also failing to fully explore the edge characteristics of sea ice melting, and the prediction granularity is relatively coarse. Ren and other researchers proposed using the lightweight intelligent prediction model SICNet and predicted the day-by-day sea ice concentration for the next 7 days using the past 7 days of data as the input [43]. A model was designed with an encoder–decoder structure that fused seasonal and trend features, which could predict the day-by-day sea ice concentration and the sea ice thickness within 45 days [44]. These works enable sea ice prediction to be further enhanced at the temporal and spatial scales.

Is it possible to use pure Transformer class models to further improve the prediction accuracy? With the development of deep learning technology, more and more scholars are trying to embed an attention module into the classical CNN or RNN models. A spatio-temporal attention module was embedded into the U-net structure, which verified that the attention mechanism has a unique advantage in learning the change characteristics of sea ice concentration [43]. A Transformer is a type of deep learning that combines the advantages of the attention mechanism. Its variants have achieved the best results in many task scenarios, but there is still a lack of work on predicting sea ice concentration using a pure Transformer as the framework. In light of this, it is possible to think of designing a Vision Transformer (ViT) [33] or variant models as encoders and connecting decoders suitable for spatio-temporal prediction. Could the best spatio-temporal prediction be achieved by retaining the powerful feature extraction capabilities of ViT models and connecting suitable downstream task output modules? This is one of the motivations for the model design in this paper.

We have designed a deep learning model based on an encoder–decoder architecture that can handle sequence prediction tasks end to end, which was named SICFormer. We maintain the resolution of the data at 448 × 304 and include year-round data in the training to reduce the possible errors in the data itself, retain the most information about the data as possible, and theoretically train the model with the year-round data to allow it to capture seasonal features as well. Our approach achieves end-to-end prediction without using data on anything other than sea ice. The aim of our research is to provide short-term predictions of the sea ice concentration in the pan-Arctic region, which could help ships to dynamically adjust their short-term voyage plans.

2. Methods

2.1. Data

The data used in this paper are from the National Snow and Ice Data Center (NSIDC). Their spatial resolution is 25 km × 25 km. The purpose of this paper is to make short-term forecasts for the next 8 days. In general, short-term forecasts are more relevant to recent data. To avoid the noise caused by early data and considering a certain sample size, daily SIC data from 2006 to 2022, totalling 6209 days, were selected. We use the complete raw imagery provided by the NSIDC, covering all 448 × 304 rasters, which increases the computational effort of the model but ensures that the information in the range is complete. Figure 1 is an example image of the sea ice concentration.

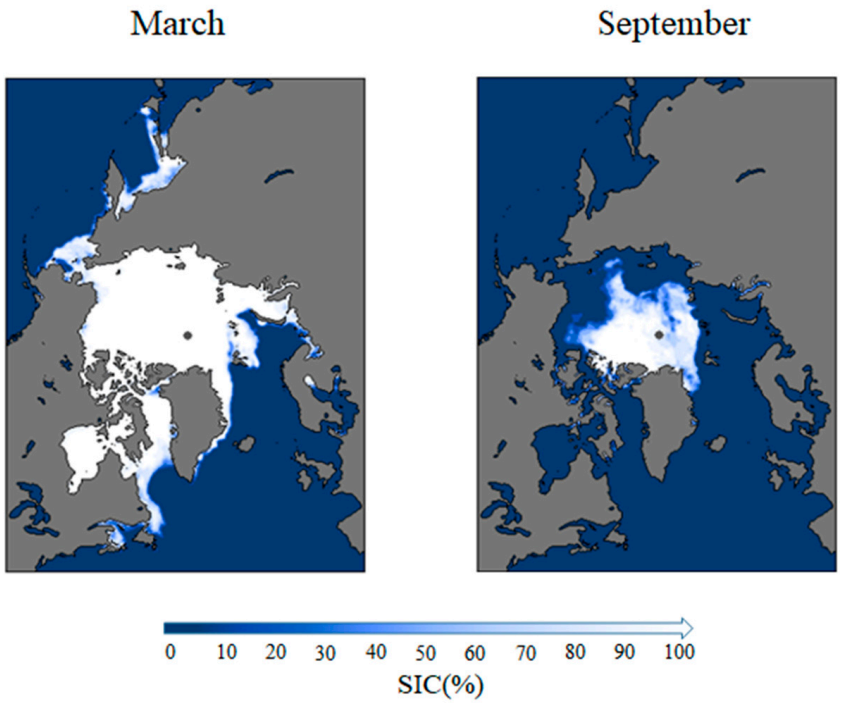

Figure 1. 1 March 2022 on the left; 1 September 2022 on the right.

2.2. Problem Definition

A multidimensional time series can be treated as a spatio-temporal sequence if multiple values at a point in time have certain spatial relationships, i.e., the relative positions of the variables affect the magnitude of the values. For example, each frame of video data is not only affected by historical frames but there is also a spatial correlation between different pixels. The spatio-temporal prediction problem is to predict the future sequence from the given historical spatio-temporal sequence, assuming that the past T frames are given at moment t, denoted as $X_{t,T} = \{x_i\}_{t-T+1}^{t}$, and our goal is to predict the future T' frames at moment t, denoted as $Y_{t,T'} = \{x_i\}_{t+1}^{t+T'}$, where $x_i \in R^{C\times H\times W}$ is the picture with channel C, height H, and width W. That is, we want to find a function $F_\Theta : X_{t,T} \mapsto Y_{t,T'}$ where Θ is learnable for all parameters. See Equation (1).

$$\Theta^* = arg\min_{\Theta} \mathcal{L}\left(\mathcal{F}_\Theta(X_{t,T}), Y_{t,T'}\right), \tag{1}$$

where $\mathcal{L}$ can be any kind of loss function, and we use the MSE loss function. The essence of learning is to find all the parameters that minimise the loss function.

2.3. The Overall Structure

The key to sea ice concentration prediction lies in extracting the spatial and temporal dependencies involved and mapping this relationship to the output according to a specific function. Therefore, it is necessary to construct a deep learning model that has a strong feature learning capability and can reflect the learned weights in the prediction results. Our proposed SICFormer first preprocesses the input image, and then the encoder module computes the attention scores of the features at different scales. Then, the intermediate module performs the convolution operation on the attention scores obtained in the previous step, and the image features can be made more stable after the convolution operation, which ensures the invariance of the features. Next, the decoder performs an upsampling operation to reconstruct the image features at the initial resolution. We also design a global shortcut that preserves the original image features and feeds them directly into the last module, which helps us to recover the lost spatio-temporal information and enables the model

to make better use of contextual information, as shown by the orange add operation in Figure 2.

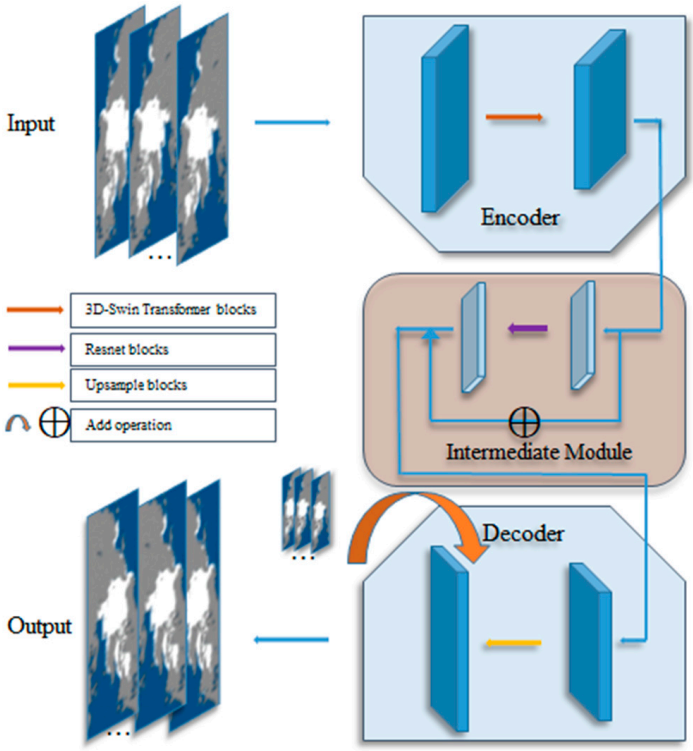

Figure 2. Overall model architecture.

Finally, by simply adjusting the shape of the image and outputting the image, we obtain the prediction result we want. SICFormer is a spatio-temporal prediction model with a ViTs-CNN-CNN structure which is clear and easy to understand; see Figure 2. The model consists of five parts, which are the input, the encoder, the intermediate module, the decoder, and the output.

2.4. The Input

In this paper, we use the SIC sequences for the past 8 days to predict the SIC sequences for the next 8 days, and the data used are a single-channel raster image, so the shape of the input data is $8 \times 1 \times 448 \times 304$. We use a normalisation method to normalise the data; see Equation (2).

$$X = \frac{DATA - DATA_{min}}{DATA_{max} - DATA_{min}}, \tag{2}$$

where $DATA_{min}$ and $DATA_{max}$ denote the minimum and maximum values for sea ice concentration, respectively.

2.5. The Encoder

The excellent performance of Transformers [45] in the field of natural language proves that the attention mechanism can learn global features and fuse them effectively. Based on this, many Transformer-based models have been constructed in the field of computer vision to deal with visual and multimodal tasks, and ViTs are one of the representative works.

The structure of a ViT strictly adheres to the architectural design of a Transformer, and its essence is the Transformer encoder. The ViT first divides the images into patches, which are equivalent to "tokens" in the Transformer. The patches are then flattened in the channel direction, and the channel is mapped to a predefined value using a linear embedding operation. At the same time, the model also adds position vectors to the

patches to preserve the spatial information of the images, and the subsequent process is consistent with the Transformer. Although the ViT achieves the best results on tasks such as image classification, its huge number of parameters and computational effort are still prohibitive. Assuming a total of $h \times w$ patches, its computational complexity in terms of self-attention is $(hw)^2$, which is not friendly enough for a pixel-level task such as sea ice concentration prediction, especially since we chose to take the NSIDC raw image size as the input. To solve this problem, we thought of the Swin Transformer, which greatly reduces the computational effort of the ViT while preserving global and local spatial features.

The Swin Transformer [46] makes two main improvements over the ViT. Firstly, it processes images through a hierarchical construction method similar to that of convolutional neural networks so that it is able to process images at different scales and extract features at different scales. Second, the Swin Transformer uses the concept of a window attention mechanism. It divides all the patches into $\frac{hw}{M^2}$ windows. M is the size of the window, each window contains M^2 patches, and the next attention computation is limited to the window. The computational complexity of the same image of size $h \times w$ patches becomes $M^2 * (hw)$, which is reduced from the square relationship of the number of patches to a linear relationship. While windows can reduce the computational complexity, they also cause a new problem, which is that patches that are not in the same window cannot compute attention, meaning that connections cannot be made without being in a window. For this reason, the Swin Transformer introduces the sliding window mechanism, which shifts each window to the lower right by a distance of $\frac{M}{2}$ patches to form a new window, which solves the problem of different windows not being able to communicate. Nevertheless, the sliding of the windows results in an increase in the total number of windows. To address this issue, a technique known as cyclic shifting is employed to restore the number of windows to its original state prior to the sliding window by repositioning and consolidating some of the newly formed windows. Subsequently, the attention score within the window is calculated. The images are returned to their original sequence once the calculation has been completed. Since the self-attention module ignores the positional information of the patches involved in the computation, the Swin Transformer uses relative positional bias information to solve this problem. The attention mechanism can be written in the form of Equation (3).

$$\text{Attention}(Q, K, V) = \sigma(\frac{QK^T}{\sqrt{d}} + B)V, \tag{3}$$

where Q, K, and V are the matrices to be learned, and σ denotes the nonlinear activation function. $B \in R^{M^2 \times M^2}$ is the relative position bias, which serves to provide relative position information for different patches within the same window.

The Swin Transformer establishes connections between non-overlapping windows without adding new windows, which takes into account both local and global features and effectively reduces the computational complexity. Our 3D-Swin Transformer [47] extends this operation to the temporal channel to implement a sliding window mechanism in space–time. It helps the model to perform window sliding in the spatial and temporal channels simultaneously and to compute spatial and temporal attention scores simultaneously, thus effectively extracting the features of a particular location at a particular time and its spatio-temporal relationship with neighbouring locations and adjacent times. This is one of the most important parts of the spatio-temporal prediction work. This design is well suited to extracting high-resolution spatio-temporal data like sea ice concentration, as shown in Figure 3

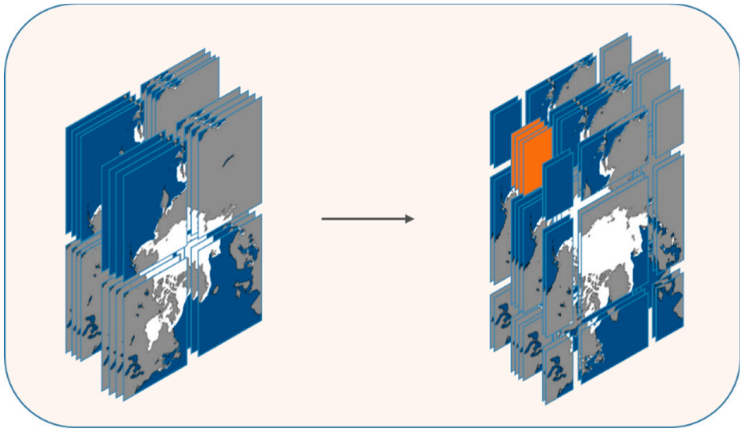

Figure 3. Sliding window mechanism to move the window by an M/2 distance to form a new window.

We form a 3D-Swin Transformer module by combining the two modules of window attention and sliding window attention—see (a) and (b) in Figure 4—with x_{l-1} being the data before they enter the module and x_{l+1} being the output data. This process can also be expressed by Equations (4)–(7).

$$x_l^{'} = 3DWMSA(LN(x_{l-1})) + x_{l-1}, \tag{4}$$

$$x_l = FFN(LN(x_l^{'})) + x_l^{'}, \tag{5}$$

$$x_{l+1}^{'} = 3DSWMSA(LN(x_l)) + x_l, \tag{6}$$

$$x_{l+1} = FFN(LN(x_{l+1}^{'})) + x_{l+1}^{'}, \tag{7}$$

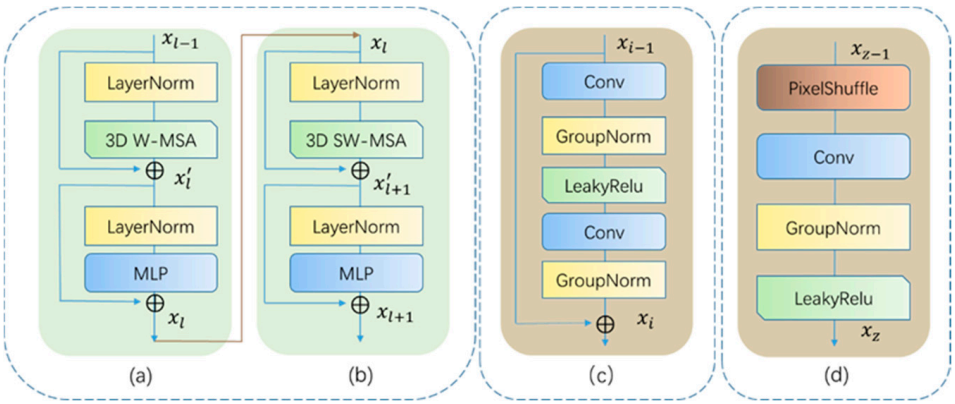

Figure 4. (**a**,**b**) are 3D-Swin Transformer blocks, (**c**) is a ResNet block, and (**d**) is an upsample block.

After one 3D-Swin Transformer module, feature extraction is completed. As previously stated, our encoder employs a hierarchical structure analogous to that of a convolutional neural network (CNN). After a specified number of 3D-Swin Transformer modules, a downsampling operation is conducted with the objective of reducing the image size in order to extract features from the data at varying scales. This downsampling is achieved through patch merging.

The whole encoder repeats this step four times to extract very rich image features from the data, without a downsampling operation during the last pass through the 3D-Swin Transformer module, and the overall process can be seen very clearly in Figure 5.

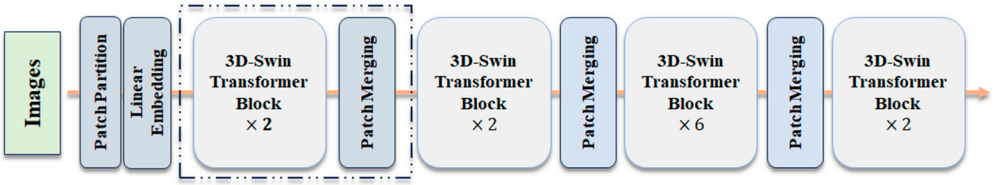

Figure 5. The overall process of the encoder.

2.6. The Intermediate Module

The main role of this part is to take the features extracted by the encoder and make the training more stable using a further convolution operation. Based on the excellent performance of residual networks [48], in this part, we first build a plug-and-play residual block; as shown in (c) of Figure 3, two convolutional layers can ensure that the number of channels remains unchanged while extracting features from the data. The size of the convolution kernel is 3×3. GroupNorm divides the channel dimensions into multiple groups and normalises the features within each group rather than based on the features of the whole batch, as BatchNorm does. It is therefore more stable and effective for small batch training scenarios. In order to avoid possible gradient vanishing phenomena, the LeakyReLU activation function is used instead of the ReLU activation function. The characteristics of these two determine that they are more suitable for prediction tasks. The process can be expressed as follows:

$$x_i = GN(Cov(\sigma(GN(Cov(x_{i-1}))))) + x_{i-1}, \tag{8}$$

where σ denotes the nonlinear activation function.

In this section, a shortcut layer is nested on top of the three residual modules; see Figure 6. The nested network structure is capable of refining the modelling capability and abstracting and transforming the input features in more detail, thereby enhancing the model's ability to capture complex patterns. This constitutes part of the hierarchical nested residual structure put forth in this paper.

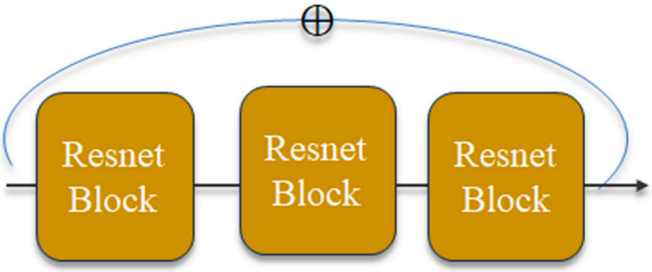

Figure 6. Residual blocks and their nested branch connections.

2.7. The Decoder

The decoder is an upsampling module with PixelShuffle [49] as its core. Its core idea is to rearrange and utilise the channel dimensions of the feature map of a low-resolution image. Its internal operation can be divided into two steps. First, a convolutional layer with a convolutional kernel of 1×1 is used to increase the number of channels of the feature map, while feature fusion and dimensionality reduction are performed. Next, the pixel reorganisation work is carried out, i.e., the elements of each channel are rearranged in the spatial dimension, and the original $H \times W \times C$ is adjusted to $r * H \times r * W \times C/r^2$. In this way, the information that was originally concentrated in the channels is dispersed to the spatial dimension, which achieves the purpose of upsampling, while more information is retained. Compared to methods such as neighbourhood interpolation or bilinear interpolation alone, PixelShuffle can be trained end to end with other convolutional layers, adapting to specific data distributions and task requirements to produce higher-quality images. It is especially good at recovering high-frequency details, thus better capturing sea

ice concentration reconstruction details. In accordance with our prediction task, there is a transformation of temporal information into spatial information, which is particularly conducive to spatio-temporal prediction. Based on this feature, we designed a CNN module after each upsampling so that an upsample block was composed; see (d) in Figure 3. The whole decoder consists of three upsample Blocks, with a CNN block connected at the end; see Figure 7. It should be noted that the data features are summed with the original data at the pixel level before they are fed into the last block, which is actually equivalent to a shortcut that spans the whole model. It is a residual structure which further ensures the stability and generalisation ability of the model.

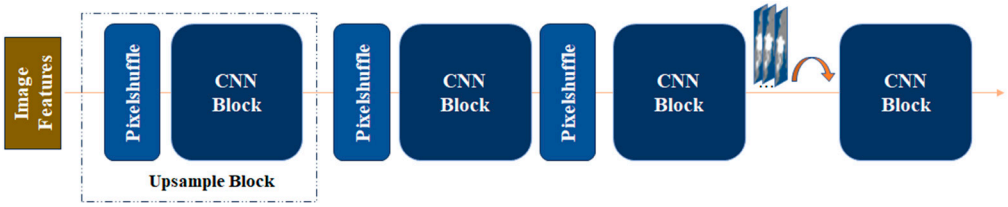

Figure 7. The overall process of the decoder.

The process for an upsample block is shown in Equation (9).

$$x_z = \sigma(GN(Cov(Pixel(x_{z-1})))), \tag{9}$$

where σ denotes the nonlinear activation function.

2.8. The Output

This part actually consists of a CNN layer with a 1×1 convolutional kernel and a Reshape operation. The purpose of Reshape is to reshape the output to match the input, which is the goal of our prediction.

2.9. The Training and Evaluation Setup

The model was run on an NVIDIA 4090 RTX GPU with 24 G of RAM and an Intel Xeon Platinum 8352 V 2.10 GHz dodeca-core processor as the CPU. The batch size of the training setup is 3, with an initial learning rate of 0.001. The learning rate change mode is set to a cosine type to dynamically adjust the learning rate to speed up the convergence. In this paper, the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Mean Absolute Percentage Error (MAPE), and the Nash–Sutcliffe efficiency coefficient (NSE) are used as the metrics to evaluate the performance of the model. MAE is the absolute error, and MAPE is the relative error. RMSE is more sensitive to outliers compared to MAE, and the NSE is used to reflect the degree of the match between the real and modelled values. These metrics are calculated separately for each forecast map. When calculating these metrics, we only calculated the non-land rasters in the entire image.

$$\text{MAE} = \text{mean}(|\text{Preds} - \text{Trues}|), \tag{10}$$

$$\text{RMSE} = \text{square root}(\text{mean}((\text{Preds} - \text{Trues})^2)), \tag{11}$$

$$\text{MAPE} = \text{mean}(\frac{|\text{Preds} - \text{Trues}|}{\text{Trues}}), \tag{12}$$

$$\text{NSE} = 1 - \frac{\sum (\text{Trues} - \text{Preds})^2}{\sum (\text{Trues} - \text{mean}(\text{Trues}))^2}, \tag{13}$$

The MAE and MAPE are calculated by spatially averaging the data, followed by temporal averaging. Preds are set to denote the predicted values of the grid, Trues denote

the ground values of the grid (NSIDC), and the formulas for the four metrics are shown in Equations (10)–(13). For MAPE, if the denominator is zero, we replace it with 0.1.

3. Results

3.1. Overall Performance

We obtained the optimal parameters for the model after 20 rounds of training on the training set and tested the prediction of the model on the test set, respectively.

We recorded the experimental results for the model in Table 1. It can be seen that the annual average error for each year is very close, with only small fluctuations. The smallest MAE is 1.87% in 2019, and the largest MAE occurs at 1.93% in 2022, with a difference of 0.06%, which indicates the better generalisation ability of the model. The RMSE, which is more sensitive to outliers, is also at a low level, and the MAPE fluctuates slightly around 4.3%. The value of the NSE is around 0.98, indicating a better forecasting accuracy. The five-year average MAE, RMSE, MAPE, and NSE values are 1.89%, 5.98%, 4.31%, and 0.98, respectively. Overall, the absolute and relative errors of SICFormer's prediction are at a low level, and the prediction accuracy is high, which indicates that our model has good prediction performance.

Table 1. Prediction error of SICFormer on test sets of different years.

Metrics	2018	2019	2020	2021	2022	Average
MAE (%)	1.87	1.87	1.89	1.91	1.93	1.89
RMSE (%)	5.96	5.92	5.96	5.94	5.96	5.98
MAPE (%)	4.27	4.34	4.35	4.31	4.29	4.31
NSE	0.98	0.98	0.97	0.98	0.98	0.98

To further explore the predictive performance of the model, Figure 8 shows line plots of the true mean, predicted mean, and error mean for single-day sea ice concentration on a non-land grid for the years 2018–2022, as well as line plots of the model's single-day error for the errors on days 1 and 8 of a forecast period.

In (a) of Figure 8, we have selected all the days in a forecast cycle (8 days) as forecasts for the corresponding date in the year. Apart from the first 8 days, which cannot be forecasted, the rest of the 357 days (358 days in a leap year) have forecasts for the corresponding date. We can see two notable features: First, both the forecast curve and the MAE curve have obvious periodicity, which is because the forecast errors in the first few days tend to be smaller than those in the following days in a forecast cycle, which is in line with common sense; second, there is an apparent process of increasing, decreasing, and then increasing forecast errors around day 150 to day 300, which is because day 150 is when the Northern Hemisphere enters the summer season, when the degree of sea ice change is particularly drastic, leading to an increase in uncertainty that makes the MAE slightly larger. As we move into summer, the sea ice concentration stabilises, at which point the uncertainty decreases and the MAE decreases. Similarly, in winter, the "refreezing" process leads to an increase in the MAE.

In (b) of Figure 8, we have selected day 1 of each prediction cycle as the predicted value for the corresponding date in that year. Since our model predicts 8 days into the future, the first 8 days of each year are unpredictable if the year is treated as a separate test set, and the last 8 days can only be selected as day 1, so there are actually 350 days (351 days in leap years) in the line graph we show. It can be seen that the MAE is very stable, and the predicted values are in almost perfect agreement with the evolutionary trend in the real values, which can accurately reflect the dynamic change process for sea ice within a year, reflecting the characteristics of freezing, melting, and refreezing within the sea ice concentration in an annual cycle and proving that our model has the basic ability to perform the task of short-term prediction of the sea ice concentration.

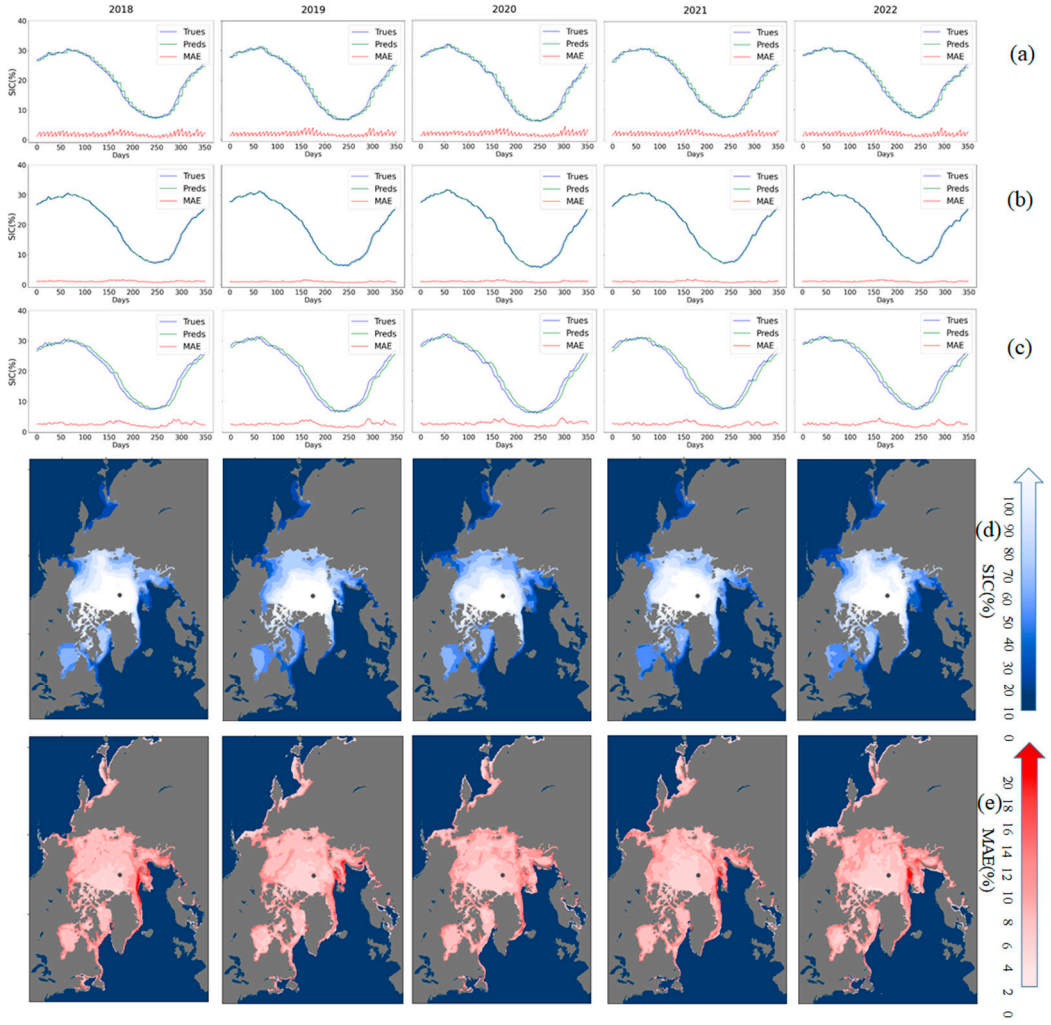

Figure 8. (**a**) Mean error over the forecast period, (**b**) mean error on day 1 of the forecast period, (**c**) mean error on day 8 of the forecast period. (**d**,**e**) are the spatial distribution of the predicted annual mean sea ice concentration and the spatial distribution of the annual mean error.

In (c) of Figure 8, the prediction error on day 8 of each prediction cycle in the test set is shown, which increases and has some error fluctuations, but the overall error is still at a low level. The error line graphs from different angles also prove that long time series prediction based on the deep learning method also has the same two major difficulties as the traditional statistical method, one being the increase in uncertainty brought by time—the longer the prediction, the higher the uncertainty—and the second being the drastic degree of change in the thing itself affecting the prediction error—the more drastic the change is, the higher the prediction uncertainty is as well.

Figure 8 also shows the spatial distribution of the predicted mean annual sea ice concentration (d) and the spatial distribution of the mean annual error (e), where each grid is the average of all the predicted days. The results show that the vast majority of errors are less than 6% and that the larger errors are concentrated in a portion of the sea ice edge region between 0° W and 40° W. Overall, SICFormer performs well in predicting the sea ice concentration for the next 8 days.

3.2. Exploring the Details

In general, the sea ice concentration in the Arctic reaches its minimum around mid-September, when the sea ice and land edge conditions are also more complex. In order to further discuss the predictive effectiveness of the model, we selected 10 September to 17 September in 2021 and 2022 as a sample for analysis.

Figure 9a,b illustrate the spatial distribution of the predicted and true non-land sea ice concentration for a single day between 10 September and 17 September 2021. Figure 9c depicts the spatial distribution of the residuals of the non-land grid for a single day, which is obtained by subtracting the predicted value from the true value. Overall, the spatial distributions of the predicted and true values exhibit a high degree of overlap, indicating the excellent prediction performance of SICFormer.

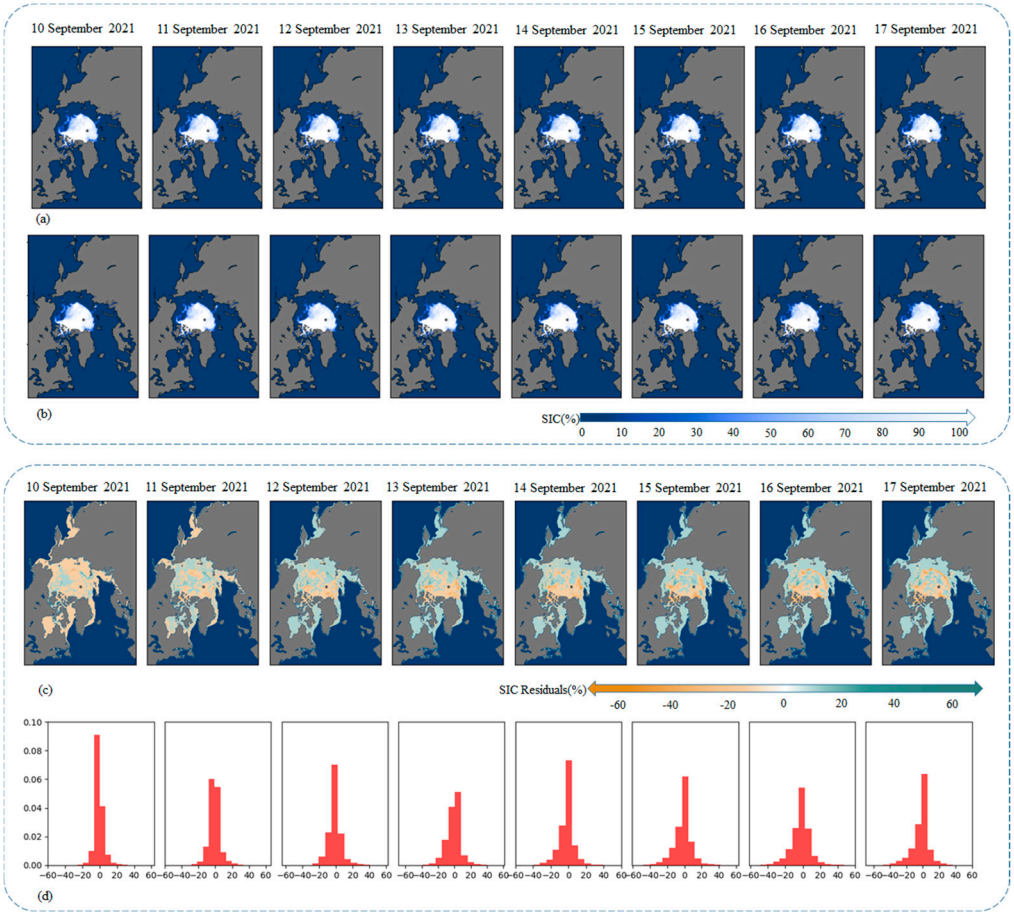

Figure 9. Indicators from 10 September to 17 September 2021. (**a**) Spatial distribution of predicted values. (**b**) Spatial distribution of true values. (**c**) Spatial distribution of residuals as predicted values-true values. (**d**) Histogram of the data distribution of the residuals.

Figure 9c illustrates that our model tends to underestimate the sea ice concentration in areas where the variability is relatively more pronounced. This is a prediction error resulting from the complex ice conditions at the common edge of the sea ice and the coastline. Conversely, our model tends to overestimate the sea ice concentration in areas of coastline and gentler variability, where these errors are relatively small and the colour blocks are lighter, and there is a component of random error in these small errors. This is also well illustrated in Figure 9d, where our error histograms demonstrate that the majority of the errors are within the range of (−20%, 20%), particularly during the initial three days, which exhibit a high concentration of errors. Although a small number of errors exceed 20 percent in the next four days, the majority of the errors remain within the range of (−20%, 20%). For the sake of clarity, the distribution of the errors is more accurately represented in Figure 9 by the exclusion of errors with values within the range (−0.5, 0.5). This range encompasses both open water and some potential random errors.

Figure 10 presents a graph of the same period in 2022. It can be observed that despite the temporal proximity, the discrepancy between the genuine values for the sea ice concentration in 2021 and 2022 is pronounced as a consequence of the elasticity cycle of sea

ice concentration. Nevertheless, the conclusions drawn from 2021 are equally applicable to 2022. As illustrated in Figure 10, the forecast plots for 2022 exhibit a high degree of agreement with the true values, with the majority of the errors falling within the range of −20% to +20%. The comparison between different years demonstrates the model's capacity for generalisation, enabling it to handle differences in the sea ice concentration across years. Consequently, the model exhibits an excellent short-term end-to-end prediction capability, capable of predicting the sea ice concentration for the subsequent eight days within seconds of model training. The bar chart in Figure 10 is generated in the same way as in Figure 9.

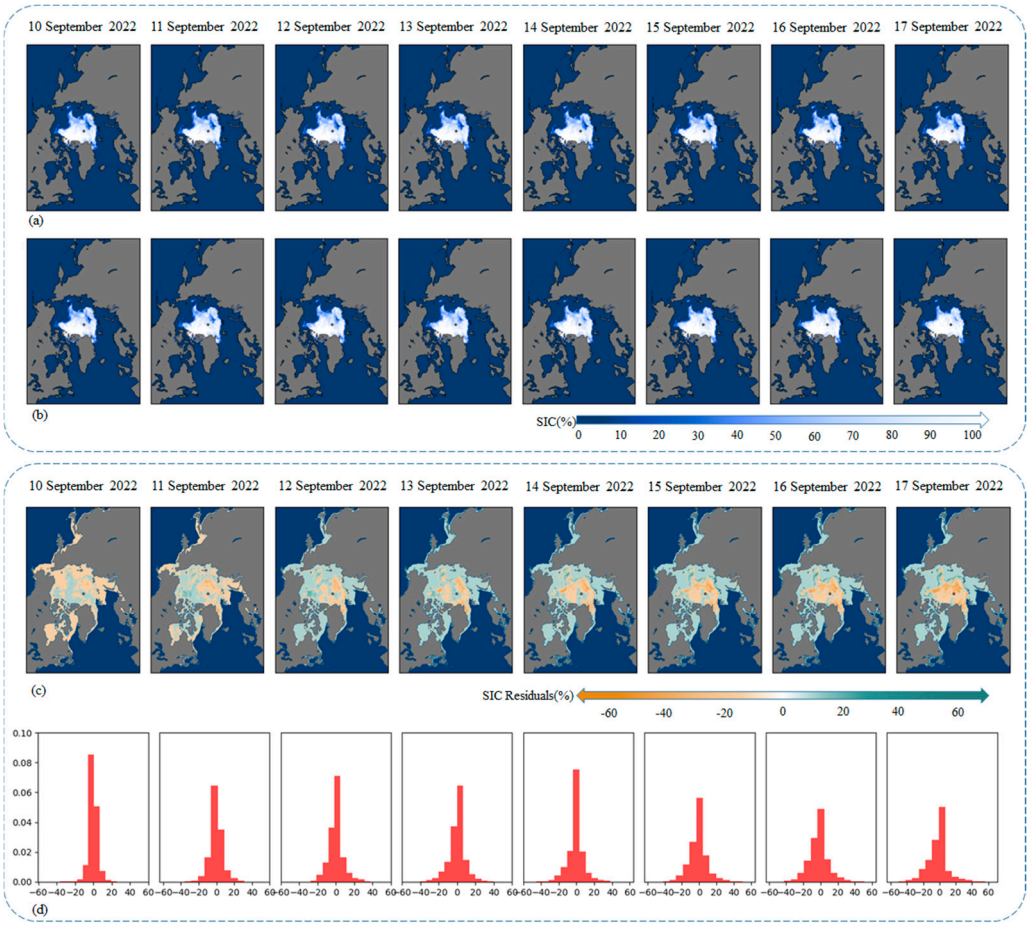

Figure 10. Indicators from 10 September to 17 September 2022. (**a**) Spatial distribution of predicted values. (**b**) Spatial distribution of true values. (**c**) Spatial distribution of residuals as predicted values-true values. (**d**) Histogram of the data distribution of the residuals.

To further demonstrate the model's capacity to capture these dynamics, we selected the periods with the largest errors in Figure 8a for display: days 168–175 and days 296–303 in the year 2021. In other words, the periods from 17 June to 24 June and from 23 October to 30 October in the same year correspond to the melting and freezing periods for sea ice, respectively. As illustrated in Figure 11, the sea ice freezing area is relatively extensive during these two periods, which may be a contributing factor to the significant error observed in this period. The predicted values are in close agreement with the true values in terms of their spatial detail. This demonstrates that the model has the capacity for generalisation.

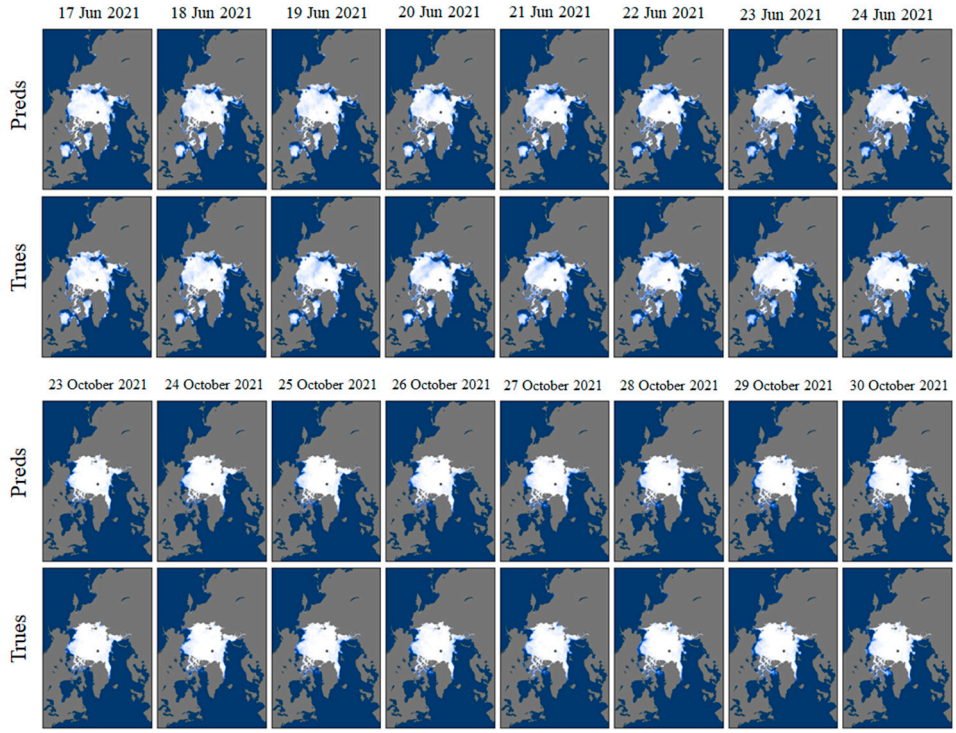

Figure 11. Predicted and real values of sea ice distribution during the ablation and freezing periods.

4. Discussion

We compare this with the existing dominant deep learning-based spatio-temporal prediction models which are capable of end-to-end prediction and have achieved good results on a number of publicly available datasets.

Table 2 presents the predictive performance of the ConvLSTM, PhyDNet, Tau, SimVP, Mau, and PredRNN++ methods, trained under identical conditions, with the results averaged over the years 2018–2022. As illustrated in Table 2, our model outperforms the competition in terms of mean absolute error (MAE) and mean absolute percentage error (MAPE) in the sea ice concentration prediction task. Nevertheless, the SimVP method demonstrated the lowest root mean square error (RMSE) and the highest Nash–Sutcliffe efficiency coefficient (NSE). Additionally, our model attained the fourth highest NSE ranking.

Table 2. Comparative test results for the models.

Methods	MAE	RMSE	MAPE	NSE
ConvLSTM [30]	2.90	6.70	11.52	0.975
TAU [50]	2.67	6.15	9.70	0.979
MAU [51]	2.13	5.65	6.75	0.982
PredRNN++ [52]	1.98	5.67	5.24	0.982
SimVP [27]	2.41	5.48	9.05	0.983
PhyDNet [32]	3.20	7.74	9.02	0.966
Ours	1.90	5.99	4.32	0.980

In order to facilitate a more comprehensive comparison of the models' predictive abilities with regard to dates that are further away, the eighth day of each prediction cycle is taken as the predicted value. Figure 12 illustrates the change in error for each model in 2022, as well as the change in error from 10 September to 17 September. The results demonstrate that our models exhibit the lowest mean absolute error (MAE) for the majority of the year, particularly during the summer months when the sea ice concentration is minimal. This is a period when the sea ice conditions are more complex, with frequently varying sea ice concentrations at the sea ice edge. It is also the time of year when Arctic activity is at its

peak, for which accurate prediction is critical. Figure 12 also shows that the prediction error increases as the prediction date increases within a prediction period, a phenomenon common to all the experimental models.

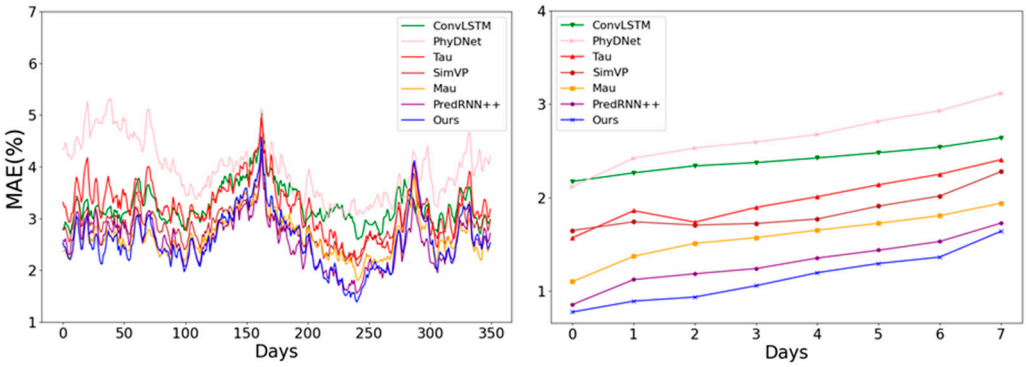

Figure 12. The left panel shows a line graph of the MAE extracted only on day 8 of the 2022 forecast cycle. The right panel shows a line graph for 10 September to 17 September 2022 in one forecast cycle.

Figure 13 illustrates the full-year RMSE fluctuations for 2022, as well as the RMSE variations observed during the 17–24 September and 23–30 October cycles. It can be observed that the model demonstrates a consistently lower level throughout the majority of the year. With the exception of a lower and higher extreme value observed around days 250 and 300, respectively, the changes are relatively consistent. The results displayed in the right graph indicate that the initial three days of a cycle exhibit the lowest values for the RMSE. The RMSE values for the final three days exhibit a slight increase in comparison to those of models such as SimVP yet remain within a comparable range. This illustrates that there is scope for further enhancement of the model in terms of its stability and long-term predictive capability. In conjunction with the MAE line graph, our model exhibits the lowest combined error and demonstrates a competitive forecasting performance.

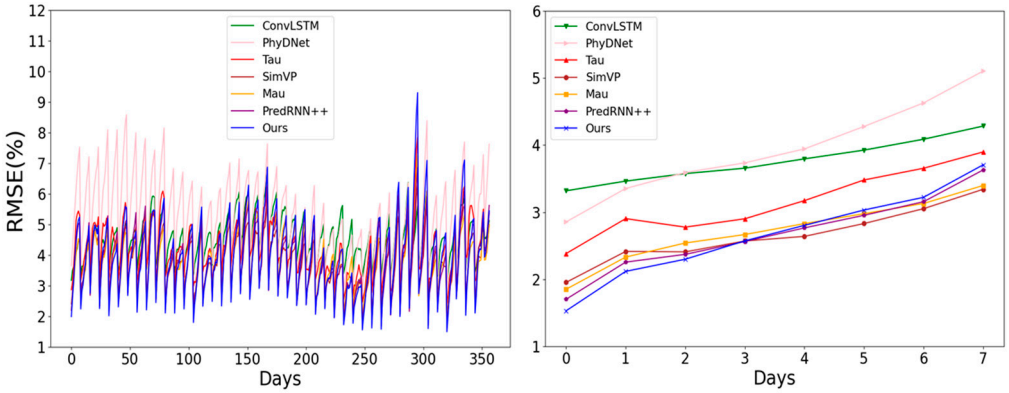

Figure 13. The left is a line graph of RMSE for the full year 2022. The right panel shows a line graph for 10 September to 17 September 2022 in one forecast cycle.

We have also further visualised the prediction results for each model; see Figure 14. It can be seen that the ConvLSTM model, which has a larger error, lacks a multi-level presentation of the sea ice concentration in its prediction images compared to the real images. The edge texture is more blurred and monolithic. This shows that it does not have a strong enough grasp of the details of the sea ice edges, and it lacks the ability to extract detailed features. The other comparative models, such as PhyDNet, have similar problems, but for ConvLSTM, this is the most significant. Our model has a strong edge texture extraction capability and can reconstruct rich concentration features well, and its grasp of detail and more hierarchical representation can fully recover the real sea ice concentration

distribution. However, the excessive attention to spatial detail also presents a limitation of our model in the form of a delayed response to changes in sea ice, resulting in smaller changes in the time series. Nevertheless, this does not preclude our model from retaining its status as the one with the smallest prediction error.

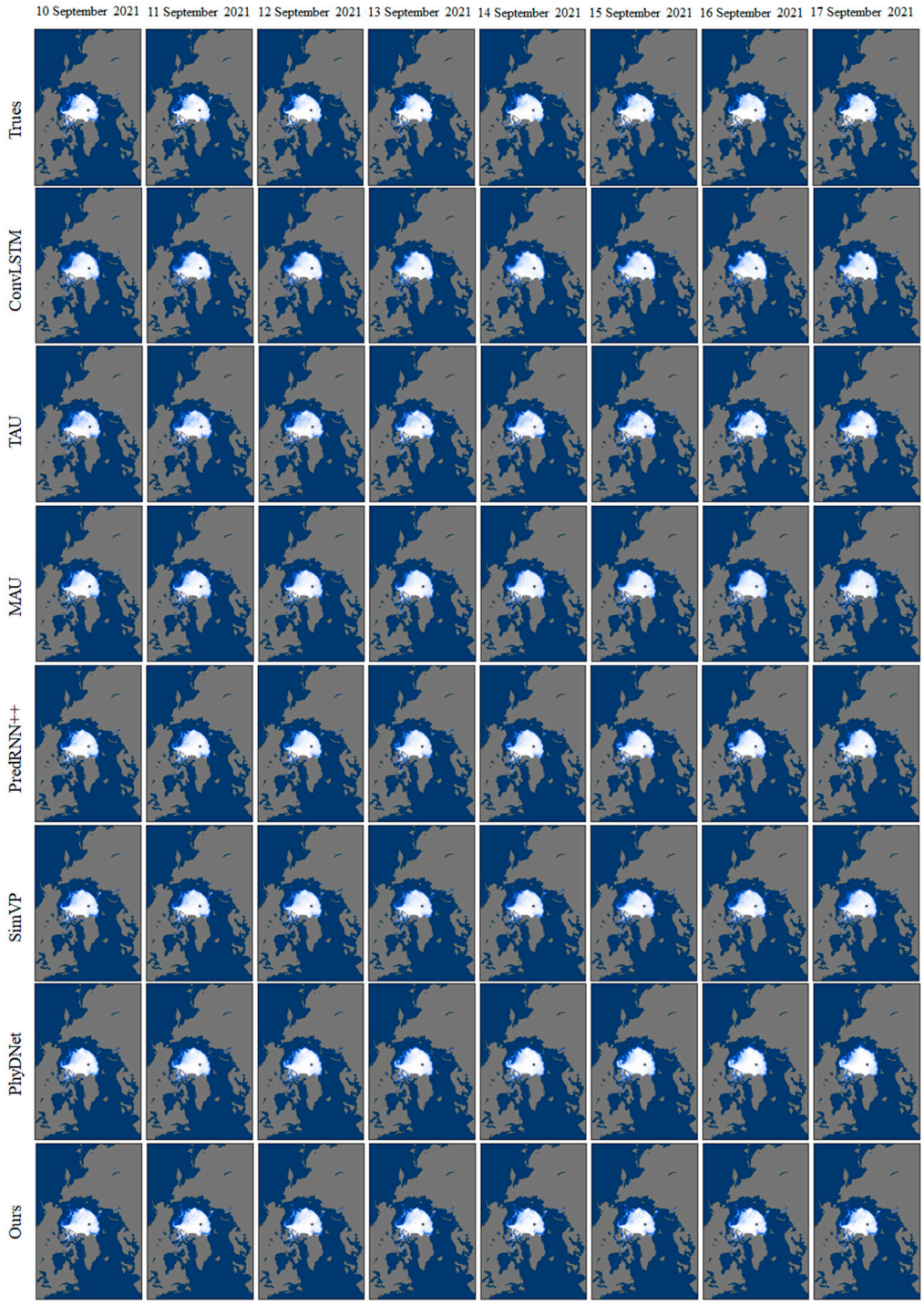

Figure 14. Comparative visualisation of the models' forecasts from 10 September to 17 September 2021.

The results of the comparative experiments show that our model achieves very competitive results in the task of short-term prediction of sea ice concentration. This represents the inaugural effective endeavour to make spatio-temporal predictions utilising the visual class Transformer.

Notwithstanding the advancements achieved, our study continues to present certain challenges. The excessive focus on minutiae has resulted in a limitation of our model. This indicates that the response to sea ice changes is somewhat sluggish, resulting in relatively minor alterations to the temporal sequence. Although the spatio-temporal attention of the 3D-Swin Transformer architecture is capable of learning both temporal and spatial dependencies, this is constrained by the smaller time scale of the data and the larger spatial scale. The model places greater emphasis on the spatial scale, which results in a delayed response in temporal prediction. Further research could attempt to calculate the attention on a time scale and a spatial scale separately and assign different weights to each of them [35]. The selected indicators do not fully reflect the actual situation of sea ice. MAE and RMSE can accurately reflect the specific differences in each prediction grid. However, they are unable to reflect or explain the intrinsic law of sea ice concentration change. In the future, an indicator can be studied and proposed which reflects the rate of the change in sea ice, similar to the rate of change in mass or the rate of change in volume in the field of material science. The magnitude of our prediction errors has been markedly diminished; however, there is a discernible tendency for these errors to increase in regions and periods of significant fluctuations in sea ice concentration. The experimental results demonstrate that as the prediction date approaches, the magnitude of the prediction error diminishes. Iteration through short-term prediction remains a promising avenue of research, and this prediction strategy has been formally adopted by the Pangu-Weather Large Model [53]. In the future, it may be possible to train benchmark models of one-day and two-day lengths, with the possibility of iterating the benchmark model while minimising the iteration error. To illustrate, a prediction of nine days could entail four instances of the two-day model and one instance of the one-day model. This permits greater flexibility in the number of days for which predictions can be made. Furthermore, our future research will also focus on utilising multi-source remote sensing data and investigating the impact of various factors to enhance the interpretability of deep learning models.

5. Conclusions

In this paper, we propose a SICFormer model using a pure Transformer-like architecture as an encoder for a short-term prediction task for sea ice concentration in the pan-Arctic region, with a prediction target of the next 8 days. Specifically, we use a Video Swin Transformer as the encoder, which can take into account both global and local image features, effectively capture spatio-temporal dependencies in the data, and improve the efficiency of training, and design a CNN-based residual block and an upsampling block for sea ice concentration prediction. To avoid noise in the early data, we used a total of 17 years of data from 2006 to 2022, while maintaining the original size of the NSIDC sea ice concentration data product. We also predicted the sea ice concentration for all periods of the year to account for potential errors that may exist in the data product itself. Our experimental results show that SICFormer is capable of predicting the sea ice concentration in the short term and is competitive with the current mainstream spatio-temporal prediction models. The main contributions of this paper's work are as follows:

- A Video Swin Transformer based on the classical Transformer framework is used as an encoder, and a ViTs-CNN-CNN architecture is proposed and applied to a sea ice density prediction task to achieve the best performance.
- A hierarchical nested residual structure is designed. The first layer is a jump connection across the whole model to add raw data to the last CNN block, and the second layer nests a shortcut layer outside the two ResNet blocks. This design ensures stable training.
- Based on the evaluation metrics selected in this paper, the MAE is reduced to 1.89%, the RMSE is 5.99%, and the MAPE is 4.32%. The NSE is 0.980, and the combined performance is the best among all the models compared.

In addition, this paper summarises the limitations of the model and suggests future research directions. Some potential avenues for improvement are suggested in terms

of the accuracy, scoring metrics, and interpretability. It has some reference value for subsequent research.

Author Contributions: Conceptualisation, H.Z. and Z.J.; methodology, H.Z. and Z.J.; software , Z.J. and B.G.; validation, Z.J. and B.G.; formal analysis, Z.J. and H.Z.; investigation, Z.J. and Y.S.; resources, B.G. and Y.J.; data curation, H.Z. and Y.S.; writing—original draft preparation, Z.J. and B.G.; writing—review and editing, Z.J. and Y.S.; visualisation, Z.J. and Y.J.; supervision, B.G. and H.Z.; project administration, H.Z. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 52101405.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at https://nsidc.org/data/ (accessed on 10 December 2023).

Acknowledgments: The SIC remote sensing observations from the National Snow and Ice Data Center are gratefully acknowledged. Special thanks are also extended to Chinese Academy of Sciences for providing excellent resources to conduct this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Parkinson, C.L. A 40-y record reveals gradual Antarctic sea ice increases followed by decreases at rates far exceeding the rates seen in the Arctic. *Proc. Nat. Acad. Sci. USA* **2019**, *116*, 3126280. Available online: https://www.ncbi.nlm.nih.gov/pubmed/31262810 (accessed on 11 October 2023). [CrossRef]
2. Olonscheck, D.; Mauritsen, T.; Notz, D. Arctic sea-ice variability is primarily driven by atmospheric temperature fluctuations. *Nat. Geosci.* **2019**, *12*, 430–434. [CrossRef]
3. Kim, Y.H.; Min, S.K.; Gillett, N.P.; Notz, D.; Malinina, E. Observationally-constrained projections of an ice-free Arctic even under a low emission scenario. *Nat. Commun.* **2023**, *14*, 3139. [CrossRef]
4. Chen, J.; Kang, S.; You, Q.; Zhang, Y.; Du, W. Projected changes in sea ice and the navigability of the Arctic passages under global warming of 2 °C and 3 °C. *Anthropocene* **2022**, *40*, 100349. [CrossRef]
5. Yang, Q.; Losa, S.N.; Losch, M. Assimilating SMOS sea ice thickness into a coupled ice-ocean model using a local SEIK filter. *J. Geophys. Res. Oceans* **2014**, *119*, 6680–6692. [CrossRef]
6. Zhang, J. Sea ice properties in high-resolution sea ice models. *J. Geophys. Res. Ocean.* **2021**, *126*, e2020JC016686. [CrossRef]
7. Adcroft, A.; Anderson, W.; Balaji, V. The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *J. Adv. Model. Earth Syst.* **2019**, *11*, 3167–3211. [CrossRef]
8. Smith, G.C.; Roy, F.; Reszka, M.; Colan, D.S.; He, Z.; Deacu, D.; Belanger, J.-M.; Skachko, S.; Liu, Y.; Dupont, F.; et al. Sea ice forecast verification in the Canadian global ice ocean prediction system. *Quart. J. Roy. Meteorolog. Soc.* **2016**, *142*, 659–671. [CrossRef]
9. Hebert, D.A. Short-term sea ice forecasting: An assessment of ice concentration and ice drift forecasts using the U.S. Navy's Arctic cap nowcast/forecast system. *J. Geophys. Res. Ocean.* **2015**, *120*, 8327–8345. [CrossRef]
10. Liang, X.; Zhao, F.; Li, C.; Zhang, L.; Li, B. Evaluation of ArcIOPS sea ice forecasting products during the ninth CHINARE-Arctic in summer 2018. *Adv. Polar Sci.* **2020**, *31*, 14–25. [CrossRef]
11. Mu, L.; Liang, X.; Yang, Q.; Liu, J.; Zheng, F. Arctic ice ocean prediction system: Evaluating sea-ice forecasts during Xuelong's first trans-arctic passage in summer 2017. *J. Glaciol.* **2019**, *65*, 813–821. [CrossRef]
12. Yang, Q.H.; Mu, L.J.; Wu, X.R. Improving Arctic sea ice seasonal outlook by ensemble prediction using an ice-ocean model. *Atmos. Res.* **2019**, *227*, 14–23. [CrossRef]
13. Wang, L.; Yuan, X.; Ting, M.; Li, C. Predicting summer Arctic sea ice concentration intraseasonal variability using a vector autoregressive Model. *J. Clim.* **2016**, *29*, 1529–1543. [CrossRef]
14. Yuan, X.; Chen, D.; Li, C.; Wang, L.; Wang, W. Arctic sea ice seasonal prediction by a linear Markov model. *J. Clim.* **2016**, *29*, 8151–8173. [CrossRef]
15. Wang, L.; Yuan, X.; Li, C. Subseasonal forecast of Arctic sea ice concentration via statistical approaches. *Clim. Dyn.* **2019**, *52*, 4953–4971. [CrossRef]
16. Ham, Y.G.; Kim, J.H.; Luo, J.J. Deep learning for multi-year ENSO forecasts. *Nature* **2019**, *573*, 568–572. [CrossRef]
17. Ren, Y.; Cheng, T.; Zhang, Y. Deep spatio-temporal residual neural networks for road-network-based data modeling. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1894–1912. [CrossRef]

18. Reichstein, M. Deep learning and process understanding for data driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
19. Zheng, G.; Li, X.; Zhang, R.H.; Liu, B. Purely satellite data–driven deep learning forecast of complicated tropical instability waves. *Sci. Adv.* **2020**, *6*, 1482. [CrossRef]
20. Li, X. Deep-learning-based information mining from ocean remote-sensing imagery. *Nat. Sci. Rev.* **2020**, *7*, 1584–1605. [CrossRef] [PubMed]
21. Zhang, X.; Li, X.; Zheng, Q. A machine-learning model for forecasting internal wave propagation in the Andaman sea. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2021**, *14*, 3095–3106. [CrossRef]
22. Zhang, X.; Li, X. Combination of satellite observations and machine learning method for internal wave forecast in the Sulu and Celebes seas. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 2822–2832. [CrossRef]
23. Liu, B.; Li, X.; Zheng, G. Coastal inundation mapping from bitemporal and dual-polarization SAR imagery based on deep convolutional neural networks. *J. Geophys. Res. Ocean.* **2019**, *124*, 9101–9113. [CrossRef]
24. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
25. Williams, R.J.; Zipser, D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Comput.* **1989**, *1*, 270–280. [CrossRef]
26. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
27. Gao, Z.Y.; Tang, C.; Wu, L.R. SimVP: Simpler yet Better Video Prediction. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3160–3170.
28. Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; Yu, P.S. Memory in memory: A predictive neural network for learning higher-order nonstationarity from spatiotemporal dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9154–9162.
29. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 879–888.
30. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in neural information processing systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
31. Castrejon, L.; Ballas, N.; Courville, A.N. Improved conditional vrnns for video prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7608–7617.
32. Guen, V.L.; Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11474–11484.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; Schmid, C. Vivit: A video vision transformer. *arXiv* **2012**, arXiv:2103.15691.
35. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? *arXiv* **2021**, arXiv:2102.05095.
36. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J. Multiscale vision transformers. *arXiv* **2021**, arXiv:2104.11227.
37. Xu, Z.; Wang, Y.; Long, M.; Wang, J. Predcnn: Predictive learning with cascade convolutions. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 2940–2947.
38. Gao, H.; Xu, H.; Cai, Q.; Wang, R. Disentangling propagation and generation for video prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9006–9015.
39. Chi, J.; Kim, H.C. Prediction of Arctic sea ice concentration using a fully data driven deep neural network. *Remote Sens.* **2017**, *9*, 1305. [CrossRef]
40. Chi, J.; Bae, J.; Kwon, Y.J. Two-stream convolutional long and short-term memory model using perceptual loss for sequence-to sequence Arctic sea ice prediction. *Remote Sens.* **2021**, *13*, 3413. [CrossRef]
41. Kim, Y.J.; Kim, H.C.; Han, D.; Lee, S. Prediction of monthly Arctic sea ice concentrations using satellite and reanalysis data based on convolutional neural networks. *Cryosphere* **2020**, *14*, 1083–1104. [CrossRef]
42. Andersson, T.R.; Hosking, J.S.; Maria, P.O. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nat. Commun.* **2021**, *12*, 5124. [CrossRef] [PubMed]
43. Ren, Y.; Li, X.; Zhang, W. A Data-Driven Deep Learning Model for Weekly Sea Ice Concentration Prediction of the Pan-Arctic During the Melting Season. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4304819. [CrossRef]
44. Zheng, Q.Y.; Wang, R.; Han, G.J. A Spatiotemporal Multiscale Deep Learning Model for Subseasonal Prediction of Arctic Sea Ice. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4300522. [CrossRef]
45. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L. Attention is All you Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017. [CrossRef]
46. Liu, Z.; Lin, Y.T.; Cao, Y. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.

47. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z. Video Swin Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3192–3201.
48. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Shi, W.Z.; Caballero, J.; Huszár, F. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
50. Tan, C.; Gao, Z.; Li, S.; Xu, Y.; Li, S. Temporal Attention Unit: Towards Efficient Spatiotemporal Predictive Learning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18770–18782.
51. Chang, Z. MAU: A Motion-Aware Unit for Video Prediction and Beyond. *Neural Inf. Process. Syst.* **2021**, *34*, 26950–26962.
52. Wang, Y.; Gao, Z.; Long, M.; Wang, J.; Yu, P.S. PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
53. Bi, K.F.; Xie, L.X.; Zhang, H.H.; Chen, X.; Gu, X.T.; Tian, Q. Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. *arXiv* **2022**, arXiv:2211.02556.