Comprehensive Transformer-Based Model Architecture for Real-World Storm Prediction

Fudong Lin[1], Xu Yuan[1(✉)], Yihe Zhang[1], Purushottam Sigdel[2], Li Chen[1], Lu Peng[3], and Nian-Feng Tzeng[1]

[1] University of Louisiana at Lafayette, Lafayette, LA 70504, USA
xu.yuan@louisiana.edu
[2] Intel Corporation, Santa Clara, CA 95054, USA
[3] Tulane University, New Orleans, LA 70118, USA

Abstract. Storm prediction provides the early alert for preparation, avoiding potential damage to property and human safety. However, a traditional storm prediction model usually incurs excessive computational overhead due to employing atmosphere physical equations and complicated data assimilation. In this work, we strive to develop a lightweight and portable Transformer-based model architecture, which takes satellite and radar images as its input, for real-world storm prediction. However, deep learning-based storm prediction models commonly have to address various challenges, including limited observational samples, intangible patterns, multi-scale resolutions of sensor images, *etc.* To tackle aforementioned challenges for efficacious learning, we separate our model architecture into two stages, *i.e.*, "representation learning" and "prediction", respectively for extracting the high-quality feature representation and for predicting weather events. Specifically, the representation learning stage employs (1) multiple masked autoencoders (MAE)-based encoders with different scalability degrees for extracting multi-scale image patterns and (2) the Word2vec tool to enact their temporal representation. In the prediction stage, a vision transformer (ViT)-based encoder receives the input sequence derived from packing the image patterns and their temporal representation together for storm prediction. Extensive experiments have been carried out, with their results exhibiting that our comprehensive transformer-based model can achieve the overall accuracy of 94.4% for predicting the occurrence of storm events, substantially outperforming its compared baselines.

Keywords: Storm Predictions · Vision Transformers · AI for Science

1 Introduction

Storms can cause areal catastrophes resulting from property damage, injuries, and even deaths. It has long been a critical and essential task for prompt and accurate storm occurrence prediction to facilitate emergency alert broadcasting in advance for early preparation actions. However, conventional physical

G. De Francisci Morales et al. (Eds.): ECML PKDD 2023, LNAI 14175, pp. 54–71, 2023.
https://doi.org/10.1007/978-3-031-43430-3_4

models for storm predictions tend to suffer from excessive computational overhead caused by vast climate data simulation and complicated data assimilation from different sources. Meanwhile, deep learning (DL) has enjoyed impressive advances in various applications [3,7,9,12,15,17,18,22,27,31,34,38,46,49], including those [1,23,33,40–42,59] for weather forecasting.

A few attempts have been made to develop DL-based models for storm prediction [8,20,58,60] with unsatisfactory outcomes. Their main obstacles are multifold, including limited observational storm samples in real-world scenarios, complicated and intangible patterns existing in typical storm data, which are usually multi-modal and multi-scalar, among others. Known prediction models often failed to address one or multiple such obstacles with inflexible and coupled structures, thus hindering their generalization to the real scenarios. To date, it remains open and challenging to harness DL-based models by effectively dealing with those obstacles for accurately predicting the occurrence of storm events.

To tackle the aforementioned obstacles, we endeavor to develop a comprehensive model architecture able to flexibly admit the satellite and radar images for real-world storm prediction, resorting to the vision transformer (ViT) [12] and masked autoencoders (MAE) [15]. In particular, we separate our model architecture into two stages, *i.e.*, "representation learning" and "prediction", respectively for learning high-quality representations of data and predicting weather events of interest. In the representation learning stage, three MAE-based encoders with different scalability degrees corresponding to multi-scale sensor images are utilized for extracting affluent image patterns. Meanwhile, the Word2vec [36] tool is employed to learn the temporal representations of weather events, with such representations viewed as the critical features of storm events. A pooling layer and a linear projection layer are designed to bridge the two stages for matching the length of the input sequence and the hidden vector size, respectively, able to significantly reduce the memory utilization and computation cost of self-attention as well. In the prediction stage, a ViT-based encoder is employed to receive latent representations constructed by packing image and temporal representations together. Similar to the original ViT, a multi-layer perceptron (MLP) is used to serve for predicting the occurrence of storm events, based on the learnable classification token. Inspired by the segment embedding in BERT [9], we also propose a novel content embedding for MAE-based and ViT-based encoders, able to differentiate the memberships of various sources of representations.

We have conducted extensive experiments on the real SEVIR dataset [47], which includes a collection of real-world satellite and radar images with different resolutions, as well as detailed weather event descriptions (*e.g.*, times and locations). The experimental results demonstrate that our Transformer-based architecture achieves 94.4% overall accuracy in predicting the occurrence of storm events. In addition, we conduct comprehensive ablation studies, whose results exhibit the significance and necessity of our novel designs on the ViT and MAE encoders for real-world storm predictions. These empirical results demonstrate the practical impact of our solution for precisely predicting storm events to avoid potential catastrophic loss and damage.

2 Related Work

This section presents prior work on vision transformers and deep learning-based weather forecasting.

Vision Transformers. Popularized by ViT [12], vision transformers have been a powerful surrogate to conventional neural networks (CNNs) for vision tasks. It splits an image into a set of patches and relies on its encoder to receive the input constructed by summing up a linear projection of patches and positional embeddings. Then, an extra learnable classification token is used for performing classification tasks. Subsequent work built upon the ViT abounds. For example, DEiT [45] addresses the original ViT's overfitting issue by appending a novel distillation token. Swin [35] tackles high-resolution inputs by adopting the hierarchical structure from CNNs. TiT-ViT [55] aggregates structure information by recursively merging neighboring tokens into one token, and MAE [15] introduces self-supervised learning to the vision domain built upon ViT backbones. Some studies, including MViT [13,30], PiT [50], PVT [19], among many others [5,11,28,29,53,54,56,57], also address the limitations of original ViT for better performance on vision tasks. Despite effectiveness in theoretical deep learning, prior studies all focus on image data with very similar resolutions only, thereby difficult to make it adapt the rich real-world satellite and radar images with varying resolutions. Although our comprehensive model architecture builds on ViT and MAE, several novel designs (*e.g.*, content embedding) are tailored to address such real-world challenges as limited observational samples, intangible patterns, multi-modality data, and multi-scale input images.

Deep Learning for Weather Predictions. Deep learning (DL) has been popularly adopted for addressing critical and challenging meteorological issues in recent years. [42] has proposed a convolutional LSTM (ConvLSTM) network, based on the fully connected LSTM (FC-LSTM) to construct an encoding-forecasting structure by concatenating several ConvLSTM layers, arriving at an end-to-end trainable model for short-term weather predictions. Motivated by ConvLSTM, subsequent studies employ various deep neural network structures, such as Autoencoders [21,32], DLWP models [1,44], LSTM [4,43,48,51,52], and others [23,40,41], for weather predictions. A few studies [8,20,58,60] also started to tackle storm predictions from the DL perspective. However, their proposed architectures are often coupled and inflexible, thereby difficult to be generalized to the real scenario. In sharp contrast, we separate our comprehensive model architecture into two loosely coupled stages, permitting multiple MAE encoders to be flexibly incorporated into or detached from our proposed architecture. Such a design approach can be flexible to tackle multi-resolution image data from different sources, capturing rich intangible patterns for accurate storm prediction.

3 Problem Statement, Challenge, and Idea

3.1 Problem Statement

In this work, we aim to develop a deep learning (DL)-based model, for effectively capturing the complex weather data patterns, to predict the occurrence

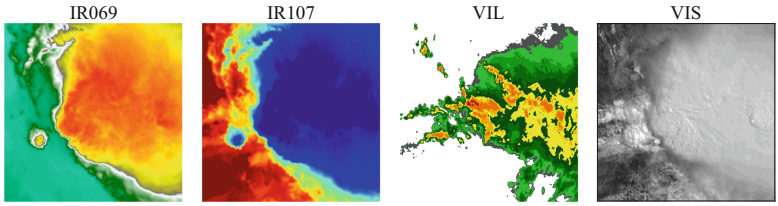


Fig. 1. Illustration of four types of sensor data for storm predictions.

Table 1. Description of the SEVIR dataset

Type	Satellite/Radar	Resolution	Description
IR069	GOES-16 C09 6.9 μm	192×192	Infrared Satellite imagery (Water Vapor)
IR107	GOES-16 C13 10.7 μm	192×192	Infrared Satellite imagery (Window)
VIL	Vertically Integrated Liquid (VIL)	384×384	NEXRAD radar mosaic of VIL
VIS	GOES-16 C02 0.64 μm	768×768	Visible satellite imagery

of storm events. Despite massive storm data available publicly (*e.g.*, the NOAA Storm Events database [39]), their tremendous sizes and extraordinary complexity usually hinder the training process of DL models. To guide our model design for storm event predictions, we employ a storm dataset downsampled from NOAA called SEVIR [47], which contains a collection of sensor images captured by satellite and radar, characterizing weather events during 2017–2019. Those sensor images can be grouped into four categories, *i.e.*, IR069, IR107, VIL, and VIS, captured respectively by GOES-16 C09 6.9 μm, GOES-16 C13 10.7 μm, Vertically Integrated Liquid, and GOES-16 C02 0.64 μm. Figure 1 depicts a set of sensor images for a weather event and Table 1 presents the details of the SEVIR dataset. This dataset also contains abundant numerical and statistical description for weather events, *e.g.*, time widow, location, *etc.* In particular, it contains $10,180$ normal and $2,559$ storm events.

The primary aim of our model is to predict whether storms will occur, deemed as a binary storm prediction, *i.e.*, either storm or normal events. Following prior studies [10,20], we frame the storm prediction as the binary classification problem.

3.2 Challenges

Limited Observational Samples. In real-world scenarios, storms belong to rare events, having fewer observational data samples than normal, non-storm events. This poses grant challenges to DL models for learning sufficient patterns, whereas the normal events' patterns dominate the data. For example, in the preprocessed SEVIR dataset, the overall storm events only include $2,559$ samples, accounting for just 20% of total events. How to develop an effective model to learn from the limited observational samples for achieving satisfactory performance remains open and challenging.

Intangible Patterns. Since the weather images typically come from the satellite and radar, they usually include erratic and intangible shapes compared to other real-world objects. This lifts the difficulty in designing the model for accurate prediction, requiring to deeply capture the hidden and common storm patterns.

Multi-scale and Multi-modal Data. The conventional DL models are only designed for taking one small-scale input. But, a storm event typically has images from different sources with multi-scale resolutions. For example, there are four types of sensor images (*i.e.*, IR069, IR107, VIL, and VIS) in the SEVIR dataset with three different resolutions, *i.e.*, 192×192, 384×384, and 768×768, as listed in Table 1. We aim to take all types of images into account to increase the data sample amounts for use. So far, how to effectively align the features from multiple types of sensor images with multi-scale resolutions remains open. Beyond sensor images, the language data (*e.g.*, time description) is also closely correlated to storm occurrences. Our model is expected to feed both image and language data concurrently, deemed as multi-modal data, whose effective processing by the DL approach is still a big challenge.

3.3 Our Idea

To tackle the aforementioned challenges, we develop a comprehensive transformer-based model architecture for storm predictions, where the predictions are made under the simultaneous consideration of all types of sensor images as well as language-based prior knowledge (*i.e.*, time description).

Our design is driven by the following three observations. *First*, as shown in Fig. 1, different types of sensor images for a weather event contain very similar high-level patterns (*e.g.*, the shape). Based on this observation, for each weather event, we can construct its comprehensive image representation by concatenating the feature embedding extracted from all types of sensor images, arriving at a higher-quality representation. *Second,* thank to vision transformer (ViT) [12], the gap between natural language processing (NLP) and computer vision (CV) has been significantly mitigated [15]. Besides, vision transformer can benefit from task-specific domain knowledge [29]. Hence, it is feasible to explore some mechanisms to incorporate language-based prior knowledge into the vision transformer, thus in turn augmenting its efficacious DL from limited observational samples. *Third,* to handle the multi-scale resolutions problem, we can tailor multi-scale transformer encoders for embedding different types of sensor images, as shown in Fig. 2 (Bottom). Meanwhile, a novel content embedding (see Figs. 3a, 3b, and 3c) can be included for differentiating the membership of various sources of representations, motivated by the segment embedding in BERT [9].

4 Method

Figure 2 illustrates the overview of our proposed model architecture for storm predictions, consisting of two stages, *i.e.*, *representation learning* and *prediction*. The representation learning stage (*i.e.*, Fig. 2 Bottom) involves three different

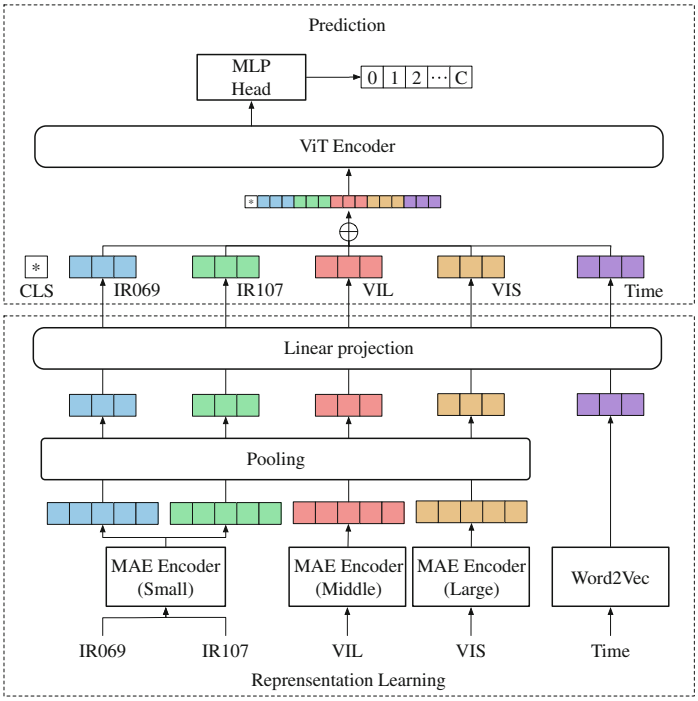


Fig. 2. Overview of our transformer-based comprehensive model architecture.

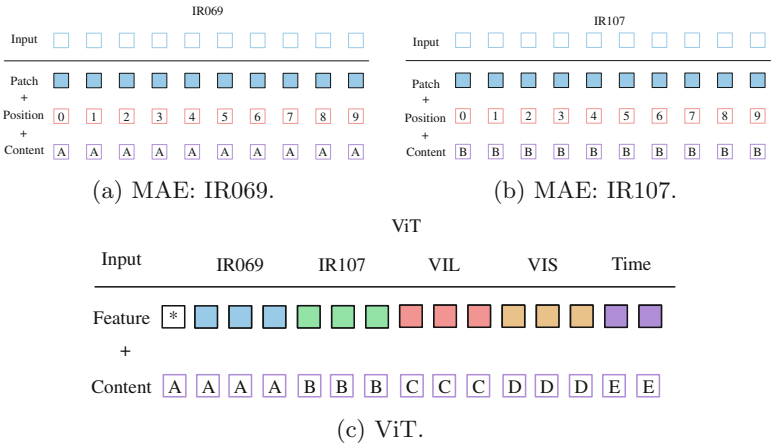
scales of transformer encoders trained by Masked Autoencoders (MAE)-denoted as MAE encoders-for extracting image representation from IR069 (or IR107), VIL, and VIS, respectively. The Word2vec tool is used to extract the temporal representation of weather events. The pooling layer and the linear projection layer serve to bridge the two stages by matching the input sequence length and the hidden vector dimension, respectively, making it possible for our model architecture to decouple the two stages to some extent.

The prediction stage (*i.e.*, Fig. 2 Top) derives the input sequence from a weather event by concatenating its comprehensive image representation extracted from four types of sensor images and its temporal representation extracted from the time description of that event. A learnable classification token is fed to the multi-layer perceptron (MLP) for storm predictions.

4.1 Representation Learning

This stage attains both the image representation and the temporal representation, respectively for the sensor data and the descriptive time data.

Image Representation. MAE encoders are applied for extracting the image representation here. To tackle input images with different resolutions, prior studies [15,25,26,29] often randomly scale up or crop input images to a fixed resolution (*e.g.*, 224×224). This simple solution is effective in conventional vision tasks



(a) MAE: IR069. (b) MAE: IR107.

(c) ViT.

Fig. 3. Illustration of how we construct the input sequence.

as they typically consider identical small-scale images only. However, given our sensor images have multi-scale resolutions (*e.g.*, 192×192 vs 768×768), scaling up (or cropping) to the same scale may add redundant (or discard important) information. To tackle this issue, we respectively design the small-, middle-, and large-scale MAE encoders for extracting image representations from IR069 (or IR107), VIL, and VIS, based on their resolutions listed in Table 1 (the 3rd column). Notably, each MAE encoder in our design can be fed with several types of sensor images with similar resolutions. Meanwhile, inspired by the segment embedding in BERT [9], we devise a learnable content embedding to differentiate the membership of various sources of images fed into the same MAE encoder for high-quality feature embeddings. Similar to MAE, we divide an image into a set of image patches and construct the patch embeddings by a linear projection of image patches. The position embedding is used to indicate positional information of image patches. The input sequence for MAE is constructed by adding the patch embedding, position embedding, and our novel content embedding. Figures 3a and 3b show examples of the input sequence for IR069 and IR107, respectively. Notably, content embeddings in the middle- or large-scale MAE encoders are removed as only one type of sensor image is fed into them.

Temporal Representation. This part is inspired by the prior study [29], which has reported that vision transformer can benefit from task-specific prior knowledge. So, we incorporate the temporal representation into the input representation sequence of the ViT encoder. The temporal representation for a storm event is extracted from its beginning date in a month interval manner. Specifically, for a given weather event, we use Word2vec [36] to embed three dates relevant to its beginning date (*i.e.*, two weeks before its beginning date, its beginning date, and two weeks after its beginning date). The temporal representation is constructed by packing together the three date embeddings to form a monthly interval. The intuition underlying this month interval manner is that using the month to cap-

ture a storm's occurrence is more informative than its specific day. That is, a storm event is more likely to happen within a specific month interval rather than a specific date.

If constructing the input representation sequence for the prediction stage naively by concatenating the image and temporal representations directly, the self-attention in our ViT encoder will incur considerable memory and computation burden. Instead, we employ a pooling layer to shrink the length of the image representation by consolidating the latent representation outputted via MAE encoders. Notably, regarding the temporal representation, which is already of small length (*i.e.*, 3), it is unnecessary to apply the pooling to it. After that, a linear projection layer is utilized to match the hidden vector sizes between the two stages. This way decouples our representation learning and prediction stages to some extent. Note that the hidden vector sizes in our two stages are different.

4.2 Prediction

In this stage, the feature embedding for a weather event is constructed by concatenating a learnable "classification token" (*i.e.*, CLS token) and its image and temporal representations, as shown in Fig. 1. Similar to the small-scale MAE encoder, the content embeddings are used for differentiating the membership of various sources of the feature embedding (*i.e.*, IR069, IR107, VIL, VIS, or event time). We remove the positional embedding here as no valuable position information exists among various sources of representation. Hence, the input for our ViT encoder is derived by summing up the feature embedding and the content embedding, as illustrated in Fig. 3c. A multi-layer perceptron (MLP) receives the CLS token output by the ViT decoder (*i.e.*, the head of the output sequence) to predict whether a storm (or a specific storm type) will occur.

Our technical contributions are summarized as follows. First, to tackle the issue of scarce observational samples for storm predictions, we enrich the latent representations of weather events by concatenating image representations extracted from different types of sensor images. As such, our ViT encoder can benefit from higher-quality representations. Second, we leverage language-based prior knowledge for storm predictions by appending temporal representations extracted from the descriptive time data of weather events. To the best of our knowledge, this is the very first work on DL-based storm predictions that addresses multi-modality data. Third, we devise a novel content embedding for both MAE and ViT encoders, benefiting Transformer-based models by indicating the membership of various input types. This can greatly improve the performance of Transformers when handling multiple types of inputs simultaneously. Fourth, although we use the SEVIR dataset to demonstrate the feasibility of our model architecture, it in effect can be generalized to deal with any type of satellite/radar image with various resolutions from NOAA for real-world storm predictions.

Table 2. Details of MAE and ViT encoders used in our design

Model	Scale	Layers	Hidden size	Heads	Patch size	Input Size	MLP Ratio
MAE encoder	Small	12	192	6	16×16	224×224	4
	Middle	12	384	12	32×32	448×448	4
	Large	12	768	16	48×48	672×672	4
ViT encoder	ViT-Base	12	768	12	–	–	4

5 Experiments and Results

We implement our proposed model architecture and conduct extensive experiments to evaluate its performance in storm prediction. We follow the 80/20 training/test to split on the SEVIR dataset, whose event counts for normal and storm events are 10,180 and 2,559, respectively.

5.1 Experimental Setting

Baselines. We take the convolutional neural networks (CNN) and vision transformer (ViT) as our baselines for comparison. Specifically, the models of ResNet-50 [16] and ViT-Base [12] are used for the two baselines. The hyperparameters are set as reported in their original studies. Since the baselines cannot take input with multiple resolutions, we consider two cases for comparison: 1) each type of sensor image is regarded as a single dataset, and 2) scaling up or cropping four types of sensor images to the same size (*i.e.*, 224×224).

Parameter Settings. We build our MAE and ViT encoders on the top of MAE's official code[1]. But due to the computational limitations, we prune their models to the relevant small models for use, with detailed parameters listed in Table 2. They in effect can be easily scaled up to large model sizes for real-world storm predictions.

We employ the AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, the weight decay of 0.05, and a batch size of 128 (or 64) for MAE encoder (or ViT encoder). Following [2, 14,15,24,37], we employ the layer-wise learning rate decay [6] of 0.75. We train MAE encoder (or ViT encoder) for 50 (or 100) epochs, with the base learning rate of $1e-3$, the linear warmup epochs of 5, and the cosine decay schedule. We train small-, middle-, and large-scale MAEs, including encoders and decoders, respectively on IR069 (and IR107), VIL, and VIS, with a masking ratio of 75%. After training, we only use the encoders to extract image representation without any masking.

For the *Word2vec*, we use its implementation by Gensim[2], with sg = 0 (*i.e.*, CBOW), vector size = 768, min count = 1, window = 3, and the training epoch of 1,000. Note that the training data for Word2vec is pre-processed by the days through the years 2017–2019.

[1] https://github.com/facebookresearch/mae.
[2] https://github.com/RaRe-Technologies/gensim.

Table 3. Storm predictions on the SEVIR dataset. *All* in the two baselines denotes the scenario that scales up or crops multi-scale images to a fixed resolution. The best results are shown in bold

Method	Image Types	Normal			Storm			Accuracy
		Precision	Recall	F1 Score	Precision	Recall	F1 Score	
Resnet-50	IR069	90.1	87.2	88.6	53.4	60.4	56.7	82.0
	IR107	88.6	94.8	91.6	69.9	49.5	58.0	86.0
	VIL	90.8	85.2	87.9	51.3	64.2	57.0	81.1
	VIS	87.7	96.2	91.8	74.1	44.3	55.4	86.1
	All	86.1	93.5	89.6	58.3	37.6	45.7	82.5
ViT-Base	IR069	90.5	91.7	91.1	63.7	60.4	62.0	85.5
	IR107	88.0	92.5	90.2	60.9	47.9	53.6	83.8
	VIL	91.9	93.9	92.9	72.2	66.0	69.0	88.4
	VIS	87.7	92.0	89.8	58.7	46.9	52.1	83.2
	All	87.9	97.1	92.3	78.9	45.1	57.4	86.9
Ours	–	**95.5**	**97.7**	**96.6**	**89.4**	**81.1**	**85.0**	**94.4**

5.2 Overall Performance Under Storm Event Predictions

We take precision, recall, and F1 score as our evaluation metrics to exhibit the performance in predicting the occurrence of storm events (*i.e.*, binary storm prediction). Table 3 presents the values of three metrics of our model architectures as well as the comparative results to the CNN (*i.e.*, ResNet-50) and vision transformer (*i.e.*, ViT-Base) baselines. For two baselines, we consider both scenarios: 1) taking each type of sensor image as the individual input and 2) scaling up or cropping all sensor images to the 224×224 for inputting as a whole dataset, denoted as *All*. We observe that our approach can always beat all baselines in predicting both normal and storm events, achieving the precision, recall, and F1 score values of $95.5\%, 97.7\%, 96.6\%$ and of $89.4\%, 81.1\%, 85.0\%$, respectively. Our overall accuracy is 94.4%, exceeding 8.3% and 6.0%, respectively, to the best results of ResNet-50 (86.1% on VIS) and of ViT-Base (88.4% on VIL). In terms of predicting normal events, both our approach and the baselines can achieve promising prediction results, with most values of three metrics more than 90%. The reason is that the normal events belong to the majority in the dataset (10180 of 12667), so their patterns are easy to be learned by both our approach and the baselines. But when predicting the storm events, the performance of two baselines degrades largely, with most values below 70%. In particular, among three metrics, the recall corresponding to the column under Storm in Table 3 is the most difficult but very important. We observe that our approach can still achieve the value of 81.1%. Regarding the ResNet-50 and ViT-Base, their best values are only 64.2% and 66.0%, respectively, largely underperforming our approach, demonstrating their limited learning ability from the storm samples. This also signifies the importance of our approach, which has novel designs of representation learning, temporal representation, and content embedding, which

Table 4. Ablation studies for different components on our comprehensive model

Method	Normal			Storm			Accuracy
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	
MAE encoder$^-$	93.0	92.7	92.8	70.3	71.0	70.7	88.5
Temporal$^-$	94.9	92.2	93.5	71.2	79.5	75.1	89.7
Content$^-$	94.0	96.5	95.2	83.7	74.4	78.8	92.2
Ours	**95.5**	**97.7**	**96.6**	**89.4**	**81.1**	**85.0**	**94.4**

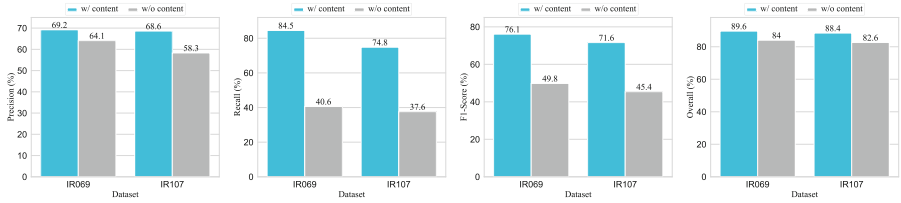
can better learn storm patterns with limited observational samples. Besides, we observe that naively scaling up or cropping multi-scale images to a fixed resolution results in the lowest recall on both baselines, *i.e.*, 37.6% on ResNet-50 (the 7th row) and 45.1% on ViT-based (the 12th row). The reason is that this naive solution may add redundant or discard valuable information.

5.3 Significance of Our Design Components for Storm Predictions

We next conduct ablation studies to show the necessity and significance of each design component in our comprehensive model, in contributing to the storm prediction. In particular, the representation learning, temporal representation, and content embedding are removed in turn, as our design variants to evaluate the performance of the remaining system. The three corresponding variants are denoted as MAE encoder$^-$, Temporal$^-$, and Content$^-$, respectively. Notably, in MAE encoder$^-$, image representations are constructed by a linear projection of image patches, similar to the original ViT. Table 4 presents our experimental results. We have three observations. First, all three variants perform worse than our original model architecture, especially for predicting storm events. Second, MAE encoder$^-$ performs worst among three variants, with its respective precision, recall, and F1 score values of 19.1%, 10.1%, and 14.3% less than ours, in terms of predicting storm events. This demonstrates the necessity of using MAE encoders to learn high-quality image representation in our design. Third, although Content$^-$ and Temporal$^-$ can achieve better performance than MAE encoder$^-$, they still perform much inferior to ours for predicting storm events. Specifically, for Content$^-$, its recall and F1 score values are respectively 6.7% and 6.2% worse than ours. For Temporal$^-$, its precision and F1 score values are respectively 18.2% and 9.9% worse than ours. This validates that both temporal representation and content embedding are also important to our design. Hence, we can conclude that all design components are necessary and important in contributing to our model architecture's prediction performance.

5.4 Necessity of Content Embedding in Our MAE Encoder

Here, we show the importance and necessity of novel content embedding in our MAE encoder (See Figs. 3a and 3b) when addressing multiple types of sensor



(a) Precision (b) Recall (c) F1-Score (d) Overall accuracy

Fig. 4. Comparative results of our MAE encoder with/without the proposed content embedding.

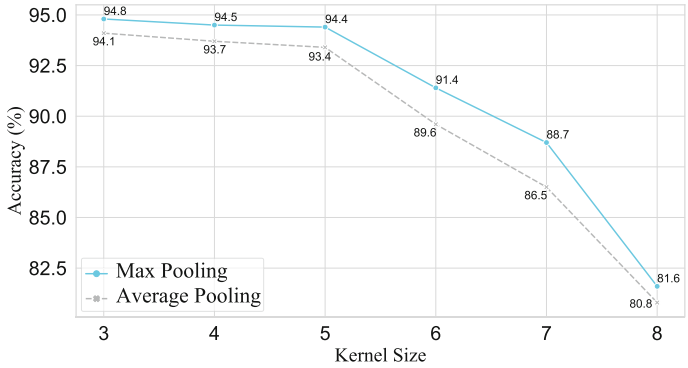
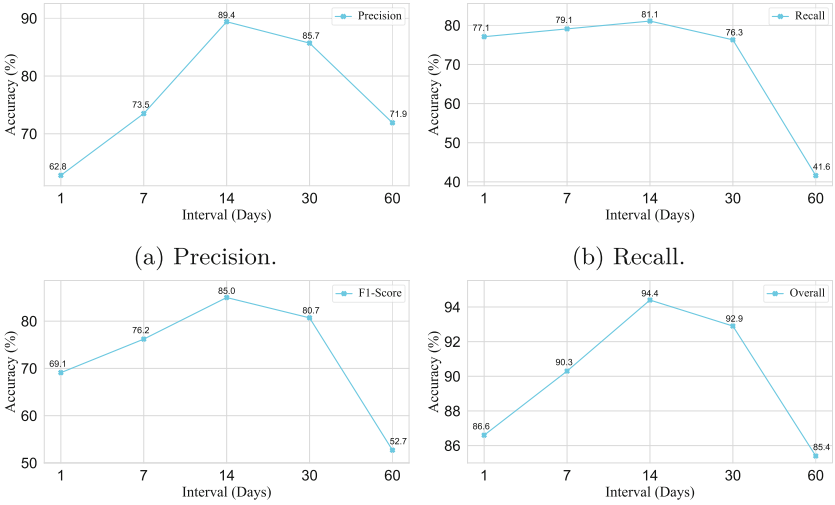


Fig. 5. Results under two pooling methods with various kernel sizes.

images with similar resolutions. The experiments were conducted on our small-scale MAE encoder only, as content embedding is not applied on our middle- and large-scale MAE encoders. Similar to our experiments on ViT-Base, we regard IR069 and IR107 as two datasets. The experimental settings are the same as "end-to-end fine-tuning" in the original MAE [15]. Notably, the experimental results in this section are obtained by removing content embedding in our MAE encoder, different from those in Sect. 5.3 whose results are obtained by removing content embedding on our ViT encoder.

Figure 4 depicts comparative results of our MAE encoder with/without the content embedding. In particular, we focus on the precision, the recall, and the F1-score of the storm event, as well as the overall accuracy, as shown in Figs. 4a, 4b, 4c and 4d, respectively. We have two observations. First, our MAE encoder with the content embedding achieves better performance under all scenarios, with overall accuracy improvements of 5.6% and of 5.8% on IR069 and IR107, respectively. Second, regarding the recall, our MAE encoder with the content embedding achieves the best improvement, with respective 43.9% and 37.2% improvements on IR069 and IR107. This statistical evidence exhibits the necessity and importance of the proposed content embedding in our MAE encoder for



(a) Precision. (b) Recall.

(c) F1-Score. (d) Overall accuracy.

Fig. 6. Experimental results of our temporal representation with different time intervals, including (a) the Precision, (b) the Recall, and (c) the F1-score of the storm event, and (d) the overall accuracy.

differentiating the membership of various input sources, which can substantially elevate our model performance.

5.5 Detailed Design Underlying Our Pooling Layer

We present the detailed design underlying our pooling layer to support our architecture mentioned in Sect. 4.1. Two common pooling methods (*i.e.*, the max and the average pooling) with different sizes of the sliding window (*i.e.*, the kernel size) are taken into account. The kernel size varies from 3 to 8[3]. The stride of the sliding window is set to the same as the kernel size. As such, a larger kernel size can reduce the memory and computation cost of self-attention. Figure 5 presents experimental results in terms of overall accuracy. We observe that the max pooling outperforms the average pooling under all scenarios. Besides, when the kernel size is greater than 5, increasing the kernel size will quickly degrade our model performance under both pooling methods. Hence, to balance the trade-off between computational efficiency and overall accuracy, we employ the max pooling with the kernel size of 5 in our pooling layer.

[3] Notably, when the kernel size equals 2, our computational resources cannot afford the computational overhead incurred by self-attention.

Table 5. Comparative results of our ViT encoder with/without the positional embedding. The best results are bold

Case	Normal			Storm			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
w/ position	95.4	97.2	96.3	87.5	80.5	83.9	94.0
w/o position	**95.5**	**97.7**	**96.6**	**89.4**	**81.1**	**85.0**	**94.4**

5.6 Constructing Temporal Representations

In this section, we conduct experiments to support our design in Sect. 4.1, where we construct the temporal representation for a weather event by embedding three dates relevant to its beginning date in a month-interval manner (*i.e.*, two weeks before its beginning date, its beginning date, and two weeks after its beginning date). We conduct experiments to exhibit the impact of various time intervals on our model performance. In particular, if the time interval is set to T days, we construct its temporal representation by embedding T days before its beginning date, its beginning date, and T days after its beginning date. Figure 6 presents our experimental results in terms of precision, recall, and F1-score of predicting storm events as well as the overall accuracy, when varying T from 1 to 60. We discover that when the time interval T equals 14 (*i.e.*, two weeks), our proposed model architecture achieves the best performance under all metrics. The reason may be due to: (i) compared to a small time interval (*i.e.*, ≤ 7 days), the month interval manner (*i.e.*, time interval $= 14$ days) is more informative; and (ii) if the time interval is too large (*i.e.*, ≥ 30 days), the relevance between weather events and temporal information is hard to learn.

5.7 Impact of Positional Embedding on Our ViT Encoder

This section conducts experiments to support our design in Sect. 4.2, where we drop the positional embedding in our ViT encoder. Hence, two cases are taken into account, *i.e.*, our ViT encoder with or without the positional embedding. Table 5 presents our experimental results with and without the positional embedding. We observe that our ViT encoder with positional embedding actually hurts the prediction performance, with a performance degradation of 0.4% in terms of the overall accuracy (*i.e.*, 94.0% with the positional embedding versus 94.4% without the positional embedding). The reason is that no valuable positional information exists among the input sequences of our ViT encoder as they come from various sources of representation. To reduce the redundancy, we remove the positional embedding in our ViT encoder.

6 Conclusion

This work has developed a comprehensive Transformer-based model architecture for real-world storm prediction. Our model architecture separates its "rep-

resentation learning" and "prediction" stages to effectively extract the high-quality representations and accurately predict the occurrence of storm events, respectively. Multiple MAE-based encoders, the Word2vec tool, and a ViT-based encoder, with a collection of novel designs (such as image representation concatenation, temporal representation, and content embedding) are incorporated into our comprehensive model architecture to tackle practical challenges associated with real-world storm prediction. Experimental results exhibit the excellent learning capability of our model architecture. Although we conduct experiments on the SEVIR dataset, our model architecture can be generalized to effectively handle any type of real-world satellite and radar image data.

Acknowledgments. This work was supported in part by NSF under Grants 1763620, 2019511, 2146447, and in part by the BoRSF under Grants LEQSF(2019-22)-RD-A-21 and LEQSF(2021-22)-RD-D-07. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agencies.

References

1. Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., Hickey, J.: Machine learning for precipitation nowcasting from radar images. arXiv preprint arXiv:1912.12132 (2019)
2. Bao, H., Dong, L., Wei, F.: BEiT: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Chen, L., Wang, W., Mordohai, P.: Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In: Computer Vision and Pattern Recognition (CVPR) (2023)
4. Chen, R., Wang, X., Zhang, W., Zhu, X., Li, A., Yang, C.: A hybrid CNN-LSTM model for typhoon formation forecasting. GeoInformatica **23**, 375–396 (2019)
5. Chen, W., et al.: A simple single-scale vision transformer for object localization and instance segmentation. arXiv preprint arXiv:2112.09747 (2021)
6. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations (ICLR) (2020)
7. Cui, Y., Yan, L., Cao, Z., Liu, D.: TF-blender: temporal feature blender for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
8. Cuomo, J., Chandrasekar, V.: Developing deep learning models for storm nowcasting. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2021)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2019)
10. Domico, K., Sheatsley, R., Beugin, Y., Burke, Q., McDaniel, P.: A machine learning and computer vision approach to geomagnetic storm forecasting. arXiv preprint arXiv:2204.05780 (2022)
11. Dong, G., Tang, M., Cai, L., Barnes, L.E., Boukhechba, M.: Semi-supervised graph instance transformer for mental health inference. In: IEEE International Conference on Machine Learning and Applications (ICMLA) (2021)

12. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
13. Fan, H., et al.: Multiscale vision transformers. In: International Conference on Computer Vision (ICCV) (2021)
14. He, J., Wang, T., Min, Y., Gu, Q.: A simple and provably efficient algorithm for asynchronous federated contextual linear bandits (2022)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. CoRR (2021)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
17. He, Y., Lin, F., Yuan, X., Tzeng, N.: Interpretable minority synthesis for imbalanced classification. In: International Joint Conference on Artificial Intelligence (IJCAI) (2021)
18. He, Y., et al.: HierCat: hierarchical query categorization from weakly supervised data at Facebook marketplace. In: ACM Web Conference (WWW) (2023)
19. Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: International Conference on Computer Vision (ICCV) (2021)
20. Hinz, R., et al.: Towards very-low latency storm nowcasting through AI-based onboard satellite data processing. In: International Conference on Information and Knowledge Management Workshop (CIKM Workshop) (2021)
21. Hossain, M., Rekabdar, B., Louis, S.J., Dascalu, S.: Forecasting the weather of Nevada: a deep learning approach. In: International Joint conference on Neural Networks (IJCNN) (2015)
22. Jumper, J., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021)
23. Klocek, S., et al.: MS-nowcasting: Operational precipitation nowcasting with convolutional LSTMs at microsoft weather. arXiv preprint arXiv:2111.09954 (2021)
24. Kong, R., et al.: Getting the most from eye-tracking: user-interaction based reading region estimation dataset and models. In: Symposium on Eye Tracking Research and Applications (2023)
25. Lai, Z., Wang, C., Cheung, S.c., Chuah, C.N.: SAR: self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning. In: Computer Vision and Pattern Recognition (CVPR) (2022)
26. Lai, Z., Wang, C., Gunawan, H., Cheung, S.C.S., Chuah, C.N.: Smoothed adaptive weighting for imbalanced semi-supervised learning: improve reliability against unknown distribution data. In: International Conference on Machine Learning (ICML), pp. 11828–11843 (2022)
27. Li, J., Wang, W., Abbas, W., Koutsoukos, X.: Distributed clustering for cooperative multi-task learning networks. IEEE Trans. Netw. Sci. Eng. **596**, 583–589 (2023)
28. Li, X., Metsis, V., Wang, H., Ngu, A.H.H.: TTS-GAN: a transformer-based time-series generative adversarial network. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds.) Artificial Intelligence in Medicine (AIME) (2022)
29. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. arXiv preprint arXiv:2203.16527 (2022)
30. Li, Y., et al.: Improved multiscale vision transformers for classification and detection. arXiv preprint arXiv:2112.01526 (2021)

31. Lin, F., Yuan, X., Peng, L., Tzeng, N.: Cascade variational auto-encoder for hierarchical disentanglement. In: International Conference on Information & Knowledge Management (CIKM) (2022)
32. Lin, S.Y., Chiang, C.C., Li, J.B., Hung, Z.S., Chao, K.M.: Dynamic fine-tuning stacked auto-encoder neural network for weather forecast. Futur. Gener. Comput. Syst. **89**, 446–454 (2018)
33. Liu, D., Cui, Y., Cao, Z., Chen, Y.: A large-scale simulation dataset: Boost the detection accuracy for special weather conditions. In: 2020 International Joint Conference on Neural Networks (IJCNN) (2020)
34. Liu, D., Cui, Y., Tan, W., Chen, Y.: SG-Net: spatial granularity network for one-stage video instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
35. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (ICCV) (2021)
36. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR), Workshop Track Proceedings (2013)
37. Min, Y., He, J., Wang, T., Gu, Q.: Learning stochastic shortest path with linear function approximation. In: International Conference on Machine Learning (ICML) (2022)
38. Min, Y., Wang, T., Zhou, D., Gu, Q.: Variance-aware off-policy evaluation with linear function approximation (2021)
39. NOAA: The NOAA storm events database. https://www.ncdc.noaa.gov/stormevents/
40. Ravuri, S., et al.: Skillful precipitation nowcasting using deep generative models of radar. Nature **597**, 672–677 (2021)
41. Samsi, S., Mattioli, C.J., Veillette, M.S.: Distributed deep learning for precipitation nowcasting. In: High Performance Extreme Computing Conference (HPEC) (2019)
42. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems (NeurIPS) (2015)
43. Shi, X., et al.: Deep learning for precipitation nowcasting: a benchmark and a new model. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
44. Sønderby, C.K., et al.: MetNet: a neural weather model for precipitation forecasting. arXiv preprint arXiv:2003.12140 (2020)
45. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (ICML) (2021)
46. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
47. Veillette, M.S., Samsi, S., Mattioli, C.J.: SEVIR: a storm event imagery dataset for deep learning applications in radar and satellite meteorology. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
48. Wang, B., et al.: Deep uncertainty quantification: a machine learning approach for weather forecasting. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019)
49. Wang, W., et al.: Real-time dense 3d mapping of underwater environments. arXiv preprint arXiv:2304.02704 (2023)
50. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: International Conference on Computer Vision (ICCV) (2021)

51. Wang, Y., Gao, Z., Long, M., Wang, J., Philip, S.Y.: PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In: International Conference on Machine Learning (2018)
52. Wang, Y., Long, M., Wang, J., Gao, Z., Yu, P.S.: PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
53. Wang, Z., Li, T., Zheng, J., Huang, B.: When CNN meet with VIT: towards semi-supervised learning for multi-class medical image semantic segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13807, pp. 424–441. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25082-8_28
54. Wang, Z., Zhao, W., Ni, Z., Zheng, Y.: Adversarial vision transformer for medical image semantic segmentation with limited annotations. In: British Machine Vision Conference 2022 (2022)
55. Yuan, L., et al.: Tokens-to-token VIT: training vision transformers from scratch on ImageNet. In: International Conference on Computer Vision (ICCV) (2021)
56. Zhang, D., Zhou, F.: Self-supervised image denoising for real-world images with context-aware transformer. IEEE Access **11**, 14340–14349 (2023)
57. Zhang, D., Zhou, F., Jiang, Y., Fu, Z.: MM-BSN: self-supervised image denoising for real-world with multi-mask based on blind-spot network. In: Computer Vision and Pattern Recognition Workshop (CVPRW) (2023)
58. Zhang, W., Zhang, R., Chen, H., He, G., Ge, Y., Han, L.: A multi-channel 3D convolutional-recurrent neural network for convective storm nowcasting. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 363–366. IEEE (2021)
59. Zhang, Y., et al.: Precise weather parameter predictions for target regions via neural networks. In: Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) (2021)
60. Zhang, Z., He, Z., Yang, J., Liu, Y., Bao, R., Gao, S.: A 3D storm motion estimation method based on point cloud learning and doppler weather radar data. IEEE Trans. Geosci. Remote Sens. **60**, 1–5 (2021)