STES: A Spatiotemporal Explanation Supervision Framework

Dazhou Yu [*] Binbin Chen [†] Yun Li [*] Suman Dhakal [‡] Yifei Zhang [*]
Zhenke Liu [*] Minxing Zhang [*] Jie Zhang [‡] Liang Zhao [*]

Abstract

Explanation supervision is a technique that guides a deep learning model to have correct attention during training and thus improve both the interpretability and predictability of the model. However, the exploration of explanation supervision methods for spatiotemporal prediction has been limited. In this paper, we propose a framework for explanation-supervised spatiotemporal forecasting which aims to explicitly incorporate human-annotated spatiotemporal explanations as supervision signals, achieved by introducing a unique objective that integrates human explanations for general spatiotemporal predictive models. Specifically, to extend the explanation supervision technique to spatiotemporal prediction, our framework addresses several inherent challenges associated with spatiotemporal data. Firstly, it tackles the difficulty of identifying and correcting the spatiotemporal reasoning process. Secondly, it addresses the challenge of handling the absence of human explanation annotation through interpolation techniques. Lastly, it handles the varying influence of different time points. To evaluate the effectiveness of our approach, we conducted extensive experiments on two real-world spatiotemporal datasets. The results demonstrate the superiority of our methods in improving the interpretability of explanations and the performance of the backbone deep neural network models, surpassing existing state-of-the-art explanation supervision methods.

Keywords: explanation supervision, spatiotemporal, interpretability

1 Introduction

Deep learning models have demonstrated exceptional performance in various domains, including spatiotemporal prediction. However, the black-box nature of these models raises concerns about transparency and interpretability, as users often need to understand the reasoning behind AI's decisions. In response to the demand for transparency and trust in machine learning models, the field of explainable Artificial Intelligence (XAI) has witnessed significant advancements in recent years [6, 18]. These advancements aim to uncover the inner workings of black-box neural networks. One common approach involves generating saliency maps like Grad-CAM [18], which highlight the most influential sub-parts or features of the input [7, 8, 19]. The highlighted areas in an image are called attention (or saliency areas), which are the important areas based on which the model makes the prediction [6]. However, a critical question is what if the model focuses on the wrong areas for making the decision? The domain of explanation supervision is hence debuted to align the model explanation with the human-annotated explanations in terms of the saliency areas.

Recently, approaches have emerged to address explanation supervision in images by aligning the model's saliency areas with human annotations [7, 8, 19]. HAICS [19] and GRADIA [8] leverage human annotations as explanation supervision signals to enhance model interpretability and performance in image classification tasks. RES [7] further improves the robustness of the explanation supervision by considering the noise in the annotations. However, these works only focus on static image classification. In this paper, beyond individual image-based prediction, we are interested in prediction based on a sequence of images, which is an important formulation for spatiotemporal prediction tasks. Hence the explanations for each prediction are the saliency areas in the sequence of images instead of a single image. As exemplified in Figure 1, we want to predict whether or not a significant solar flare would happen based on the sequence of images in previous days, while identifying the areas that are indicative of the flare as early as possible. Astronomers have annotated the significant regions in this example with red-colored boxes, which serve as indicators for future solar flares. The color intensity within these boxes encodes the relative importance of these areas. Consequently, the model should prioritize its attention toward these human-annotated explanations to identify patterns and insights crucial for predicting solar flares.

While we possess annotated attention for spatiotemporal prediction tasks, the current explanation supervision methods are tailored for individual static images and cannot effectively address our explanation supervision needs within a sequence of images. This limitation arises due to specific technical challenges as follows: 1) **Difficulty in identifying and correcting the spatiotemporal reasoning process**. In Figure

[*]Emory University, {dyu62, yli230, yzh3443, zliu365, mzha329, lzhao41}@emory.edu
[†]University of Pennsylvania, chenbb@seas.upenn.edu
[‡]George Mason University, {sdhakal2, jzhang7}@gmu.edu

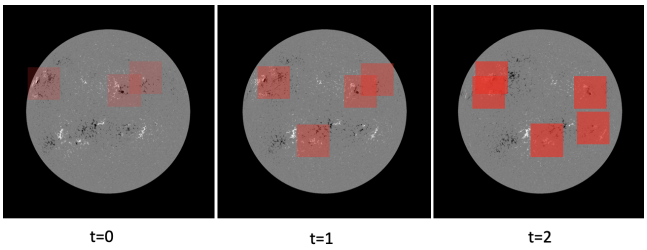


Figure 1: An example of the spatiotemporal explanation for solar flare prediction. (Human annotation only has the bounding box, color is not included.) Model attention should align with the annotated areas, and the influence weight of different timestamps should align with the color intensity.

1, we observe that the annotated area in the spatiotemporal data changes in a complex pattern, and only a small portion of the input requires focused attention. Indeed, noisy input can make deep learning models more vulnerable to potential overfitting issues. When a model is exposed to such noisy data during training, it may memorize the noise instead of learning the underlying patterns or features essential for generalization, leading to incorrect predictions. 2) **Difficulty in handling missing annotations.** In the case of a sequence of images over time, the necessary annotations are multiplied by the length of the sequence. However, it is important to consider that human expert annotation is costly. Furthermore, for specific tasks, the cadence of the input data might change, resulting in situations where annotations are available only at certain intervals, such as every 24 hours, while the input data is based on a 12-hour interval. There may be instances where the annotation is unavailable, making it difficult to determine the important area in that particular frame. 3) **Difficulty in handling changing influence at different time points.** The impact of different time points on the current event varies. However, evaluating the specific influence and assigning appropriate importance to each timestamp is a complex task, as it depends on the nature of the specific prediction task.

In this work, we propose the spatiotemporal explanation supervision (STES) as a generic explanation model designed for a sequence of images (i.e., rasterized spatial data) to tackle the above inherent difficulties associated with spatiotemporal tasks. We summarize our contributions as follows:

1. **We propose an innovative explanation supervision framework for spatiotemporal prediction tasks.** Unlike existing methods, our framework integrates human-annotated explanations as supervisory signals across all timestamps, ensuring temporal consistency in the attention of important areas. It is compatible with various predictive task models such as convolutional neural networks followed by recurrent neural networks, making it versatile and applicable in different scenarios.
2. **We propose a self-supervised explanation interpolation encoder-decoder module.** This addresses the challenge of missing human annotations in the time series data. Unlike traditional interpolation methods that rely on fixed-time relationships and neglect long-term dependencies, our method learns the evolving trends of the entire time series. It reconstructs annotations when they are available and interpolates missing ones, enabling our framework to maintain a consistent guiding mechanism over time.
3. **We introduce a time importance weighting mechanism.** The importance weighting mechanism assigns different weights to each timestamp based on the importance derived from the model's attention during the explanation loss calculation. This mechanism leverages a predefined decay function and adaptively considers the model attention for a specific sample.
4. **We conduct extensive experiments.** To demonstrate the efficacy of our method, We conduct extensive experiments on real-world datasets for solar flare prediction and hurricane tasks. The results demonstrate the effectiveness of our method, as it outperforms state-of-the-art models in terms of both interpretability and performance. These experiments provide empirical evidence to support the efficacy and superiority of our proposed framework.

2 Related Works

2.1 Spatiotemporal Prediction for Raster Data The increasing availability of spatiotemporal data and advancements in related techniques have sparked a growing interest in spatiotemporal data mining [13, 23, 26], particularly in the field of spatiotemporal prediction or forecasting [14, 20]. In spatiotemporal prediction, the goal is to develop a model that can predict a response variable based on explanatory features across different locations and times. The focus of this paper is on analyzing time-series raster data, which is characterized by its grid-like structure. The prediction of spatiotemporal raster data is crucial in fields like meteorology, where it's used for weather forecasting [10], and in environmental sciences for monitoring changes in land use or vegetation cover [4]. The methods often involve handling spatial and temporal information effectively, ranging from traditional statistical techniques to machine learning methods. Specifically, Convolutional Neural Networks (CNNs) are the most widely used method for spatial information learning [16] and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for temporal dynamics. Hybrid models that combine CNN and sequential models to jointly learn the spatial and temporal information are also popular like ConvLSTM [20] for precipitation nowcasting and wind speed forecasting [9].

2.2 Explanation Supervision The increasing use of complex neural networks has highlighted the need for explainable machine learning approaches to comprehend these opaque models [1]. Local explanations, which offer insights into

individual decisions made by the model, have garnered significant attention. Various feature attribution methods, such as gradient-based [21], surrogate [17], perturbation-based [15], and sensitivity analysis methods [3], have been developed to generate these local explanations. While these methods shed light on the decision-making process of the model, ongoing debates about their effectiveness and interpretability continue to shape the future development of explanation techniques [12, 25]. Beyond their usefulness for human understanding, learned explanations can also enhance model performance by incorporating explanatory information during the training process. Explanation supervision techniques [5], which utilize local explanations as supervisory signals, have been extensively studied in image [8, 19, 27], text [24], graph [6], and attribute [22] data domains. One such work is the HAICS [19] framework, which presents a conceptual framework for explanation supervision. HAICS has been further applied in image classification tasks, utilizing human annotation in the form of scribble annotations serving as explanation supervision signals. GRADIA [8] facilitates human involvement in identifying instances with unsatisfactory local explanations and directly adjusting them. By incorporating feedback from human users, this approach enhances both the performance and quality of explanations. RES framework [7] is devised to address the challenges of dealing with noisy annotations. This framework facilitates explanation supervision on Deep Neural Networks by incorporating both positive and negative explanation labels. However, the application of such explanation supervision techniques in the spatiotemporal domain, specifically with time-series raster data, requires further exploration. By developing and refining these techniques, we can bridge the gap between complex model behavior and human comprehension in spatiotemporal data analysis and prediction tasks.

3 Problem Formalization

The spatial raster data at a given timestamp t is denoted as $x^{(t)} \in \mathbb{R}^{C\times H\times W}$, which represents C observations over an $H \times W$ sized grid space. Let $y^{(t)}$ denote the label indicating whether an event will occur in the future from t. We assume that for each timestamp of interest t, we have access to a temporal sequence of length T including time t: $X = (x^{(t-T+1)}, x^{(t-T+2)}, \cdots, x^{(t)}) \in \mathbb{R}^{T\times C\times H\times W}$. The general goal of the spatiotemporal prediction problem is to predict the corresponding label $y^{(t)}$ by learning a mapping function f which leverages the information encoded in X:

$$\left[x^{(t-T+1)}, x^{(t-T+2)}, \cdots, x^{(t)}\right] \xrightarrow{f} y^{(t)}. \tag{3.1}$$

$M^{(i)} = (m^{(i,t-T+1)}, m^{(i,t-T+2)}, \cdots, m^{(i,t)})$ denotes the model explanation for the ith sample, which is a series of saliency maps generated by a post-hoc model explainer like Grad-CAM [18]. Each saliency map has the same size as the input raster $m^{(i,\cdot)} \in \mathbb{R}^{H\times W}$. The ground truth explanation $E^{(i)} = (e^{(i,t-T+1)}, e^{(i,t-T+2)}, \cdots, e^{(i,t)})$ is annotated by a human expert, where each annotation is a binary-valued mask $e^{(i,\cdot)} \in \{0,1\}^{H\times W}$ showing the areas of interest for the task. A trustworthy model's attention area should align with the human annotation.

Effectively extracting information from historical data is a complex task that involves overcoming several unique challenges. These challenges include: **1) Difficulty in identifying and correcting spatiotemporal reasoning process.** Predicting complex spatiotemporal events, like solar flare breakouts, involves the joint identification of important times and locations. To achieve accurate predictions, the model must rapidly adjust its attention $M^{(i)}$ over time and consistently focus on crucial areas $E^{(i)}$. Otherwise, the model may learn from irrelevant areas, incorporating noisy information, and increasing the risk of overfitting. Consequently, this could lead to incorrect predictions due to erroneous learning. **2) Difficulty in handling missing annotations.** $e^{(i,\cdot)}$ can have all-zero values in a missing case. The absence of annotations can hinder the understanding and interpretation of historical data, making it difficult to draw accurate conclusions or make reliable predictions. **3) Difficulty in handling changing influence at different time points.** The impact of various timestamps changes throughout the entire spatiotemporal process. $e^{(i,\cdot)}$ only gives the interest area and requires a further weighting mechanism to acquire its weight at the given time. Tackling this obstacle necessitates the development of methodologies capable of effectively identifying and filtering out irrelevant long-term influences.

4 Methodology

To overcome the challenges discussed earlier, we introduce a framework called Spatiotemporal Explanation Supervision (STES), depicted in Figure 2. In the subsequent sections, we outline the critical components of our framework. In Section 4.1, we present a general spatiotemporal explanation-supervised method that incorporates expert knowledge in a time-consistent way. This model serves as the foundation for our framework and enables the utilization of human-annotated explanations to guide the prediction process. In Section 4.2 and Section 4.3, we provide a detailed description of the different components comprising our complete framework. These sections outline the specific techniques and methodologies used to enhance our model's prediction accuracy and interpretability. By combining these components within the STES framework, we aim to address the limitations of existing approaches and achieve improved performance in spatiotemporal prediction tasks.

4.1 Framework As illustrated in Figure 2, the STES framework is composed of three modules. The upper part of Figure 2 depicts the first module, which consists of a deep

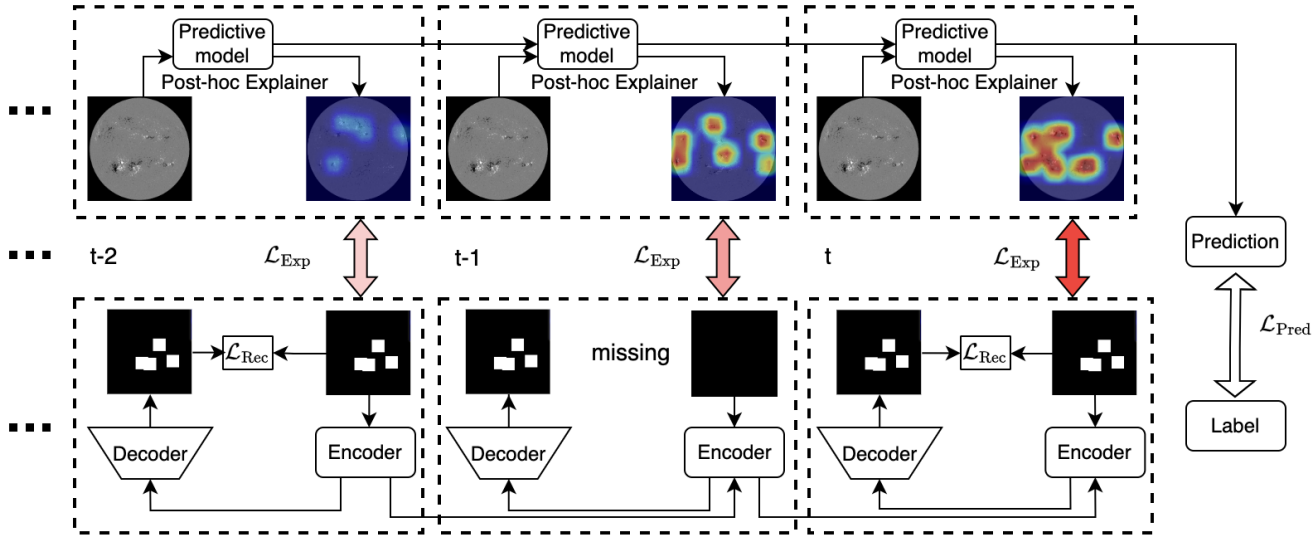


Figure 2: Illustration of the proposed STES framework. The upper part depicts a general spatiotemporal predictive model and the model attention saliency maps generated by a post-hoc model explainer. The model attention is compared with the human annotation. The color intensity of the arrows in the middle encodes the importance weights of different time points. The bottom part shows an interpolation encoder-decoder module for missing annotation interpolation.

neural network model f for spatiotemporal prediction and a post-hoc model explainer (e.g., Grad-CAM [18]). The explainer outputs a saliency map that indicates the important input areas at each timestamp. The middle part shows the time-aware explanation comparison module, responsible for calculating the explanation loss, as indicated by the colored arrows. Lastly, the bottom part represents the annotation interpolation encoder-decoder module, which interpolates missing annotations. The structure of f is flexible where we can use an encoder such as convolutional layers to learn spatial patterns which are then sequentially learned by recurrent networks such as a gated recurrent unit [2].

Since the input raster contains a lot of irrelevant areas that can impede the model from extracting meaningful representations, STES leverages time-consistent explanation supervision to mitigate their influence. By incorporating human-annotated explanation areas at each timestamp as supervision, our framework enables the model to learn valuable information consistently from the input time-series raster, thereby enhancing both its predictability and explainability. For each timestamp, a post-hoc explainer is employed to extract and visualize the attention of the model, which serves as the model explanation. Subsequently, the model explanation is normalized and compared to the human expert annotation, and an explanation loss is computed based on the comparison outcome to provide additional supervision.

On the other hand, the scarcity of image-annotation pairs poses a challenge to ensuring time-consistent explanation supervision. To overcome this limitation, our proposed approach introduces a spatiotemporal interpolation encoder-decoder module, which addresses the issue of missing annotations by providing interpolation. During training, the model utilizes an encoder to extract the spatiotemporal embeddings of the annotations. Subsequently, the embeddings are deconvolutionized to reconstruct annotations. By comparing these generated annotations with the input annotations, the model can be self-supervised to ensure its effectiveness. Based on the aforementioned statements, our framework encompasses three distinct losses: the prediction loss of the general spatiotemporal model, the explanation loss of the model, and the training loss of the interpolation encoder-decoder module. The overall objective function of the STES framework can be formulated as follows:

(4.2)
$$\min \sum_{i}^{N} \underbrace{\mathcal{L}_{\text{Pred}}\left(f(X^{(i)}), y^{(i,t)}\right)}_{\text{prediction loss}} + \underbrace{\sigma_1 \mathcal{L}_{\text{Exp}}\left(M^{(i)}, E^{(i)}\right)}_{\text{explanation loss}} + \underbrace{\sigma_2 \mathcal{L}_{\text{Rec}}\left(\hat{E}^{(i)}, E^{(i)}, \delta_s(E^{(i)})\right)}_{\text{reconstruction loss}}$$

The first term is the common prediction loss, where $f(X^{(i)})$ is the model prediction for the i-th input sequence $X^{(i)}$, $y^{(i,t)}$ is the label for the timestamp of interest, the loss is typically computed using a common loss function like cross-entropy. The second term is the explanation loss, often measured using metrics like L_1 loss, which quantifies the difference between the model explanation $M^{(i)}$ and the ground truth explanation $E^{(i)}$. The last term is the re-

construction loss, typically calculated using L_1 loss, for training the interpolation encoder-decoder module. Here, $\hat{E}^{(i)} = (\hat{e}^{(i,t-T+1)}, \hat{e}^{(i,t-T+2)}, \cdots, \hat{e}^{(i,t)})$ denotes the reconstructed human annotation from the decoder. The interpolation generation and training process are discussed in Section 4.2 and 4.3. $\delta_s(E^{(i)})$ is a filtering vector that selects the timestamps containing nonzero human annotations. The σ_1, σ_2 are two hyperparameters that control the loss balance. We also analyze the time complexity of our framework and the conclusion is that the added complexity compared to a simple predictive model f only arises from a second-order backpropagation, please see the appendix for details.

4.2 Self-supervised Explanation Interpolation While existing methods of explanation supervision lack the ability to utilize historical annotation, our framework addresses this limitation by ensuring that the model consistently focuses on the correct areas, allowing past explanations to contribute to the current predictions. However, since expert-annotated explanations are often missing for certain timestamps, we face the challenge of incomplete annotations throughout the time series, which hampers the model's ability to receive consistent guidance. To overcome this challenge and guide the model to continuously prioritize important areas in the input, we introduce an interpolation encoder-decoder module. In real-world scenarios, annotations can be missing at any time, and the number of missed annotations can vary. Traditional interpolation methods, which rely on fixed-time relationships and overlook long-term relationships, are inadequate for this task. In our approach, we propose an interpolation encoder-decoder module that learns the evolving trend of the entire time series raster data and interpolates missing annotations for any given timestamp. Specifically, for the timestamp of interest, denoted as t, if any of the annotations for the input rasters before t are missing, we replace them with 0, resulting in the time-series input annotation data denoted as $E^{(i)}$. We then feed $E^{(i)}$ into a convolutional model followed by a recurrent model to get the spatiotemporal representation, and use a deconvolution model to project the embedding back to the 2D raster space:

$$\hat{E}^{(i)} = g_{\text{Dec}}(\dot{E}^{(i)}) = g_{\text{Dec}} \circ g_{\text{RNN}} \circ g_{\text{CNN}}(E^{(i)}), \tag{4.3}$$

where $\dot{E}^{(i)} = (\dot{e}^{(i,t-T+1)}, \dot{e}^{(i,t-T+2)}, \cdots, \dot{e}^{(i,t)})$ is the learned latent representation of the annotation at each timestamp considering both the spatial and temporal information, $\hat{E}^{(i)}$ is the learned annotation projected from the latent embedding, $\circ$ denotes function composition, g_{Dec} denotes the deconvolution model that projects the learned latent embedding to the 2D raster space, g_{RNN} represents the recurrent module in the interpolater, g_{CNN} is a convolutional function that extracts the spatial information from 2D input raster. For timestamps where the annotation is not missing, the reconstruction output is used as supervision to train the interpolation encoder-decoder module. In cases where the annotation is missing, the model performs interpolation, and the resulting output is utilized to supervise the model's explanation.

$$\mathcal{L}_{\text{Rec}} = \sum_{t=1}^{T} |\hat{e}^{(i,t)} - e^{(i,t)}| \cdot \mathbf{1}(e^{(i,t)} \neq 0) \tag{4.4}$$

By incorporating the interpolation encoder-decoder module and leveraging the reconstruction loss and interpolated outputs, our framework effectively addresses the challenge of missing annotations, allowing the model to receive consistent guidance and improving the overall performance of the explanation-supervised framework.

4.3 Adaptive Time-aware Explanation Supervision Considering the potentially long duration of time series data, assigning equal importance to all past timestamps is inappropriate when calculating the explanation supervision loss. As the temporal influence decreases as the time distance increases, we need a time decay function to assign higher weights to time points that are closer to the present. Depending on the nature of the event, there are many different types of decay functions including linear, quadratic, and exponential. However, they all rely on a hyperparameter named decay coefficient to control the decay rate, which needs specific tuning and is hard to generalize to unseen data. Furthermore, spatiotemporal events like solar flares can manifest either suddenly or gradually, necessitating our consideration of varying decay rates in different instances. The attention maps generated by the explainer contain important information that can be utilized to learn a proper decay rate. Specifically, if the model's attention consistently exhibits significant saliency areas at each timestamp, it likely indicates that the event requires a lengthy period to develop, and we should employ a low decay rate. To discern the underlying pattern, we first extract the maximum from the saliency map at each timestamp using maximum pooling, then we feed them into an MLP model h_{d} to produce the decay coefficient parameter for the sample. Here we introduce the adaptive time decay function to account for the different influences of different timestamps: $\theta^{(i,t)} = \phi(h_{\text{d}}(\tilde{m}^{(i,t-T+1)}, \cdots, \tilde{m}^{(i,t)}), t)$, where ϕ is a general decay function whose output decreases with time t, $\tilde{m}^{(i,\cdot)} = \text{MAX}(m^{(i,\cdot)})$ is the maximum of the saliency map. This enables the model to adaptively assign different weights to each timestamp based on its temporal proximity to the interest timestamp. Moreover, since the human annotations are rough bounding boxes, even an ideal model would not have 100% overlapped attention with the annotation, so we do not require a strict alignment between them. To achieve this, we introduce a threshold α to selectively penalize timestamps with a loss that exceeds this threshold. The explanation loss calculation is formulated as a summation over all timestamps, where we take the maximum between the product of the time decay function and the explanation loss

for each timestamp minus α, and zero: $\mathcal{L}_{\text{Exp}}(M^{(i)}, E^{(i)}) = \sum_{t=1}^{T} \max\left\{\theta^{(i,t)} \cdot \mathcal{L}_{\text{Exp}}(m^{(i,t)}, e^{(i,t)}) - \alpha, 0\right\}$. This enables the model to capture the varying degrees of influence from different timestamps, leading to improved performance.

5 Experiment

We first introduce the experimental settings and the metric results. After that, we perform qualitative analysis, ablation studies, efficiency analysis, and the sensitivity of hyperparameters to help further explore the characteristics of the proposed framework. We randomly split each dataset into train/val/test with a radio of 6/2/2. Further details of our implementations can be found at https://github.com/dyu62/solar_share. All the experiments are conducted on a 64-bit machine with an NVIDIA A5000 GPU.

5.1 Experiment Setting

5.1.1 Dataset 1) **SolarFlare**: Line of sight magnetic field shown in the format of images. A sample is labeled as '1' if an M- or X- class flare occurs within 24 hours else labeled as '0'. The dataset is unbalanced with a positive rate of 17.8%. The period for the dataset is from 01/01/2012 to 05/31/2012. 2) **Hurricane**: Tropical Cyclone for Image-to-intensity Classification dataset contains tropical cyclone data from satellites to support the estimation of tropical cyclone rapid intensification. Only 3% of the samples are positive that show rapid intensification. In this study, frames from all tropical cyclones in Eastern North Pacific in 2004 are collected every 3 hours. Each frame contains 201*201 data points around the center of a tropical cyclone, and each frame is associated with a binary label indicating whether the tropical cyclone undergoes rapid intensification.

5.1.2 Comparison Methods We conduct a comprehensive performance comparison of the STES framework against three existing explanation supervision methods alongside the vanilla backbone model as the baseline. Specifically, the following methods were examined: (1) Baseline: As the backbone of other explanation-supervision models, this is a simple spatiotemporal predictive model composed of a ResNet18 [11] and a single layer GRU [2] that is solely trained with the prediction loss. (2) GRADIA [8]: This framework trains the backbone model using both the prediction loss and a traditional L_1 loss, directly minimizing the distance between the model explanation and the explanation labels at the target time point. (3) HAICS [19]: In this approach, the backbone model is trained using both the prediction loss and a conventional Binary Cross-Entropy (BCE) loss, directly minimizing the distance between the model explanation and explanation labels at the target time point. (4) RES [7]: A framework designed to handle the inaccurate boundary, incomplete region, and inconsistent distribution when applying noisy annotations. We used publicly available implementations of the comparison methods introduced in [7, 8].

5.2 Effectiveness Analysis

5.2.1 Evaluation Metrics We evaluate the framework on both its interpretability and performance using the following metrics: 1) Intersection over Union (IoU) score: a measurement of the overlap between model attention and annotated explanation mask, to determine the quality of explanation. 2) Accuracy (Acc). 3) Area Under the Curve (AUC), under the Receiver Operating Characteristic (ROC) curve, which represents the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR). 4) Precision. 5) Recall. 6) F1 score. 7) True Skill Statistic (TSS): the difference between the TPR and FPR. It is a normalized measure of accuracy that is insensitive to class-imbalanced datasets.

5.2.2 Performance The metric results are summarized and presented in Table 1. The best result for each dataset is highlighted in bold, and the standard deviation over three runs is reported following the $\pm$ mark. The superiority of the proposed method becomes apparent in the results, as it consistently achieves the highest scores for AUC, F1, and TSS on all datasets. Notably, for the solar flare prediction task, STES improves the F1 score by 30% compared to the best comparison method. Additionally, STES attains the best IoU and Acc scores on the SolarFlare dataset. Although the comparison methods occasionally outperform STES in certain metrics (e.g., GRADIA and HAICS delivering the best Precision scores on the SolarFlare and Hurricane datasets respectively), their overall prediction performance is not comparable to that of STES. These findings indicate that STES effectively guides the model to enhance both interpretability and performance, resulting in superior outcomes.

5.3 Qualitative Analysis of the Explanation To better demonstrate the model explanation improvements, here we provide a case study of the model-generated explanation for the solar flare dataset. We present the model-generated explanations as heatmaps overlaid on the original input samples, where warmer color is applied to more important areas. The explanation of the model trained using STES (shown in Figure 3) demonstrates superior accuracy in consistently highlighting the crucial areas over time for identifying solar flare breakouts. In contrast, both the baseline model and the three comparison methods failed to produce accurate explanations that remained consistent across time. For instance, the explanations generated by the baseline, HAICS, and RES methods assigned importance primarily to the last time point, disregarding the preceding timestamps. On the other hand, while GRADIA attends to past timestamps, it fails to accurately focus on the correct area compared

Table 1: The performance of the proposed model and the comparison methods.

SolarFlare							
Method	IoU	Acc	AUC	Precision	Recall	F1	TSS
Baseline	0.073±0.007	0.862±0.016	0.882±0.016	0.619±0.013	0.635±0.073	0.628±0.053	0.535±0.073
GRADIA	0.511±0.117	0.871±0.025	0.843±0.016	**0.992±0.090**	0.286±0.055	0.444±0.064	0.286±0.056
HAICS	0.023±0.046	0.819±0.006	0.396±0.035	0.181±0.244	**1.000±0.000**	0.307±0.106	0.100±0.050
RES	0.396±0.081	0.810±0.019	0.810±0.021	0.484±0.150	0.761±0.083	0.577±0.066	0.546±0.008
STES	**0.717±0.061**	**0.931±0.001**	**0.968±0.010**	0.783±0.072	0.857±0.052	**0.818±0.161**	**0.805±0.052**
Hurricane							
Method	IoU	Acc	AUC	Precision	Recall	F1	TSS
Baseline	0.069±0.024	0.965±0.021	0.945±0.028	0.952±0.083	0.476±0.360	0.569±0.305	0.459±0.356
GRADIA	0.500±0.139	0.961±0.015	0.927±0.011	0.833±0.289	0.429±0.297	0.571±0.101	0.411±0.278
HAICS	0.041±0.029	0.972±0.006	0.931±0.010	**1.000±0.000**	0.476±0.165	0.603±0.163	0.476±0.165
RES	**0.586±0.115**	**0.974±0.017**	0.946±0.082	0.809±0.330	0.333±0.165	0.426±0.175	0.326±0.159
STES	0.535±0.111	0.972±0.009	**0.951±0.003**	0.869±0.227	**0.556±0.220**	**0.631±0.055**	**0.517±0.156**

to our method. Based on the visualizations from the case study, we argue that models trained with the STES framework possess greater robustness and improved generalizability to downstream predictive tasks by effectively learning to assign importance to the critical areas presented in the data samples in a time-consistent manner.

5.4 Ablation Study To investigate the benefits gained from providing consistent guidance through interpolating missing annotations and considering the time decay effect by assigning different weights to timestamps, we conducted two ablation experiments. In the first experiment, we remove the interpolation module $g_{\text{Dec}} \circ g_{\text{RNN}} \circ g_{\text{CNN}}$ from our model, resulting in a version named STES-I. In the second experiment, we remove $\theta^{(i)}$ and treat each timestamp equally for guiding the model, leading to a version called STES-T. We evaluate the performance of these two ablation models on the SolarFlare dataset, and the results are reported in Table 2. Although the two ablation models achieve better scores on certain metrics, both the F1 score and TSS scores show a decrease compared to the original model. The STES-I method achieves the highest AUC and recall score, but these metrics alone cannot fully reflect the overall model performance. In terms of the more comprehensive F1 and TSS scores, STES outperforms STES-I. On the other hand, STES-T achieves the highest IoU score as it disregards the time decay effect and treats all timestamps equally. Consequently, it performs best only on the IoU score but fails to deliver satisfactory results on all other metrics.

5.5 Efficiency Analysis Here we evaluate the average training runtime per epoch for all methods. As indicated in Table 3, the forward process demonstrates significantly faster performance compared to backpropagation. Among the static explanation supervision methods, their runtimes are almost identical, while our method slightly lags behind them. This discrepancy arises from the need to calculate the attention map for all timestamps in our approach. However, the overall increase in the runtime of our method is not substantial and is almost the same as the backward time of the baseline model. The total runtime exceeds the combined time for the forward and backward processes due to the utilization of batch training and the constant time required for data loading, which applies to all methods.

5.6 Sensitivity Analysis We conduct sensitivity analysis on three hyperparameters: the explanation loss coefficient σ_1, the reconstruction loss coefficient σ_2, and the penalty threshold α. We report the changes in both explanation performance (IoU) and task performance (F1, TSS) as shown in Figure 4. The baseline performance is denoted as the yellow dashed line. As can be seen in Figure 4a, the optimal value for σ_1 is around 0.1 and 1. The general "n" shape is potentially reasonable as the model needs to balance between explanation and performance. The sensitivity of σ_2 is shown in Figure 4b. This curve is flat, showing that this parameter is not that sensitive. When σ_2 is large, the model puts more effort into training the interpolation encoder-decoder module, which can provide a more consistent guide when the annotation is missing, leading to a better explanation. What's more, we notice the performance is always above the baseline, which proves our model is robust in terms of this hyperparameter. The sensitivity of α is shown in Figure 4c. This figure has the sharpest shape, showing that this parameter is the most sensitive one. When α is large, the model doesn't penalize the explanation loss, which cannot provide any guide. When α is small, the model is too strict and may penalize correct explanations due to the coarse annotation.

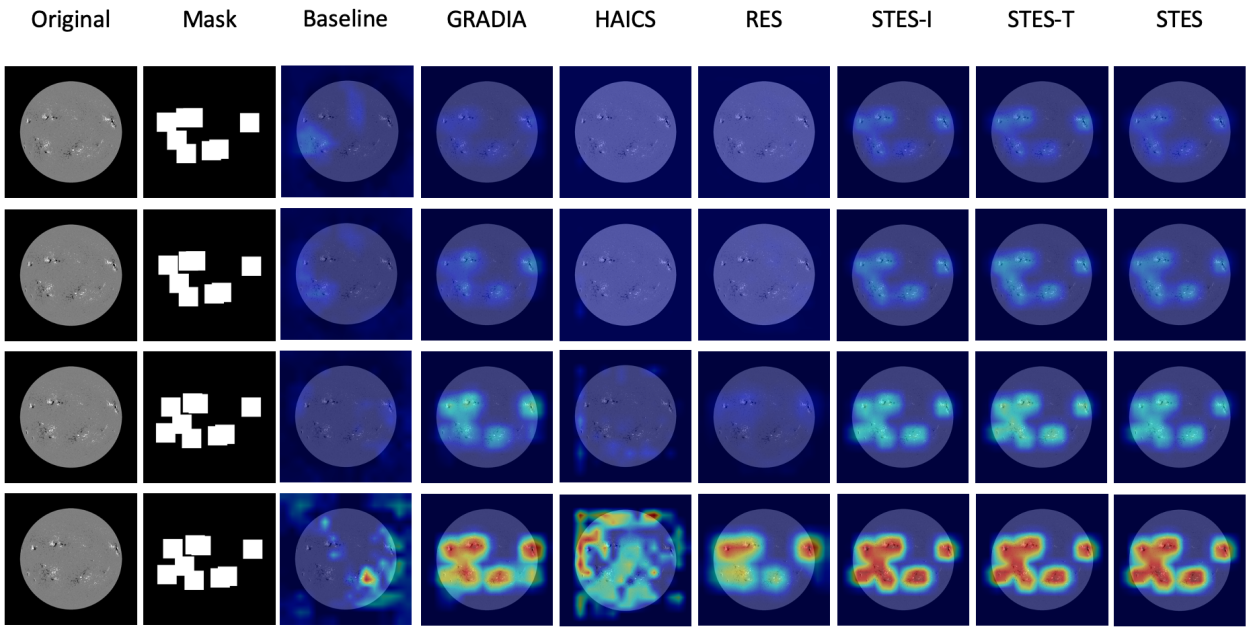
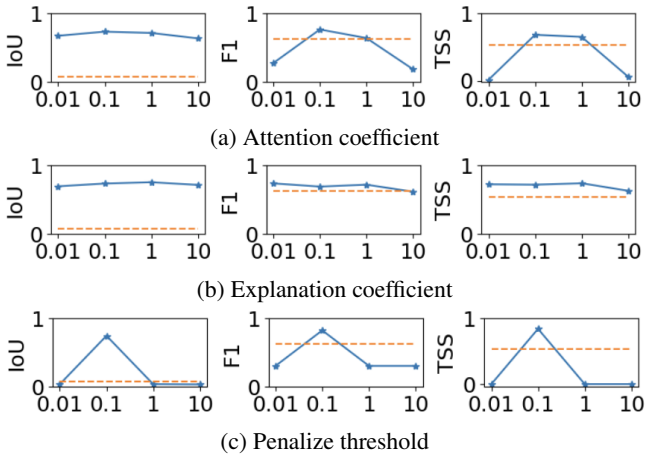


Figure 3: Model explanation visualization for the SolarFlare dataset. The original column is the input raster sequence (time length is 4 in this case) and the Mask column is the human-annotated ground truth explanation.

Table 2: The performance of the ablation variants of STES.

Variants	IoU	Acc	AUC	Precision	Recall	F1	TSS
STES-I	0.758±0.054	0.897±0.043	**0.979±0.004**	0.655±0.161	**0.905±0.028**	0.760±0.077	0.799±0.042
STES-T	**0.772±0.085**	0.914±0.026	0.960±0.002	0.739±0.155	0.810±0.051	0.773±0.084	0.746±0.005
STES	0.717±0.061	**0.931±0.001**	0.968±0.010	**0.783±0.072**	0.857±0.052	**0.818±0.161**	**0.805±0.052**



(a) Attention coefficient

(b) Explanation coefficient

(c) Penalize threshold

Figure 4: Sensitivity analysis for hyperparameters

Table 3: Running time for Training and Testing

	Forward (s)	Backward (s)	Total (s)
Baseline	0.9	6.0	12.5
GRADIA	1.5	6.3	13.3
HAICS	1.4	6.3	13.1
RES	1.7	6.2	13.6
STES	1.9	11.0	18.6

6 Conclusion

We propose STES, a novel explanation-supervision framework for spatiotemporal prediction with an interpolation encoder-decoder module and the time importance weighting mechanism. It provides consistent guidance and improves the deep model's explainability and performance at the same time. Experiments on two real-world spatiotemporal image datasets, SolarFlare and Hurricane, show that while existing explanation supervision models encounter limitations during the spatiotemporal reasoning process, our framework can achieve superior explainability and predictability.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[2] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[3] Paulo Cortez and Mark J Embrechts. Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348. IEEE, 2011.

[4] Evidence Chinedu Enoguanbhor, Florian Gollnow, Jonas Ostergaard Nielsen, Tobia Lakes, and Blake Byron Walker. Land cover change in the abuja city-region, nigeria: Integrating gis and remotely sensed data to support land use planning. *Sustainability*, 11(5):1313, 2019.

[5] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *arXiv preprint arXiv:2212.03954*, 2022.

[6] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. Gnes: Learning to explain graph neural networks. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 131–140. IEEE, 2021.

[7] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 432–442, 2022.

[8] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022.

[9] Amir Ghaderi, Borhan M Sanandaji, and Faezeh Ghaderi. Deep forecast: Deep learning-based spatio-temporal forecasting. *arXiv preprint arXiv:1707.08110*, 2017.

[10] Alessio Golzio, Silvia Ferrarese, Claudio Cassardo, Gugliemina Adele Diolaiuti, and Manuela Pelfini. Land-use improvements in the weather research and forecasting model over complex mountainous terrain and comparison of different grid sizes. *Boundary-Layer Meteorology*, 180(2):319–351, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[13] Zhe Jiang. A survey on spatial and spatiotemporal prediction methods. *arXiv preprint arXiv:2012.13384*, 2020.

[14] Ting Li, Junbo Zhang, Kainan Bao, Yuxuan Liang, Yexin Li, and Yu Zheng. Autost: Efficient neural architecture search for spatio-temporal prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 794–802, 2020.

[15] Holger R Maier and Graeme C Dandy. The use of artificial neural networks for the prediction of water quality parameters. *Water resources research*, 32(4):1013–1022, 1996.

[16] Ngoc-Thanh Nguyen, Minh-Son Dao, and Koji Zettsu. Leveraging 3d-raster-images and deepcnn with multi-source urban sensing data for traffic congestion prediction. In *Database and Expert Systems Applications: 31st International Conference, DEXA 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings, Part II 31*, pages 396–406. Springer, 2020.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[19] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton Van Den Hengel, and Johan W Verjans. Human-ai interactive and continuous sensemaking: A case study of image classification using scribble attention maps. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2021.

[20] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[22] Roman Visotsky, Yuval Atzmon, and Gal Chechik. Few-shot learning with per-sample rich supervision. *arXiv preprint arXiv:1906.03859*, 2019.

[23] Senzhang Wang, Jiannong Cao, and S Yu Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020.

[24] Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*, 2021.

[25] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[26] Minxing Zhang, Dazhou Yu, Yun Li, and Liang Zhao. Deep spatial prediction via heterogeneous multi-source self-supervision. *ACM Transactions on Spatial Algorithms and Systems*, 9(3):1–26, 2023.

[27] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1977–1987, 2023.