Local Geomagnetic Component Modeling of Auroral Images Based on Local-Global Feature

WANG Bo[1*], *ZHANG Yuanshu*[1], *CHENG Wei*[2], *TIAN Xinqin*[1], *SHENG Qinghong*[1], *LI Jun*[1], *LING Xiao*[1], *LIU Xiang*[1]

1. College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China;
2. Beijing Institute of Applied Meteorology, Beijing 100029, P. R. China

(Received 5 September 2025; revised 22 November 2025; accepted 3 December 2025)

Abstract: Accurately predicting geomagnetic field is of great significance for space environment monitoring and space weather forecasting worldwide. This paper proposes a vision Transformer (ViT) hybrid model that leverages aurora images to predict local geomagnetic station component, breaking the spatial limitations of geomagnetic stations. Our method utilizes the ViT backbone model in combination with convolutional networks to capture both the large-scale spatial correlation and distinct local feature correlation between aurora images and geomagnetic station data. Essentially, the model comprises a visual geometry group (VGG) image feature extraction network, a ViT-based encoder network, and a regression prediction network. Our experimental findings indicate that global features of aurora images play a more substantial role in predicting geomagnetic data than local features. Specifically, the hybrid model achieves a 39.1% reduction in root mean square error compared to the VGG model, a 29.5% reduction compared to the ViT model and a 35.3% reduction relative to the residual network (ResNet) model. Moreover, the fitting accuracy of the model surpasses that of the VGG, ViT, and ResNet models by 2.14% 1.58%, and 4.1%, respectively.

Key words: ultraviolet aurora image; geomagnetic field prediction; vision Transformer (ViT) hybrid model

CLC number: P352 **Document code:** A **Article ID:** 1005-1120(2025)06-0710-18

0 Introduction

Solar wind, influenced by the interplanetary magnetic field, propagates through the vast space between the Sun and Earth and subsequently interacts with Earth's magnetic field. This interaction leads to a series of disturbances in the magnetosphere, ionosphere, and auroral zone, including magnetic storms, substorms, and auroral phenomena. As a result, predicting solar-terrestrial phenomena and related magnetic activity has become a major focus of space research[1-3]. The impact of geomagnetic disturbances such as magnetic storms and substorms extends to various critical systems, including satellites, space stations, power grids, communications, navigation, and aviation. Consequently, the monitoring and prediction of geomagnetic activity and the development of relevant models are crucial aspects of space weather research[4-6].

The monitoring of geomagnetic stations is a highly effective approach to continuously and comprehensively evaluate global magnetospheric activity with a high temporal resolution. This datum obtained from the geomagnetic stations serves as a critical parameter for researchers, enabling them to gain insights into the spatial environment and elucidate the energy coupling between the magnetosphere and the ionosphere. The geomagnetic index, as determined from station measurements, is limited in its ability to assess geomagnetic disturbances

*Corresponding author, E-mail address: wangbo_nuaa@nuaa.edu.cn.

How to cite this article: WANG Bo, ZHANG Yuanshu, CHENG Wei, et al. Local geomagnetic component modeling of auroral images based on local-global feature[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025, 42(6): 710-727.

http://dx.doi.org/10.16356/j.1005-1120.2025.06.002

only within the coverage area of the station. Consequently, significant aurora events occurring beyond the station's coverage area might not be detected[7]. Efforts have been made by Newell and others to enhance the spatial resolution and detection rate of substorm indices by expanding the number of stations and leveraging data from approximately 100 magnetic observatories in the northern hemisphere[8-10]. However, the pursuit of improved prediction accuracy of geomagnetic indices has been hindered by the constraints of land availability, leading to a saturation in the accuracy of geomagnetic index prediction, and a difficulty in furthering the substorm identification rate[11]. Over the years, there has been significant progress in monitoring geomagnetic changes within specific areas at rapid speeds; however, the challenge remains in providing such data on a global scale. The process of substorm occurrence is often accompanied by dramatic changes in aurora morphology and brightness, revealing the connection between the aurora phenomenon and geomagnetic data. As a sensor of solar wind acting on the geomagnetic field, the aurora is a significant manifestation of geomagnetic disturbance, especially geomagnetic substorms, and another manifestation of magnetospheric activity[12]. Satellite-borne optical imagers, for example POLAR and IMAGE, have the capability to obtain information that cannot be provided by ground-based optical imagers, such as the polar and equatorial boundaries of the aurora egg, the overall morphology of the auroral oval, and the spatial distribution of the intensity of the auroral oval. Moreover, these satellites can perform multi-band imaging of the aurora, detect plasma entering the polar region and magnetotail of the magnetosphere, plasma entering and exiting the ionosphere, and the energy of particles sinking into the ionosphere and the upper atmosphere[13]. With the increase in the number of stations and data processing capabilities, correlating fixed-point observation of geomagnetic with the large-scale imaging of aurora has become an easier problem to deal with. The abundant satellite aurora imaging data and geomagnetic station monitoring data provide opportunities for constructing new data-driven geomagnetic index models.

The successful application of data-driven deep learning methods in computer vision, natural language processing, and other fields has paved the way for a new technological trend in the field of remote sensing: spatial-temporal data mining based on deep learning methods. Auroral images, as typical spatial-temporal data, have the capability to capture a relatively complete auroral oval in a relatively short period of time, making them highly desirable for various purposes. Leveraging the powerful nonlinear mapping and learning abilities of artificial neural networks, a Satellite image data-driven model between aurora intensity and geomagnetic data is established as a significant supplement to traditional methods. Aurora intensity variation has been extensively studied and found to be modulated by interplanetary magnetic field and solar wind parameters[14-15]. Meng et al.[16] introduced the global auroral power (GAP) as a new indicator of geospace activities. Subsequent research has demonstrated a strong correlation between the GAP index and the one-minute rapid observation auroral electrojet (AE) index. Liou et al.[17] conducted a comparative analysis of a large number of auroral images and revealed a robust correlation between the auroral power (AP) and the AE index, particularly showing a better correlation in winter than in summer. Liu et al.[18] found a high correlation coefficient of 0.76 between the mean energy of auroral precipitation particles (Pa) and the geomagnetic AE index, based on their analysis. In addition, Mitchell et al.[19] introduced the OVATION-SM model, which divided auroral intensity into a grid of 0.25 magnetic local time (MLT) $\times$0.5 magnetic latitude (MLAT). The model utilizes multiple linear regression and stepwise regression to express the auroral intensity of each grid as a linear combination of the SME index, the time of the last substorm occurrence, and the time of the next substorm occurrence. However, OVATION-SM is constructed based on location-independent variables and does not capture the detailed auroral morphology, which is more closely linked to MLT variables. Yang et al.[15] employed six space environmental parameters to model the boundary of the auroral oval, enabling prediction of its spatial location but not detailed information such

as the spatial distribution of auroral intensity. In contrast, Hu et al.[20] characterized the distribution of auroral intensity by using curve fitting methods and grid method to construct a database of auroral intensity characteristics in the polar region. From ultraviolet (UV) image data, they extracted the curve characteristics of auroral intensity along the magnetic latitude direction and constructed two auroral intensity prediction models with interplanetary/solar wind parameters and AE index as input parameters. The correlation between the aurora phenomenon and the geomagnetic index is evident in the research findings, despite the absence of a direct causal relationship. To forecast the detailed characteristics of the auroral oval, particularly its spatial intensity distribution, more refined data from geomagnetic stations with higher spatial resolution are necessary. Conversely, the auroral oval encompasses a broader spatial range, and the detailed spatial distribution of aurora intensity offers assistance in predicting geomagnetic station data through aurora images.

Convolutional neural networks (CNNs) have demonstrated outstanding performance in computer vision tasks, largely attributed to their utilization of the convolution operation. This operation facilitates the collection of local features in a hierarchical manner, ultimately leading to improved image representations. While CNNs excel in local feature extraction, they are often found deficient in their ability to capture global representations. In recent years, the Transformer model, which is based on the self-attention mechanism, initially demonstrated remarkable performance in the field of natural language processing[21]. This capability has sparked numerous studies seeking to leverage the potent modeling capabilities of the Transformer model in computer vision and multimodal remote sensing data analysis tasks[22-24]. The exceptional modeling ability of long-distance correlation and emphasis on global features in input data positions the Transformer model as an exemplary solution for language translation and related domains. This relevance has been furthered with the introduction of the visual Transformer (ViT) structure, bringing the Transformer into the arena of computer vision[25]. The ViT model represents a fully self-attention-based image classification system and stands as the pioneering work that replaces the standard convolution with the Transformer. The ViT method achieves this through the segmentation of images into patches and the subsequent generation of tokens with position embeddings, followed by the extraction of parameterized vectors as visual representations using a Transformer block. This breakthrough has led to the emergence of several visual Transformers, such as DeiT and Swin Transformer, which have found applications in diverse computer vision-based tasks[26-28]. Notably, the performance of visual Transformers has been found to be comparable to, or even surpassing, that of CNN, thereby solidifying their importance in this domain. However, Transformer focus on global features leads to a neglect of local feature details, resulting in decreased discriminability between background and foreground. Consequently, several approaches have emerged aiming to enhance representation learning by fusing local features from convolutional neural networks with global representations from Transformers. Notable models embodying this integration include the Conformer model[29], the CMT model[30], and so on.

The relationship between aurora image data and geomagnetic station data manifests not only in the broad spatial distribution of aurora intensity, but also in the nuanced characteristics of aurora morphology. As a result, this study utilizes the ViT structure as the core model, integrates convolutional networks to capture aurora image features, and employs regression prediction methods to forecast geomagnetic station data. In contrast to conventional image classification, object extraction, and change monitoring tasks, the prediction task in this research can be segmented into two distinct components: Feature extraction and regression prediction. While the Transformer model has demonstrated efficacy in a wide range of related tasks across different disciplines, its direct application to the prediction of geomagnetic data using aurora images as input presents the following challenges:

(1) Aurora data are a two-dimensional image sequence obtained from satellite imaging, while geomagnetic data are a one-dimensional array sequence composed of measurements from multiple ground

stations, and these data do not correspond in latitude. In order to predict geomagnetic data, the features of the aurora image are initially extracted and then encoded into a one-dimensional feature vector.

(2) There is a large spatial correlation and time dependency between aurora images and magnetometer monitoring values, and the local variation characteristics of aurora morphology and brightness are also strongly correlated with the strength of magnetometer monitoring values. The Transformer model structure has a natural and good modeling ability for global features and long-distance correlations of input information, but its ability to obtain local information is not as strong as CNN.

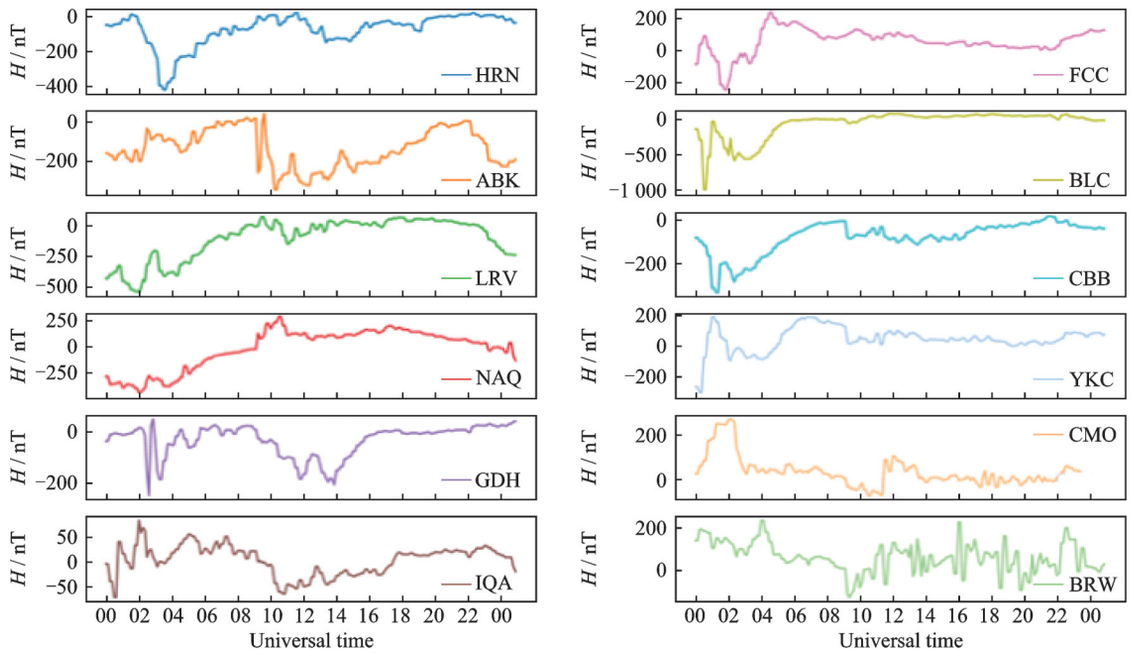
This paper presents a deep learning model for predicting local geomagnetic station component using aurora images, incorporating the ViT structure as the foundational framework and integrating it with convolutional networks. By leveraging this approach, the model has the capacity to effectively capture both large-scale spatial correlations and small-scale local features in the relationship between aurora images and geomagnetic station data, achieving ground-based geomagnetic data prediction based on aurora image sequences.

1 Data

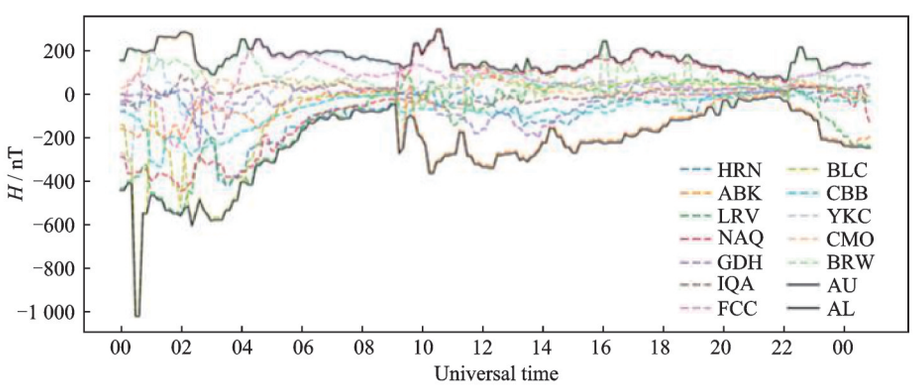
This paper focuses on establishing a preliminary correlation between auroral images and the horizontal H component of the magnetometers at local geomagnetic stations. Auroral satellite images are able to capture most, if not all, of the auroral eggs and have a wide range of coverage, while local geomagnetic station component refer to the horizontal H component of the magnetometers at geomagnetic stations. It is important to note that only geomagnetic data that correspond to the same moment of the image capture can be obtained through predictive models of geomagnetic data based on auroral images, and their temporal resolution is identical to that of the auroral images. The variation curves of the model's output data (geomagnetic data of each station) throughout 1 January, 1997, are demonstrated in Fig.1, and the location information is shown in Table 1. The goal of this paper is to focus on the extraction of auroral image features based on deep learning models for regression prediction of geomagnetic data. The prediction task within this context can be described as a regression prediction problem, where a deep learning model is trained with large amounts of image and geomagnetic data to predict the corresponding geomagnetic data from a given auroral image. It is crucial to emphasize that the forecasts in this study do not include time lead times and are not time series predictions in the traditional sense.

As shown in Fig.2, the distribution of high-latitude stations in the Arctic region is uneven due to land availability constraints, with a limited number of stations distributed within the range of 12UT—21UT. In addition, the lack of data from most stations in 1997 has resulted in a limited number of available station data. Therefore, where data are available, the stations selected in this article are evenly distributed along magnetic longitude as far as possible, and the data from stations at the same magnetic longitude will be compared and discussed. The distribution of the 12 stations selected in this article is shown as the green origin in Fig.2. The local geomagnetic data is obtained from the World Data Center (WDC, http://wdc.kugi.kyoto-u.ac.jp/), with a time resolution of 1 min. The preprocessing of local geomagnetic station component adopts the same baseline removal method as the AE index, using the average change of the five international magnetic quiet days per month to eliminate the change of quiet days. Before model training, it is also necessary to perform time matching between aurora images and local geomagnetic station component.

NASA's SPDF website (https://spdf.gsfc.nasa.gov/pub/data/polar/) provides the ultraviolet index (UVI) level data product under the UVI sensor carried by the POLAR satellite. The image size of the auroral oval in this product measures 200 pixel $\times$228 pixel, with a spatial resolution of approximately 0.04° per pixel. Due to the significant impact of oxygen Schumann Rungeband absorption on the Lyman-Birge-Hopfield short (LBHS) band, this study utilizes the Lyman-Birge-Hopfield long (LB-



(a) Time-varying geomagnetic variations at multiple stations and derived indices on 28 January, 1997



(b) Composite time-varying curves of geomagnetic data from multiple stations and derived AU/AL indices during the period of 28 January, 1997

Fig.1 Time-varying geomagnetic variations at multiple stations and derived indices on 28 January, 1997

Table 1 Detailed information of geomagnetic stations

Observatory(Abbr.)	Geomagnetic		Geographic	
	Latitude	Longitude	Latitude	Longitude
Hornsund(HRN)	74.17	123.95	77.00	15.55
Abisko(ABK)	66.14	113.53	68.35	18.82
Leirvogur(LRV)	68.75	69.83	64.18	338.30
Narsarsuaq(NAQ)	68.99	38.25	61.20	314.60
Godhavn(GDH)	77.64	33.10	69.25	306.47
Iqaluit(IQA)	72.97	6.33	63.75	291.48
Fort Churchil(FCC)	67.12	330.71	58.75	265.91
Baker Lake(BLC)	72.44	325.12	64.31	263.98
Cambridge Bay(CBB)	76.05	307.29	69.12	254.96
Yellowknife(YKC)	68.48	302.44	62.48	245.51
College(CMO)	65.50	264.45	64.87	212.14
Barrow(BRW)	69.97	249.27	71.32	203.38

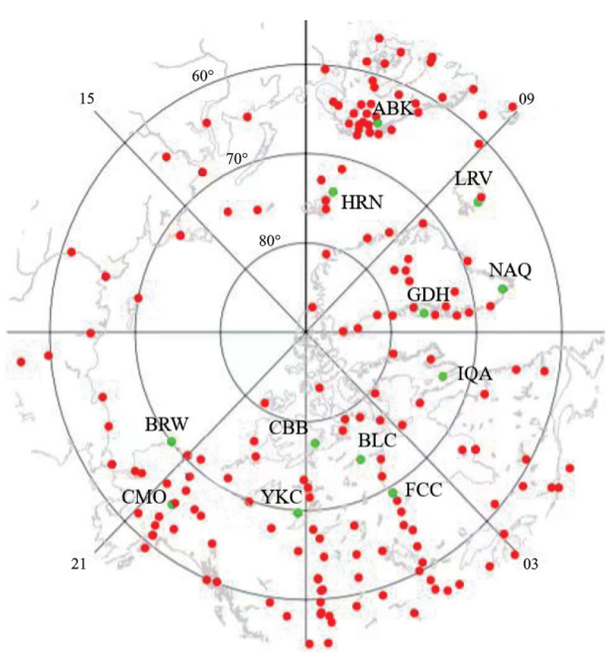


Fig.2 Distribution of geomagnetic stations

HL) band (160—180 nm) of the POLAR satellite. In the normal observation mode, the time resolution between two consecutive LBHL image ranges from 0.5 min to 3 min. To mitigate the effects of solar glare on the UVI image data, the dataset comprises 32 603 LBHL band (~170 nm) UVI images depicting complete auroral ovals over a three-month period from January to December 1997. Importantly, the auroral oval region in the northern hemisphere is situated in the polar night zone during this period, thereby minimizing the impact of solar glare on the aurora image. Prior to utilizing UVI data for modeling purposes, it is imperative to conduct preprocessing measures. Using the ENVI software and the interactive data language (IDL), we imported the image and sensor platform data from the ".cdf" files. By selecting the LBHL filter, we obtained image data in the LBHL wavelength band at each time step. The extracted data included five variables: image_t (image intensity), glat_t (geographic latitude), glon_t (geographic longitude), mlat_t (magnetic latitude), and mlon_t (magnetic longitude), all represented as 200×228 matrices. Due to satellite noise, the image_t data may contain negative values, which were reset to zero. Given that each pixel in the image is associated with corresponding magnetic coordinates, the coordinate transformation was applied to project the images onto the magnetic coordinate system. The transformed auroral images are uniformly represented in the magnetic coordinate frame, with magnetic latitudes ranging from 50° to 90° MLAT and magnetic local times spanning from 0 to 24 MLT. The final image resolution is 241 pixel×241 pixel. Fig.3 serves to illustrate a comparison between the aurora image before and after preprocessing.

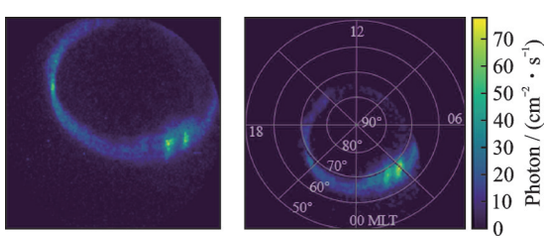


Fig.3 Comparison of ultraviolet aurora images before and after preprocessing

After data preprocessing and cleaning, the dataset used in this study consists of 32 603 auroral images along with the corresponding magnetometer readings from 12 ground stations at the same time points. From this dataset, auroral images and the matching magnetometer data from January 28, 1997, were separated for model validation, while the remaining data were reserved for model training. To address the limited data volume, the separated dataset was further partitioned using the K-fold cross-validation method. Specifically, the dataset was randomly shuffled and divided into eight equal parts, each containing 12.5% of the original data. In each iteration, seven parts were used for training and one part for validation. This process was repeated such that each fold served as the validation set exactly once. Fig.4 illustrates the schematic of the K-fold cross-validation procedure.

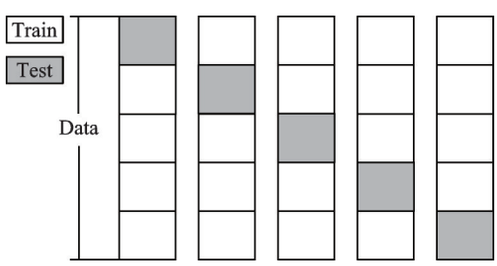


Fig.4 K-fold cross-validation procedure

2 Methods

The imaging range of the aurora imaging data covers the entire polar region, with a specific spatial resolution and a minimum latitude of 50°, resulting in image sequence data denoted as $\boldsymbol{I}\in\mathbf{R}^{m\times n}$ at a given time. Concurrently, the ground magnetic data comprises an array sequence from 12 stations within the polar region, where at a given time, a vector $\boldsymbol{y}=(y_1, y_2, \cdots, y_{12})\in\mathbf{R}^{12}$ represents the geomagnetic data of these stations. Consequently, the prediction task addressed in this paper formulates as a regression prediction problem, entailing the training of a deep learning model with substantial quantities of both image and magnetic data. The objective is to enable the model to predict the corresponding local geomagnetic station component from a given aurora image.

The model proposed in this article comprises three main components, as depicted in Fig.5. They are the visual geometry group (VGG) image feature extraction network, the ViT-based encoder network, and the regression prediction network. Initially, the VGG image feature extraction network employs a serial convolutional neural network to extract local features from aurora images while preserving position information, yielding the deep feature map for time t. Subsequently, the feature map is flattened into a two-dimensional feature matrix, which is then fed into the Transformer encoding module to model large-scale spatial dependencies within the aurora data, resulting in an encoder feature map of the same shape. Finally, a three-layer fully connected network is employed to transform the two-dimensional feature matrix into a one-dimensional feature vector, and the output sequence $\boldsymbol{y}$ is ultimately predicted through the hidden layer and output layer regression.

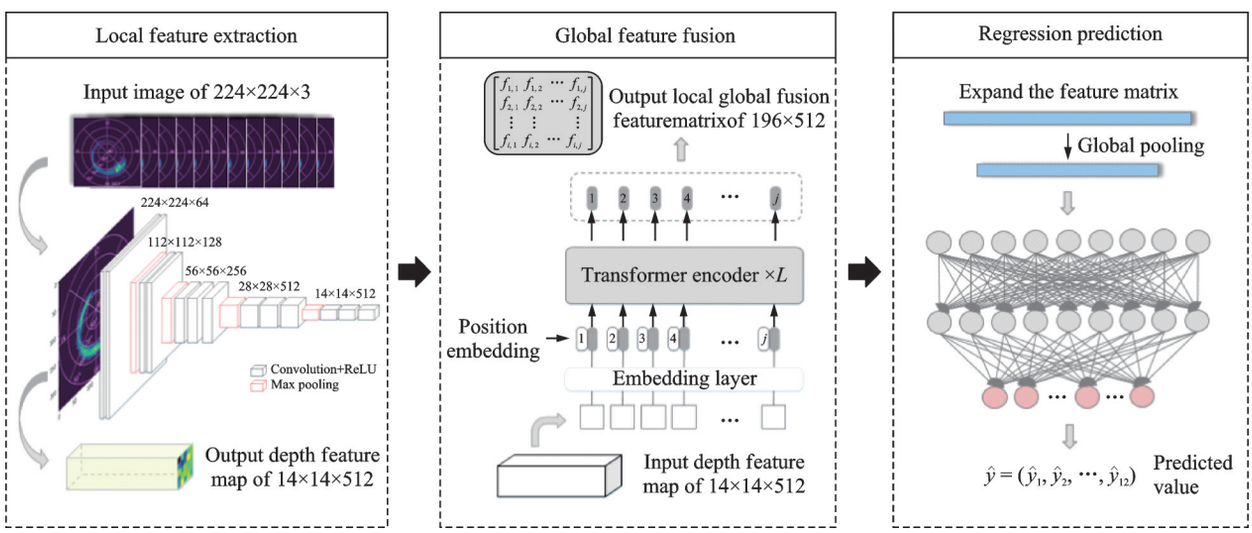


Fig.5 Framework of the model

2.1 Image feature extraction network based on VGG

In order to preserve the positional relationships in the image and extract local features, a serial convolutional neural network structure is employed. VGG, introduced by the University of Oxford in 2014, demonstrated strong performance in the ImageNet Large Scale Visual Recognition Challenge in the same year and has since gained extensive adoption[31]. VGG-16 typically denotes a network architecture incorporating 13 convolutional layers and three fully connected layers. It utilizes a concise and stackable pattern of convolutional blocks, which has proven effective on various datasets. The VGG network represents the maximum depth achievable by traditional serial networks and its significant innovation lies in the widespread use of small-sized convolutional kernels. This involves replacing a 5×5 convolutional kernel with two stacked 3×3 convolutional kernels, and replacing a 7×7 convolutional kernel with three stacked 3×3 convolutional kernels, thereby reducing the network's parameters without compromising the receptive field.

The VGG network consists of five blocks, each of which includes convolutional and pooling layers. Finally, three fully connected layers are linked for classification. In fact, the convolutional layers of the VGG network have good feature extraction capabilities. The network structure of the convolutional and pooling layers of the VGG network is considered as the feature extraction network in this paper. To transform an image of 224×224×3 into a 14×14×512 feature map, the standard VGG network is truncated in this paper to create a new feature extraction network. The network structure is shown in the Fig.5. The output feature dimensions of this network can be transformed into fixed-length embedding vectors through linear map-

ping, aligning with the input paradigm of the ViT model and facilitating subsequent processing.

The feature extraction network in this article retains the convolutional and pooling layers of the VGG network to extract local features of aurora images. After five blocks, the 224×224×3 image is transformed into a 14×14×512 feature map. The network structure is shown in the Fig.6. Compared with the first five blocks of VGG-16, the network in this article removes the last pooling layer, which refers to the data shape when the ViT model enters the encoder, and is convenient for subsequent comparative experiments.

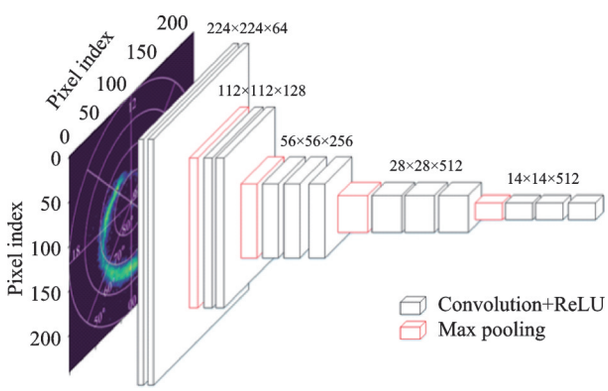


Fig.6 VGG image feature extraction network structure

2.2 Encoder network based on ViT

The Transformer encoder network performs global interaction on the local features extracted by the VGG feature extraction network to further learn the large-scale spatial dependencies between data. Transformer was originally proposed in natural language processing and had subsequently been widely applied to time series, computer vision, and other fields. The Transformer model, in comparison to CNN-based methods, excels at capturing complex spatial transformations and long-distance feature dependencies. It achieves this by effectively learning the relationship between input elements, enabling it to capture global interactions. Additionally, the Transformer model has the ability to flexibly adjust its receptive field to combat interference in the data and learn effective feature representations.

The Transformer model's core is the attention mechanism, which can be described as the process of mapping a query and a set of key-value pairs to an output. Given a query matrix $\boldsymbol{Q}\in\mathbf{R}^{N_q\times d}$ and key-value matrices $\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{N_k\times d}$, with N_q denoting the number of query tokens, N_k the number of key-value tokens, and d the dimensionality of the feature embeddings, the output matrix is calculated by applying the attention function as follows

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \quad (1)$$

Multi-head attention is an extension of attention mechanism that parallelly runs k attention operations by projecting queries, keys, and values into k different subspaces through k learnable linear transformations. Then, the outputs of these k attentions are concatenated and transformed by another learnable linear transformation to obtain the final output

$$\text{MultiHeadAttentions}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(h_1,h_2,\cdots,h_k)\boldsymbol{W}^O \quad (2)$$

$$h_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \quad (3)$$

where $\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K,\boldsymbol{W}_i^V\in\mathbf{R}^{d\times d_h}$ are the parameter matrices for the linear transformations of the query, key, and value, respectively, and $\boldsymbol{W}^O\in\mathbf{R}^{kd_h\times d}$ is the parameter matrix for the final linear transformation of the multi-head attention mechanism. Typically, d_h is set to d/k.

The difficulty of applying Transformer to the field of CV mainly lies in how to convert 2D image data into 1D data. In 2020, Dosovitskiy et al.[25] proposed a visual transformer. This model constructs a series of tokens by segmenting each image into patches with position embeddings, and then extracts parameterized vectors as visual representations using Transformer blocks. Taking ViT-B/16 as an example, a convolutional layer with a convolution kernel size of 16 ×16, a stride of 16, and a convolution kernel number of 768 can be used to achieve this. Through convolution, each 16×16×3 patch maps to a 768-dimensional vector, termed as a token, which is then transformed into a fixed-length embedding vector through linear mapping and fed into a standard Transformer module. When the input image size is 224×224×3, the resulting embedding vector dimension from the embedding layer is 196×768. Additionally, the aurora image feature map extracted by the VGG network can be input as

a token to the Transformer enconder, ensuring that the resulting embedding vector dimension aligns with the input paradigm. Furthermore, the feature matrix $F \in \mathbf{R}^{N \times d}$ undergoes a standard multi-head attention function, and the results of the attention operation are rearranged to obtain two-dimensional encoder features. Notably, for optimal utilization of image features, this study reformulates all the outputs of the encoder as encoder features for subsequent regression prediction, in contrast to classification tasks.

The ViT encoder network, as depicted in Fig.7, comprises an embedding layer and an encoding layer. Initially, the VGG deep feature map is converted into fixed-length embedding vectors by the embedding layer. Subsequently, in order to retain positional information, each patch undergoes the addition of positional encoding information prior to being input into the transformer encoder. The encoding layer then executes multi-head attention in order to process the embedded vectors. Specifically, the Transformer encoding layer is composed of L_x standard Transformer Encoder modules (where x denotes the number of repeated layers in the Transformer encoder), with each module being comprised of the layer normalization (LN), a multi-head self-attention module (MHSA), a multi-layer perceptron (MLP), and residual connections. The MLP further encompasses two convolutional functions along with a rectified linear unit (ReLU) activation function. In Fig.7, Z_l' and Z_l represent the output features of MHSA and MLP in the lth module, and the calculation process is as follows

$$Z_l' = \mathrm{MHSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1} \tag{4}$$

$$Z_l = \mathrm{MLP}(\mathrm{LN}(Z_l')) + Z_l' \tag{5}$$

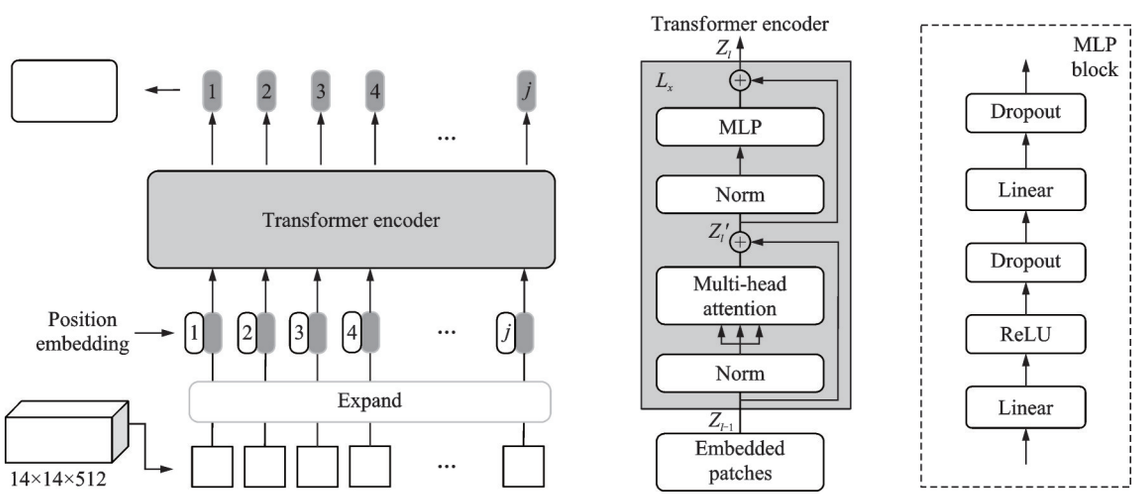


Fig.7 ViT encoder network

2.3 Regression prediction network

The encoded features obtained by the ViT module are expanded and then enter the three-layer fully connected regression prediction network, consisting of two hidden layers and an output layer, with the number of neurons being 1 024, 1 024, and 12, respectively. This network is responsible for generating the geomagnetic monitoring values of 12 stations. To optimize the performance of the network, a standard ADAM optimizer is employed, with a learning rate set to 0.000 2 and a batch size set to 24. When y_i represents the true sequence of geomagnetic station data and $\hat{y}_i$ represents the predicted sequence, the loss function is defined as the mean squared error between the predicted value and the true value of the geomagnetic station data, shown as

$$\mathrm{Loss} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6}$$

2.4 Accuracy evaluation

The model evaluation criteria include root mean square error (RMSE), average relative variance (ARV), and coefficient of determination (R^2). The closer R^2 is to 1, the better the fit of the regres-

sion line to the observed values. The criteria can be defined as

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=0}^{n-1}(y_i-\hat{y}_i)^2}{n}} \tag{7}$$

$$\mathrm{ARV}=\frac{\sum_{i=0}^{n-1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n-1}\left(\frac{\sum_{i=0}^{n-1}y_i}{n}-y_i\right)^2} \tag{8}$$

$$R^2=1-\left|\frac{\sum_{i=0}^{n-1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n-1}y_i}\right| \tag{9}$$

3 Discussion

In order to assess model effectiveness and performance, a comparative analysis of the prediction results from eight models was conducted, including the VGG network, the ViT network, and the hybrid model proposed in this article. The models evaluated encompass ViT-B/16, ViT-L/16, VGG-16, VGG-19, VGG-16+ViT-B/16, VGG-16+ViT-L/16, VGG-19+ViT-B/16, and VGG-19+ViT-L/16. Given the inherently limited size of the auroral image dataset, it was crucial to ensure that each model could achieve effective feature learning despite the scarcity of training samples. However, since all auroral images were projected onto a unified geomagnetic coordinate grid before being fed into the network, each image inherently contained consistent geographic priors. Under such conditions, conventional data augmentation techniques (e.g., rotation, translation, scaling) would disrupt the coordinate consistency and completeness of the data rather than enhance its diversity. Moreover, since auroral images exhibit highly variable noise levels depending on atmospheric and instrumental conditions, introducing additional synthetic noise would not yield meaningful augmentation and could even degrade data fidelity. To address these limitations and improve model generalization, all backbone networks (both VGG and ViT) were initialized with pre-trained weights from the ImageNet dataset. This transfer learning strategy effectively compensated for the restricted data diversity by endowing the models with rich, transferable visual representations, thereby providing a robust foundation for subsequent fine-tuning on auroral imagery. The two network structures under the VGG framework are shown in Table 2. The two model parameters under the ViT framework are shown in Table 3, comprising layer, hidden size, MLP size, and head. Layer denotes the number of times the encoder block is repeatedly stacked in the Transformer; Hidden size signifies the dimension (vector length) of each token after passing through the embedding layer; MLP size corresponds to the number of fully connected nodes in the first MLP block of the Transformer encoder (four times the hidden size); Head represents the number of heads in the multi-head attention of the Transformer. Notably, all three model types leverage a three-layer fully connected regression prediction network, following the feature learning from the aurora images, to achieve the prediction task of geomagnetic data from 12 stations. This article uses the 11th Gen Intel(R) Core(TM) i9-11900KF processor, Nvidia GeForce RTX3080 graphics processor, with a clock speed of 3.5 GHz, 32 GB of memory, and the operating system is Windows10.

Table 2 VGG model structure

VGG-16	VGG-19
13 weight layers	16 weight layers
Input(224×224×3 image)	Input(224×224×3 image)
Conv3-64	Conv3-64
Conv3-64	Conv3-64
Maxpool	Maxpool
Conv3-128	Conv3-128
Conv3-128	Conv3-128
Maxpool	Maxpool
Conv3-256	Conv3-256
Conv3-256	Conv3-256
Conv3-256	Conv3-256
	Conv3-256
Maxpool	Maxpool
Conv3-512	Conv3-512
Conv3-512	Conv3-512
Conv3-512	Conv3-512
	Conv3-512
Maxpool(14×14×512)	Maxpool(14×14×512)
FC-1 024	FC-1 024
FC-1 024	FC-1 024
FC-12	FC-12

Table 3 ViT model parameters

Model	Layer	Hidden size	MLP size	Head
ViT-B/16	12	768	3 072	12
ViT-L/16	24	1 024	4 096	16

Table 4 presents the normalized RMSE values between the predicted and actual values of geomagnetic data at 12 stations across different models. The error curves depicting the performance of these models are provided in Fig.8. The results demonstrate that the model leveraging ViT excels compared to the convolutional network model. The superior performance of the ViT model can be attributed to its adeptness in modeling global features and long-range dependencies within the input data. The pre-trained ViT model enables good performance even for prediction tasks with small data sizes. Conversely, the inherent inductive bias inherent in convolutional networks restricts their capacity to enhance model performance, as they are constrained to focusing solely on local image features. Empirical evidence further indicates that the predictive influence of large-scale global features within aurora images outweighs that of local features in modeling geomagnetic data.

Table 4 RMSE of geomagnetic data from 12 stations in the predicted results of each model

Model	HRN	ABK	LRV	NAQ	GDH	IQA	FCC	BLC	CBB	YKC	CMO	BRW
VGG-16	0.065 4	0.060 4	0.063 6	0.054 1	0.073 1	0.051 7	0.054 7	0.069 1	0.064 8	0.068 9	0.048 8	0.067 7
VGG-19	0.062 2	0.058 5	0.060 9	0.051 2	0.071	0.045 2	0.054 2	0.064 7	0.064	0.066 8	0.047 3	0.064 6
ViT-B/16	0.053 8	0.050 3	0.056 8	0.046 3	0.066 3	0.045 9	0.050 8	0.061 3	0.060 3	0.062 8	0.044 5	0.061 1
ViT-L/16	0.052 7	0.049 8	0.053 1	0.045 2	0.065 4	0.039 8	0.048 4	0.055 6	0.055 9	0.060 2	0.042 9	0.058 4
VGG-16+ ViT-B/16	0.046 6	0.039 7	0.043 0	0.035 0	0.053 1	0.029 9	0.040 8	0.049 1	0.048 3	0.049 7	0.032 3	0.049 6
VGG-16+ ViT-L/16	0.045 0	0.042 6	0.041 7	0.033 6	0.052 0	0.026 7	0.038 9	0.048 5	0.044 9	0.048 6	0.031 2	0.047 9
VGG-19+ ViT-B/16	0.041 9	0.036 5	0.037 6	0.076 5	0.048 5	**0.020 7**	0.032 0	0.043 1	0.040 5	**0.044 2**	**0.028 9**	**0.041 1**
VGG-19+ ViT-L/16	**0.040 3**	**0.035 7**	**0.036 4**	**0.026 8**	**0.046 9**	0.021 9	**0.031 3**	**0.042 4**	**0.039 4**	0.045 1	0.029 6	0.044 6

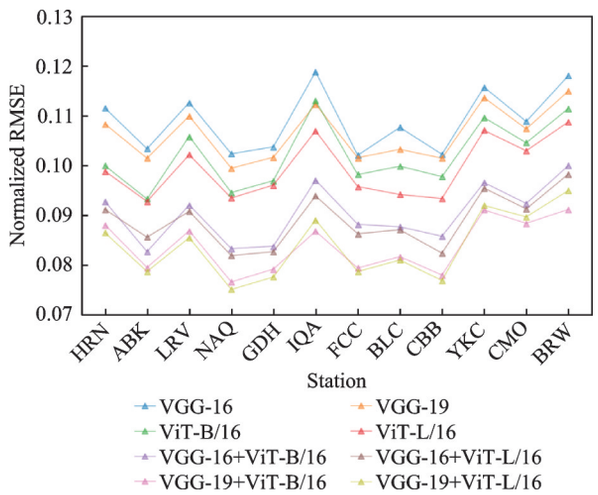


Fig.8 Error curves depicting the performance of each models

Furthermore, compared with single-architecture models, the hybrid model proposed in this study exhibits a stronger ability to capture auroral image characteristics, which results in overall lower prediction errors. This finding suggests that combining convolutional locality with Transformer-based global reasoning provides a more balanced and effective representation, thereby enhancing prediction accuracy.

On 28 January, 1997, the predicted results of the mixed model VGG-19+ViT-L/16 are displayed in Fig.9, where the orange curve depicts the true values and the blue curve represents the model's predictions. The prediction results indicate that the hybrid model's predictive values closely align with the overall trends of data from various magnetic stations, demonstrating good performance even during strong geomagnetic disturbances. It is evident that the model exhibits a smaller prediction error during stationary periods of magnetic field disturbances, while a larger prediction error is observed during substorm expansion periods. Fig.10 depicts the scatter density plot of the mixed model VGG-19+ViT-L/16 for predicting data from 12 stations, with the true values on the horizontal axis and the model's predicted values on the vertical axis. The plot reveals a strong linear relationship between the pre-

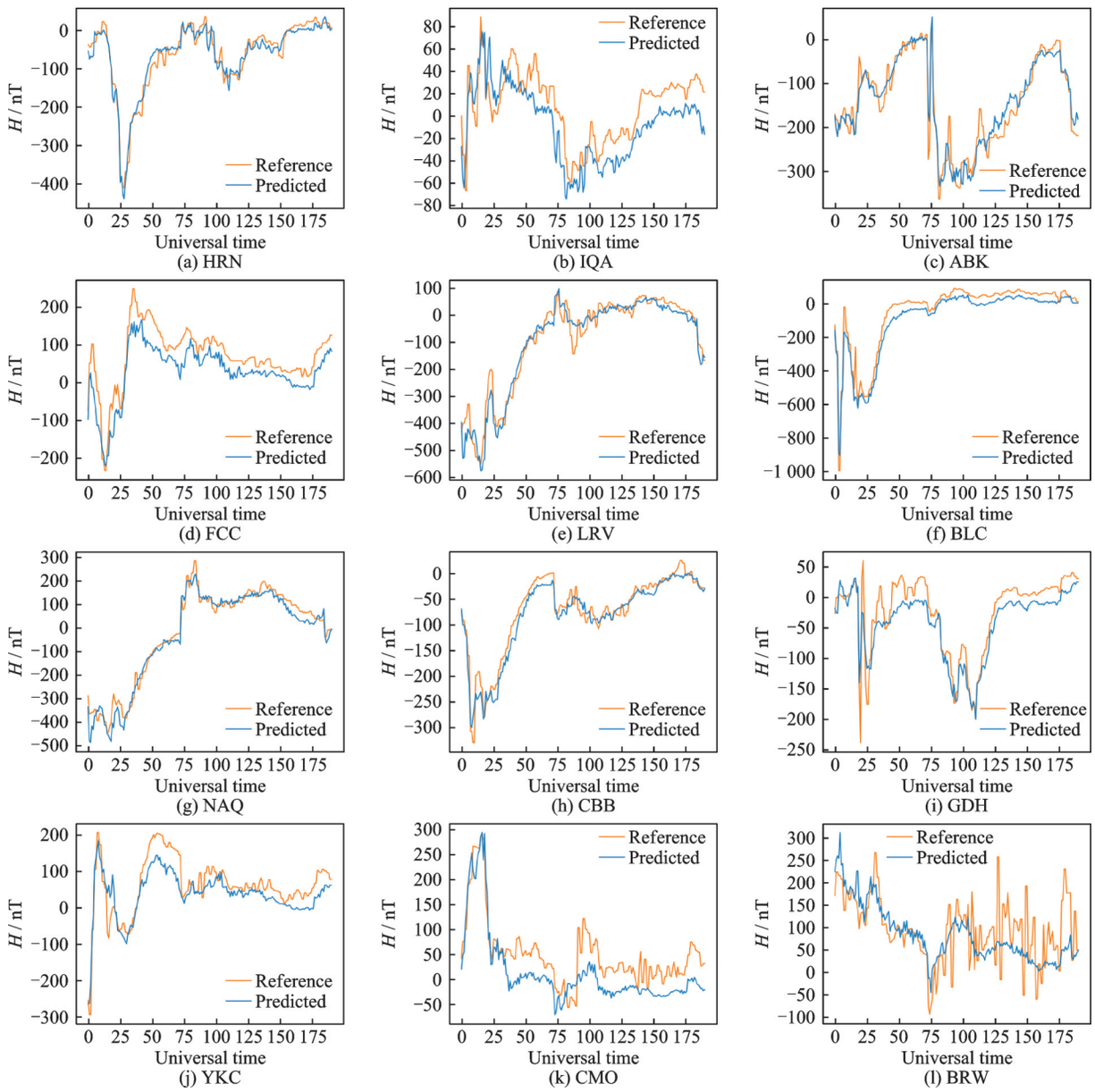


Fig.9 Predicted results of the mixed model VGG-19+ViT-L/16 on 28 January, 1997

dicted results of each station and the true values, with the majority of the predicted results concentrated near the $y=x$ line, indicative of a high correlation between the model's predictions and the true values. Notably, the areas with the highest scattered point density at each station are centered near 0 nT, attributed to the baseline processing of geomagnetic station data to eliminate diurnal variations. Among them, the fitted lines for GDH, NAQ, CMO, and BRW show significant deviations from the $y = x$ line, while FCC, CBB, and YKC exhibit minor offsets. The results for LRV are the most accurate.

The emergence of Transformer in the visual domain has posed a significant challenge to the long-standing dominance of convolutional networks in this area. This is primarily attributed to Transformer's larger receptive field, more flexible weight settings, and stronger global feature modeling capabilities in feature learning, compared to convolutional networks. Consequently, backbone networks based on Transformer hold the potential to deliver higher quality feature inputs for downstream tasks. Notably, ViT stands out as a prominent algorithm that leverages Transformer as a backbone network to encode image features[25]. ViT's global interaction

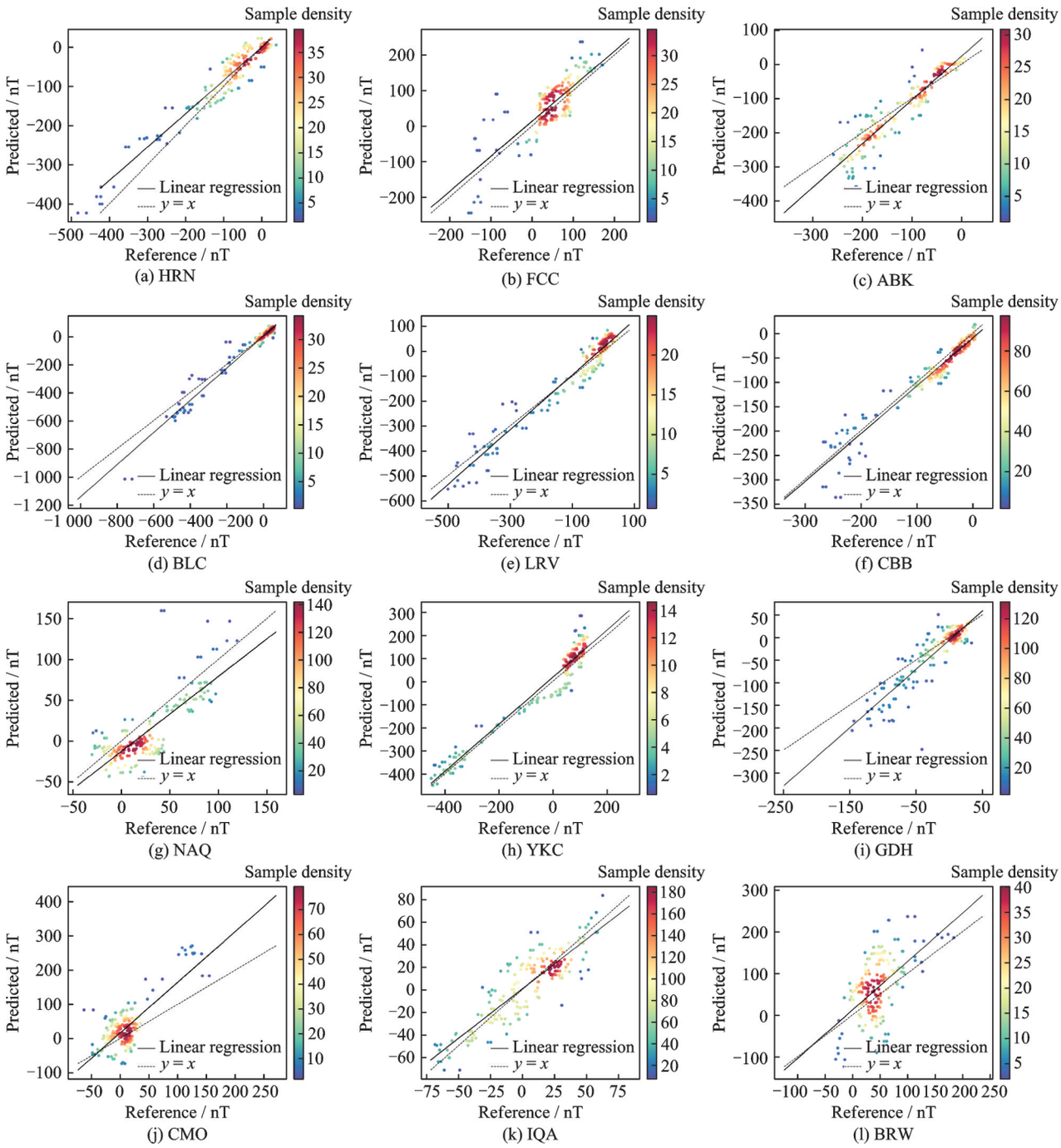


Fig.10 Scatter density plot of the mixed model VGG-19+ViT-L/16 for predicting data from 12 stations

ability with the serialized image input allows for the encoding of image features on a global scale. Moreover, integrating Transformer's global interaction capacity with the spatial locality of convolutional networks presents an opportunity to enhance feature diversity.

In the context of auroral image-driven geomagnetic prediction, VGGs are particularly effective in capturing local spatial textures, brightness gradients, and fine-scale auroral arcs, which are closely related to localized energy precipitation and short-range geomagnetic perturbations. In contrast, the Transformer component excels at modeling large-scale morphological structures and global spatial correlations across the auroral oval, which are essential for representing the overall geomagnetic field response. Therefore, integrating the convolutional network's spatial locality with the Transformer's

global interaction capacity enables the model to learn a more comprehensive and physically consistent representation of auroral patterns.

To further substantiate the advantages of the hybrid VGG-ViT design, an additional comparison was conducted by selecting another convolution-based residual network (ResNet-50 and ResNet-34) as the local feature extractor. Compared with VGG, which extracts features through a strictly hierarchical and spatially smooth convolutional process, ResNet enhances feature extraction by introducing identity skip connections that allow multi-level feature reuse and deeper gradient propagation. This design enables ResNet to capture richer high-level semantic representations[32], while VGG tends to preserve more fine-grained spatial and texture information in its feature maps[31].

As shown in Table 5, the experimental results illustrate the comparative performance of different hybrid structures. When used individually, the ResNet models outperform the VGG models but remain slightly inferior to the ViT models in all three metrics. However, when ResNet is combined with ViT, the hybrid models exhibit a decline in performance across RMSE, ARV, and R^2. In this specific application—Predicting local geomagnetic components from auroral images, this phenomenon can be attributed to the inherent characteristics of the ResNet architecture. While ResNet's residual connections facilitate deeper feature extraction and improve gradient propagation, thereby enhancing standalone performance, they also tend to emphasize high-level semantic abstraction at the expense of spatial precision. When integrated with ViT, which already captures global contextual dependencies, this redundancy in abstract feature representation may reduce the diversity of complementary information between the convolutional and Transformer components. In contrast, the VGG backbone, with its strictly hierarchical and spatially smooth convolutional structure, preserves more fine-grained spatial and texture details. These localized cues are particularly valuable in this application, as geomagnetic disturbances are often reflected in subtle spatial variations of auroral emissions. Such features complement ViT's global feature reasoning more effectively, leading to better overall integration and the superior performance observed in the VGG-ViT hybrid models.

Table 5 Performance of various models in predicting geomagnetic data for 12 stations

Model	RMSE	ARV	R^2
Resnet-34	0.061 9	0.282 4	0.699 3
Resnet-50	0.061 0	0.269 1	0.710 1
VGG-16	0.064 8	0.270 7	0.706 3
VGG-19	0.064 0	0.254 8	0.723 8
ViT-B/16	0.060 3	0.253 2	0.727 8
ViT-L/16	0.055 9	0.270 3	0.707 4
ResNet-34+ViT-B/16	0.066 8	0.293 0	0.672 7
ResNet-34+ViT-L/16	0.069 1	0.279 9	0.700 1
ResNet-50+ViT-B/16	0.065 4	0.280 1	0.699 8
ResNet-50+ViT-L/16	0.071 3	0.315 1	0.613 2
VGG-16+ViT-B/16	0.048 3	0.288 6	0.698 2
VGG-16+ViT-L/16	0.044 9	0.273 6	0.734 5
VGG-19+ViT-B/16	0.040 5	0.261 8	0.739 1
VGG-19+ViT-L/16	**0.039 4**	**0.250 3**	**0.739 3**

Table 5 highlights that the VGG-ViT hybrid architectures achieve the best overall performance among all tested models, with the lowest RMSE, ARV, and the highest R^2 values. Specifically, the hybrid model reduces the RMSE by approximately 39.1% compared to the VGG model, 29.5% compared to the ViT model and 35.3% compared to the ResNet model. Additionally, the goodness of fit of the model is enhanced by approximately 2.14% compared to the VGG model, 1.58% compared to the ViT model, and 4.1% compared to the ResNet model.

In conclusion, this study utilizes aurora satellite images to forecast geomagnetic indices and employs deep learning models to extract features from the images. The research findings indicate that this approach can effectively predict geomagnetic data at specific locations when sufficient data are available. Furthermore, aurora satellite images have the potential to forecast geomagnetic indices at any latitude and longitude within their coverage range. Future research endeavors could involve the utilization of

time series models to unlock the long-term memory characteristics of the data, contributing to advanced predictions spanning 1 h, 2 h, or even longer. Additionally, broader validation on a more extensive dataset would enhance the robustness of the findings. Given that aurora satellite images provide wide spatial coverage, while local geomagnetic station component offer continuous temporal observations, they complement each other. In the future, the joint index of aurora satellite images and geomagnetic station data may be able to leverage the advantages of both and compensate for the limitations of both. This index could potentially resolve the index saturation problem arising from the non-uniform distribution of geomagnetic stations, thereby enhancing the accuracy of predicting the timing and location of substorm events.

4 Conclusions

Based on the local and global feature correlations between aurora images and geomagnetic variations, this paper proposes a ViT-based hybrid framework for predicting local geomagnetic station components from auroral observations. The experimental evaluation is conducted using a combined dataset consisting of POLAR satellite LBHL-band aurora images and geomagnetic monitoring data collected from 12 stations located in the high- and mid-latitude regions of the Arctic. By integrating convolutional neural networks for local feature extraction with a Transformer encoder for global feature modeling, the proposed model is able to capture both fine-scale spatial details and large-scale auroral morphological structures in a unified manner. The experimental results indicate that geomagnetic prediction benefits more substantially from the global morphological characteristics of aurora images than from local features alone, highlighting the importance of large-scale spatial context in aurora-geomagnetic coupling. Moreover, the complementary integration of convolutional inductive bias and transformer-based global attention enables the model to effectively integrate local feature sensitivity with global auroral morphology representation, resulting in more robust feature representations. Overall, the proposed hybrid architecture demonstrates consistent advantages over single-model approaches, underscoring its effectiveness for geomagnetic parameter estimation based on auroral imaging data. These findings suggest that hybrid CNN-Transformer frameworks offer a promising and extensible paradigm for data-driven space weather analysis and related geospace monitoring tasks.

The main contributions of this study can be summarized as follows:

(1) This work is the first to propose a model that predicts local geomagnetic station components directly from auroral images, establishing a reverse mapping from optical auroral observations to geomagnetic responses. This approach complements existing research that primarily focuses on predicting auroral intensity from geomagnetic data.

(2) The proposed framework overcomes the geographical constraints of ground-based geomagnetic stations, enabling geomagnetic variation prediction even in regions without dense observational networks.

(3) A hybrid ViT-based architecture is introduced, combining the fine-grained local feature extraction capability of convolutional networks with the global contextual modeling power of the Transformer, resulting in improved prediction accuracy and generalization performance.

Although the auroral image dataset spans full year, it is important to note that not all images are equally suitable for modeling. In particular, the presence of solar glare (dayglow) during periods of sunlight exposure introduces significant interference in the ultraviolet auroral observations, thereby affecting data quality and reducing model accuracy. This limitation restricts the effective use of data to nighttime auroral observations, particularly in the polar night region. To further improve the robustness and generalization of the model, future work will focus on implementing additional preprocessing steps to identify and remove solar contamination. Techniques such as radiometric correction, dayglow filtering, or machine-learning-based noise detection

will be explored to mitigate these effects. By enhancing data quality, we aim to extend the temporal availability of usable auroral images and improve the consistency of geomagnetic prediction across seasons.

References

[1] AKASOFU S I, LUI A T Y, MENG C I. Importance of auroral features in the search for substorm onset processes[J]. Journal of Geophysical Research: Space Physics, 2010, 115(A8): 2009JA014960.

[2] SINGH A K, SIINGH D, SINGH R P. Space weather: Physics, effects and predictability[J]. Surveys in Geophysics, 2010, 31(6): 581-638.

[3] MANDEA M, CHAMBODUT A. Geomagnetic field processes and their implications for space weather[J]. Surveys in Geophysics, 2020, 41(6): 1611-1627.

[4] TSUJI Y, SHINBORI A, KIKUCHI T, et al. Magnetic latitude and local time distributions of ionospheric currents during a geomagnetic storm[J]. Journal of Geophysical Research: Space Physics, 2012, 117(A7): 2012JA017566.

[5] MYAGKOVA I N, SHIROKII V R, VLADIMIROV R D, et al. Prediction of the DST geomagnetic index using adaptive methods[J]. Russian Meteorology and Hydrology, 2021, 46(3): 157-162.

[6] SINGH P K. Prediction of intensity of moderate and intense geomagnetic storms using artificial neural network during two complete solar cycles 23 and 24[J]. Indian Journal of Physics, 2022, 96(8): 2235-2242.

[7] NEWELL P T, GJERLOEV J W. Evaluation of SuperMAG auroral electrojet indices as indicators of substorms and auroral power[J]. Journal of Geophysical Research: Space Physics, 2011, 116(A12): 2011JA016779.

[8] NEWELL P T, GJERLOEV J W. Substorm and magnetosphere characteristic scales inferred from the SuperMAG auroral electrojet indices[J]. Journal of Geophysical Research: Space Physics, 2011, 116(A12): 2011JA016936.

[9] SINGH A K, RAWAT R, PATHAN B M. On the UT and seasonal variations of the standard and supermag auroral electrojet indices[J]. Journal of Geophysical Research: Space Physics, 2013, 118(8): 5059-5067.

[10] NEWELL P T, GJERLOEV J W. Local geomagnetic indices and the prediction of auroral power[J]. Journal of Geophysical Research: Space Physics, 2014, 119(12): 9790-9803.

[11] WATERS C L, GJERLOEV J W, DUPONT M, et al. Global maps of ground magnetometer data[J]. Journal of Geophysical Research: Space Physics, 2015, 120(11): 9651-9660.

[12] MANSHOUR P, BALASIS G, CONSOLINI G, et al. Causality and information transfer between the solar wind and the magnetosphere-ionosphere system[J]. Entropy, 2021, 23(4): 390.

[13] ALFONSI L, BERGEOT N, CILLIERS P J, et al. Review of environmental monitoring by means of radio waves in the polar regions: From atmosphere to geospace[J]. Surveys in Geophysics, 2022, 43(6): 1609-1698.

[14] HU Z J, YANG Q J, LIANG J M, et al. Variation and modeling of ultraviolet auroral oval boundaries associated with interplanetary and geomagnetic parameters[J]. Space Weather, 2017, 15(4): 606-622.

[15] YANG Qiujun, HU Zejun, HAN Desheng, et al. Modeling and prediction of ultraviolet auroral oval boundaries based on IMF/solar wind and geomagnetic parameters[J]. Chinese Journal of Geophysics, 2016, 59(2): 426-439. (in Chinese)

[16] MENG C I, LIOU K. Global auroral power as an index for geospace disturbances[J]. Geophysical Research Letters, 2002, 29(12): 41-1-41-4.

[17] LIOU K, CARBARY J F, NEWELL P T, et al. Correlation of auroral power with the polar cap index[J]. Journal of Geophysical Research: Space Physics, 2003, 108(A3): 2002JA009556.

[18] LIU Xiaocan, CHEN Gengxiong, XU Wenyao, et al. Relationships of the auroral precipitating particle power with AE and DST indices[J]. Chinese Journal of Geophysics, 2008, 51(4): 968-975. (in Chinese)

[19] MITCHELL E J, NEWELL P T, GJERLOEV J W, et al. OVATION-SM: A model of auroral precipitation based on SuperMAG generalized auroral electrojet and substorm onset times[J]. Journal of Geophysical Research: Space Physics, 2013, 118(6): 3747-3759.

[20] HU Z J, HAN B, ZHANG Y, et al. Modeling of ultraviolet aurora intensity associated with interplanetary and geomagnetic parameters based on neural networks[J]. Space Weather, 2021, 19(11): e2021SW002751.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems. Long

Beach, USA: [s.n.], 2017: 5999-6009.

[22] XIANG Deping, ZHANG Pu, XIANG Shiming, et al. Multi-modal meteorological forecasting based on Transformer[J]. Computer Engineering and Applications, 2023, 59(10): 94-103. (in Chinese)

[23] LI J, DU J Q, ZHU Y C, et al. Survey of Transformer-based object detection algorithms[J]. Computer Engineering and Applications, 2023, 59(10): 48-64.

[24] SHI Lei, JI Qingyu, CHEN Qingwei, et al. Review of research on application of vision Transformer in medical image analysis[J]. Computer Engineering and Applications, 2023, 59(8): 41-55. (in Chinese)

[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. (2020-10-05). https://arxiv.org/abs/2010.11929.

[26] HO J, KALCHBRENNER N, WEISSENBORN D, et al. Axial attention in multidimensional Transformers[EB/OL]. (2019-12-12). https://arxiv. org/abs/1912.12180.

[27] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[EB/OL]. (2021-02-08). https://arxiv.org/abs/2102.05095.

[28] ZHU Daiyin, LV Jiming, ZHOU Peng, et al. Detection and recognition method of small UAV SAR ground targets[J]. Journal of Nanjing University of Aeronautics & Astronautics (Natural Science Edition), 2025, 57(5): 781-798. (in Chinese)

[29] PENG Z, HUANG W, GU S, et al. Conformer: Local features coupling global representations for visual recognition[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2022: 357-366.

[30] GUO J, HAN K, WU H, et al. CMT: Convolutional neural networks meet vision Transformers[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 12165-12175.

[31] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-10). https://arxiv. org/abs/1409.1556.

[32] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.41471381); the General Project of Jiangsu Natural Science Foundation (No.BK20171410); and the Major Scientific and Technological Achievements Cultivation Fund of Nanjing University of Aeronautics and Astronautics(No.1011-XBD23002).

Author

The first/corresponding author Dr. **WANG Bo** received the B.S. degree in remote sensing science and technology, the M.S. degree in surveying engineering, and the Ph.D. degree in remote sensing for photogrammetry from Wuhan University, Wuhan, China, in 2010, 2012, and 2015, respectively. From 2015 to 2021, he was a lecturer at Nanjing University of Aeronautics and Astronautics (NUAA). Since 2021, he has been an associate professor with College of Astronautics, NUAA, where he also serves as the Director of the Department. His research interests include satellite remote sensing, photogrammetry, and computer vision, with a particular focus on satellite image measurement, SAR and optical data fusion, and image matching error elimination methods.

Author contributions Dr. **WANG Bo** designed the study, compiled the models, conducted the analysis, interpreted the results, and wrote the manuscript. Prof. **SHENG Qinghong** supervised the study, guided the methodology, and revised the manuscript. Dr. **LI Jun** contributed to data processing, conducted experiments, and assisted with analysis. Dr. **LING Xiao** contributed to data integration, model implementation, and manuscript preparation. Dr. **LIU Xiang** compiled model components, conducted validation, and assisted with interpretation. Dr. **CHENG Wei** contributed to the analysis, provided technical support, and assisted with data interpretation. Mr. **ZHANG Yuanshu** contributed to data collection, performed preliminary analysis, and assisted in writing. Ms. **TIAN Xinqin** contributed to image processing, conducted supporting experiments, and assisted in manuscript editing. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

基于局部-全局特征的极光图像中局部地磁场分量建模

王 博[1], 张元舒[1], 成 巍[2], 田馨沁[1], 盛庆红[1],
李 俊[1], 凌 霄[1], 刘 祥[1]

(1.南京航空航天大学航天学院, 南京 211106, 中国; 2.北京应用气象研究所,北京 100029, 中国)

摘要:准确预测地磁场对于全球范围内的空间环境监测和空间天气预报具有重要意义。本文提出了一种利用极光图像来预测当地地磁站分量的ViT(Vision Transformer)混合模型,打破了地磁站的空间局限性。本文方法将ViT骨干模型与卷积网络相结合,以捕捉极光图像与地磁站数据之间的大规模空间相关性和明显的局部特征相关性。本质上,该模型由1个VGG (Visual geometry group)图像特征提取网络、1个基于ViT的编码器网络和1个回归预测网络组成。本文实验结果表明,极光图像的全局特征在预测地磁数据方面比局部特征发挥的作用更为显著。具体而言,该混合模型与VGG模型相比,均方根误差降低了39.1%,与ViT模型相比降低了29.5%,与ResNet (Residual network)模型相比降低了35.3%。此外,该模型的拟合精度分别比VGG、ViT和ResNet模型高出2.14%、1.58%和4.1%。

关键词:紫外极光图像;地磁场预测;ViT混合模型