*Proceedings of the Thirty-fourth (2024) International Ocean and Polar Engineering Conference* **www.isope.org**
*Rhodes, Greece, June 16–21, 2024*

*ISBN 978-1-880653-78-4; ISSN 1098-6189*



# A Real-time Inversion Framework for Carbon Equivalent Emissions in Oil and Gas Extraction based on Vision Transformer

*Zehua Song[a], Xiaoyang Yu[c], Yu Song[a], Jin Yang[b], Dongsheng Xu[b], Kejin Chen[b], Fangfei Huang[b], Bin Chen[b], Yanwei Song[d]*
[a]College of Information Science and Engineering / College of Artificial Intelligence, China University of Petroleum (Beijing)
Beijing, China
[b]College of Safety and Ocean Engineering, China University of Petroleum (Beijing)
Beijing, China
[c]ByteDance
Culver, CA, USA
[d]Haikou Ecological Environment Agency
Haikou, Hainan, China

## ABSTRACT

This study developed a framework based on Vision Transformer UNet (ViT-UNet) aimed at accurately inverting carbon equivalent emissions caused by oil and gas extraction. The research spanned the collection and preprocessing of high-resolution remote sensing imagery from specific areas in the Gulf of Mexico from 2021 to 2023, including spatial resampling, cropping, and extraction of band reflectance values, leading to the precise calculation and spatial analysis of carbon equivalent emissions. The developed ViT-UNet model leverages the feature extraction capabilities of Vision Transformer (ViT) and the spatial information reconstruction advantages of U-Net. It optimized the training process through adaptive learning rate adjustments and data augmentation techniques, achieving high-precision pixel-level carbon equivalent emission predictions ($R^2$=0.8042). This comprehensive framework provides a scientific basis for monitoring and managing carbon emissions from oil and gas fields, demonstrating the potential of deep learning in environmental monitoring.

KEY WORDS: Carbon Equivalent Emissions; Vision Transformer UNet; Remote Sensing Imagery; Emission Concentration Calibration; Spatial Analysis

## INTRODUCTION

The deepening progression of global industrialization has led to a significant increase in the concentration of greenhouse gases in the atmosphere due to the large-scale extraction and consumption of fossil fuels, thereby intensifying the trend of global warming. According to data from the International Energy Agency (IEA), the combustion and extraction activities of fossil fuels are the single largest source of global carbon dioxide emissions, accounting for over 60% of the total emissions (Liu et al., 2023). In this context, the oil and gas industry, as a crucial component of the global energy structure, has drawn widespread attention for its role in the emissions of carbon dioxide ($CO_2$), methane ($CH_4$), and carbon monoxide (CO). These gases not only affect the chemical properties of the atmosphere but are also among the primary drivers of global climate change (Lee et al., 2023). Therefore, accurately monitoring greenhouse gas emissions during the extraction process of oil and gas to estimate carbon equivalent emissions is vital for a deep understanding of the global carbon cycle and for formulating effective environmental policies.

Traditional methods for monitoring and evaluating greenhouse gas emissions on a global scale, such as ground observation stations and laboratory sample analyses, offer advantages in local accuracy. However, they face limitations in global representation and timeliness. Ground stations can provide precise $CO_2$ and $CH_4$ concentration data but are limited by geographic coverage, making it challenging to immediately reflect global carbon emission dynamics (Juselius, 2023). Similarly, laboratory analyses face challenges in timeliness and comprehensiveness when determining concentrations of gases like CO (Burgués et al., 2023). In contrast, the development of satellite remote sensing technology has offered a new perspective for global climate monitoring. Although advanced monitoring systems like Sentinel-5P can monitor various greenhouse gases, they still have limitations in monitoring key gases such as $CO_2$ (Zhao et al., 2023). These limitations directly impact our comprehensive understanding of greenhouse gas emission dynamics, especially in accurately estimating carbon equivalent emissions. Given these restrictions in satellite remote sensing technology, recent research efforts have made significant progress, expanding the scope of greenhouse gas monitoring and estimation. For example, Shi and others (Shi et al., 2023) have successfully estimated point-source $CO_2$ emissions using Differential Absorption Lidar (DIAL) technology, providing an important method for precisely monitoring local emissions. Additionally, Wu and others (Wu et al., 2023) estimated the global and China's land carbon flux for 2019 using the GEOS-Chem model. Duman and others (Duman et al., 2023) studied the spatiotemporal evolution of energy consumption carbon emissions in China from 2010 to 2020 using multi-source remote sensing data. In terms of methane emissions, Erkkilä and others (Erkkilä et al., 2023) observed methane emissions from northern lakes using remote sensing technology, while Liang and

others (Liang et al., 2023) conducted a high-resolution inversion of China's methane emissions using TROPOMI satellite observation data. Su and others (Su et al., 2023) improved the simulation and source attribution of carbon monoxide in the GEOS-Chem model. These studies highlight the potential of remote sensing technology in monitoring and estimating key greenhouse gas emissions, providing important scientific evidence for understanding the global carbon cycle and climate change. However, existing technologies still need improvement in comprehensively monitoring various emission components and effectively integrating different data sources to estimate carbon equivalent emissions. This is crucial for a comprehensive and accurate understanding of the global carbon cycle and its impact on climate change.

Against this backdrop, the study introduces an inversion framework based on the ViT-UNet deep learning model. This framework aims to estimate carbon equivalent emissions from oil and gas extraction activities in the South Marsh Island Area and Eugene Island Area of the Gulf of Mexico. It integrates high-resolution remote sensing images collected every 10 days from March 6, 2021, to February 24, 2023, with emission component data ($CH_4$, $CO_2$, CO) from 2020 to 2022. This method transcends the limitations of traditional monitoring techniques, offering a comprehensive analysis of various emission components. This provides a new perspective for accurately estimating carbon equivalent emissions. Leveraging the advanced processing capabilities of the ViT-UNet model, significant progress was achieved in enhancing the precision and efficiency of carbon emission monitoring in oil and gas extraction areas. This holds important scientific significance for a deeper understanding of the global carbon cycle and its impact on climate change.

## WORKFLOW

**Step 1 Data Acquisition and Processing:** The research initially focused on acquiring high-resolution remote sensing images of the South Marsh Island Area and Eugene Island Area in the Gulf of Mexico. These images were collected every 10 days starting from March 6, 2021, until February 24, 2023. Concurrently, data on key emission components such as $CH_4$, $CO_2$, and CO were gathered for the same period. In the preprocessing phase, the original remote sensing data underwent a spatial resampling to 20 meters to optimize spatial resolution for specific analysis needs. Moreover, by cropping specific areas within the images, data volume was effectively reduced, enhancing the efficiency of subsequent processing. Finally, band reflectance values were extracted from the remote sensing images and associated with latitude and longitude coordinates, providing precise input data for the deep learning model.

**Step 2 Calibration of Emission Concentrations and Conversion to Carbon Equivalent:** This phase focused on developing MethaNet, CarboNet, and MonoNet, deep learning models for accurately predicting $CH_4$, $CO_2$, and CO concentrations at the pixel level. Utilizing the Multilayer Perceptron (MLP) architecture, these models processed complex features from remote sensing imagery to estimate emission levels. Training involved supervised learning, optimization with AdamW, and early stopping to enhance accuracy and generalizability. Following prediction, emission data were converted to carbon equivalent emissions using physicochemical calculations and Global Warming Potential (GWP) values. Spatial analysis techniques then transformed this data into detailed geographic emission maps for the South Marsh Island Area and Eugene Island Area, facilitating accurate carbon emission visualization and analysis.

**Step 3 Design and Deployment of ViT-UNet Model:** This stage involved the original design of the ViT-UNet model, which uniquely combines the ViT's feature extraction capabilities with U-Net's spatial information reconstruction advantages for precise pixel-level carbon equivalent emission predictions. The process included transforming multispectral remote sensing imagery into a compatible three-channel format through 1x1 convolutions, preparing the data specifically for ViT-UNet. With a custom-configured ViT-B/16 encoder and an innovative decoder, the model adeptly managed detailed image analysis. Adaptive learning rate strategies, data augmentation techniques, and early stopping measures were implemented to refine performance and ensure model robustness. The Mean Squared Error (MSE) metric evaluated the accuracy of predictions, facilitating the model's ability to accurately delineate carbon equivalent emissions across specified geographic locales based on remote sensing data.

**Step 4 Model Evaluation and Hotspot Identification:** The ViT-UNet model underwent rigorous evaluation to validate its precision in predicting carbon equivalent emissions at the pixel level. Utilizing both qualitative and quantitative methods, the model's accuracy and utility were thoroughly assessed. Qualitatively, visual comparisons between model predictions and actual ground measurements showcased the model's ability to accurately map emission distributions. Quantitative analysis involved a remapping process and $R^2$ scores, enabling the conversion of image colors into quantifiable carbon emission data. This dual approach facilitated the identification of emission hotspots near oil and gas platforms and allowed for pixel-level precision in carbon emission quantification, enhancing the model's optimization for practical applications.

## EMISSION COMPONENT CONCENTRATION BASELINE

### Data Preparation

#### *Spectral Band Configuration*

The Sentinel-2B series of remote sensing images, part of the European Space Agency's (ESA) Copernicus program, aims to offer continuous monitoring of the Earth's surface. The Multi-Spectral Instrument (MSI) on the Sentinel-2B satellite provides 13 spectral bands covering from visible to near-infrared regions, with resolutions ranging from 10 to 60 meters. Table 1 displays the 13 spectral bands along with their resolution, central wavelength, and description. These capabilities enable the capture of subtle changes on the Earth's surface (Elstohy et al., 2023). Such data are vital for understanding and monitoring the interactions between the Earth's surface and the atmosphere. After radiometric calibration and atmospheric correction, Sentinel-2B imagery is converted to Level 2A data, offering detailed information about surface reflectance (Gorroño et al., 2023). This information is particularly important for analyzing the spatial distribution and temporal variations of emission component concentrations. Due to its high data quality and accuracy, Level 2A data form the basis of this study's analysis.

**Table 1** Sentinel-2B MSI Spectral Bands and Characteristics

| Band | Resolution | Central Wavelength | Description |
|---|---|---|---|
| B1 | 60m | 443nm | Coastal/Aerosol (ultra-blue) |
| B2 | 10m | 490nm | Blue |
| B3 | 10m | 560nm | Green |
| B4 | 10m | 665nm | Red |
| B5 | 20m | 705nm | Vegetation Red Edge (VNIR) |
| B6 | 20m | 740nm | Vegetation Red Edge (VNIR) |
| B7 | 20m | 783nm | Vegetation Red Edge (VNIR) |
| B8 | 10m | 842nm | VNIR |
| B8a | 20m | 865nm | VNIR |

| | | | |
|---|---|---|---|
| B9 | 60m | 940nm | Water Vapor (SWIR) |
| B10 | 60m | 1375nm | SWIR |
| B11 | 20m | 1610nm | SWIR |
| B12 | 20m | 2190nm | SWIR |

### *Emission Component Data Collection*

The measurement data for emission components were obtained from flights by the Mooney aircraft over the Gulf of Mexico and the Permian and Eagle Ford regions of Texas, USA. The Mooney aircraft is equipped with a series of high-precision sensors capable of measuring key components in the atmosphere, such as $CH_4$, $CO_2$, CO, and water vapor ($H_2O$). The measurement of $CH_4$ and $CO_2$ uses the Picarro 2301-f instrument, known for its high sensitivity and precise results based on wavelength-scanned cavity ring-down spectroscopy (Crosson, 2008). The performance metrics of the Picarro 2301-f instrument are detailed in Table 2 Despite the presence of measurement noise, the sensitivity of the instrument at the ppb level allows for the approximate neglection of such noise in the context of this study.

**Table 2** Picarro 2301-f Instrument Performance Metrics

| Parameter | $CO_2$ | $CH_4$ | CO |
|---|---|---|---|
| Precision | ≤150/50 ppb | ≤1/0.7 ppb | ≤0.6/0.3 ppb |
| Max Drift at STP (over 24 hours/1 month) | ≤150/500 ppb | ≤1/3 ppb | ≤0.6/2 ppb |
| Measurement Frequency | ≥0.2 Hz | ≥0.2 Hz | ≥0.2 Hz |
| Gas Response in Measurement Cell | ≥0.33 Hz | ≥0.33 Hz | ≥0.33 Hz |
| Guaranteed Specification Range | 300-500 ppm | 1-9 ppm | 0.5-1.5 ppm |
| Operating Range | 0-1000 ppm | 0-20 ppm | 0-10 ppm |

Ozone ($O_3$) measurements are conducted with the 2B 205 dual-beam ozone monitor, which can accurately measure ozone concentrations in both ozonized and non-ozonized air (Hjellbrekke et al., 2022). Nitric oxide (NO) and nitrogen dioxide ($NO_2$) measurements are performed using the ECOphysics 88 $NO_c$ and Teledyne T5000U instruments, respectively. The aircraft's GPS antenna records the precise location of the collected data, as well as flight altitude, heading, and speed parameters. These details are crucial for spatial positioning and time stamping of the data, ensuring accuracy and reliability in subsequent data processing. Fig. 1 shows the aircraft monitoring points for emission component data.

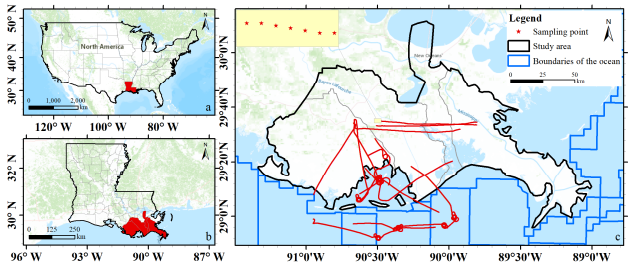


**Fig. 1** Aircraft Monitoring Points for Emission Component Data

### *Data Selection and Temporal-Spatial Synchronization*

For the baseline calibration and study of emission component concentrations, specific image datasets were carefully selected. To establish the emission component concentration baseline, the focus was on remote sensing images from five key dates in 2022: April 17, April 25, November 1, November 6, and November 8. These images contained key sampling point data for emission component concentrations, crucial for establishing the emission component concentration model. These data points provided a series of high-quality reference values for subsequent model calibration and verification. Additionally, for a comprehensive temporal-spatial analysis, 292 Sentinel-2B remote sensing images from March 6, 2021, to February 24, 2023, were selected. These images covered the South Marsh Island Area and Eugene Island Area in the Gulf of Mexico, providing a broad geographic range and time series data. The selection of this data aimed to support and validate the emission component concentration baseline established through the images from the aforementioned five key dates. To ensure effective synchronization between emission component data and remote sensing image data, strict time stamping and geographic coordinate matching were employed. By precisely aligning the emission component data from ground monitoring sites with the corresponding remote sensing images, consistency in time and space was ensured. This step was crucial for subsequent data processing and model training, as accuracy in reflecting the complexities of the real world relies on precisely aligned data (Campbell et al., 2011).

Through the rigorous data preparation described above, a solid data foundation was provided for the calibration of the emission component concentration baseline. Selecting high-quality remote sensing image data sources, precise emission component measurement data, and ensuring temporal and spatial synchronization collectively ensured the accuracy and effectiveness of subsequent analyses and model training.

## Data Processing

### *Sentinel-2B Remote Sensing Image Data Preprocessing*

Resampling: Given the spatial resolution differences among bands in the original Sentinel-2B images, a bicubic interpolation method was applied for resampling all bands to a uniform spatial resolution of 20 meters. Bicubic interpolation, a technique widely used in image processing, estimates new pixel values by considering 16 pixels (4x4) in the pixel neighborhood. Compared to nearest neighbor or bilinear interpolation methods, bicubic interpolation better maintains image smoothness and detail (Castleman, 1996). This process aimed to optimize spatial consistency of the dataset, facilitating complex large-scale environmental analysis. The choice of a 20m resolution balanced the need for high resolution with the extent of the study area.

Band Fusion: During the band fusion stage, bands b1 to b12 underwent fine spectral fusion to capture comprehensive surface and atmospheric property information. This fusion leveraged the unique spectral characteristics of each band, providing multidimensional information for subsequent atmospheric concentration baseline calibration (Lennon, 2002).

Stitching: For image stitching, given that each set of required images actually included four simultaneous-point images, the Scale-Invariant Feature Transform (SIFT) algorithm was used to identify and match common feature points between images (Lowe, 2004). The SIFT algorithm can identify unique feature points in different images and accurately stitch images by comparing the positions and orientations of these feature points. This method ensured complete and continuous coverage of the South Marsh Island Area and Eugene Island Area. It not only preserved the details of each image but also guaranteed data consistency across a large area.

Cropping: Using shapefiles (shp) defined on the Official Plats of Survey (OPD) and Lease Maps (LM) provided by geospatial services, stitched images were precisely cropped to define the specific study area. This step, employing GIS technology, enhanced the accurate delineation of the study area, improving the geographic precision of subsequent analyses (Longley, 2005). The preprocessing workflow for Sentinel-2B remote sensing image data in the South Marsh Island Area and Eugene

Island Area is shown in Fig. 2.

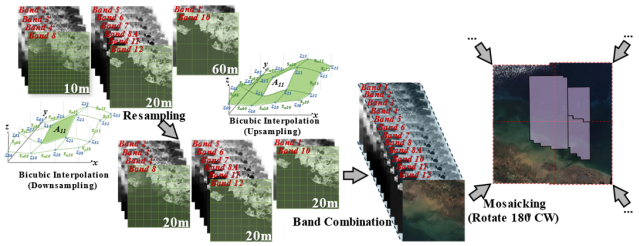


**Fig. 2** Preprocessing Workflow for Remote Sensing Image Data in the South Marsh Island Area and Eugene Island Area

### *Coordinate Extraction and Data Synchronization*

For the images from five key dates in 2022, latitude and longitude coordinates containing emission component concentration data sampling points and their band reflectance values were precisely extracted. This ensured high consistency between surface characteristics derived from remote sensing images and emission component data from ground monitoring sites. For the 73 processed images between March 6, 2021, and February 24, 2023, a systematic spatial data sampling was conducted using quantitative grid analysis techniques. This method combined principles of grid data processing and spatial analysis, creating virtual sampling points at fixed intervals of every 220 meters. It ensured efficient and precise collection of spatial data, such as latitude and longitude points and corresponding band reflectance values, across the extensive areas of the South Marsh Island Area and Eugene Island Area.

### *Correction of Emission Component Data*

In terms of emission component data correction, a multi-stage data cleaning and quality control strategy was implemented. The initial phase involved a comprehensive review of the raw data to eliminate significant anomalies or errors. Subsequently, statistical analysis methods based on standard deviation and skewness were utilized to identify and remove outliers from the dataset, reducing the impact of data noise on the analysis results (Iglewicz et al., 1993). The implementation of these steps ensured the high quality of the dataset and the reliability of the analysis.

## Emission Component Concentration Calibration

A key component was the development of fully connected network models capable of accurately predicting emission component concentrations. For this purpose, three specialized models named MethaNet, CarboNet, and MonoNet were developed to predict concentrations of $CH_4$, $CO_2$, and CO, respectively. Based on the MLP architecture, these models aimed to efficiently process complex input data and accurately predict the corresponding emission component concentrations.

Each model comprised an input layer, several hidden layers, and an output layer. The input layer received latitude and longitude coordinates and reflectance values of 12 bands extracted from remote sensing images. This data underwent thorough preprocessing, including normalization, to ensure consistency in model training and prediction. Features used in the normalization process were saved for applying the same treatment to new data during the prediction phase. The number of hidden layers was set to three to extract deep features from geospatial data. Each layer employed the ReLU activation function to enhance the model's capability to process nonlinear relationships (Goodfellow et al., 2016). This architecture aimed to identify and learn complex patterns in the input data, crucial for understanding and predicting emission component concentrations. The output layer was responsible for generating predictions of emission component concentrations at each location, converting advanced features extracted by hidden layers into specific concentration values, directly supporting the calibration of emission component concentration baselines.

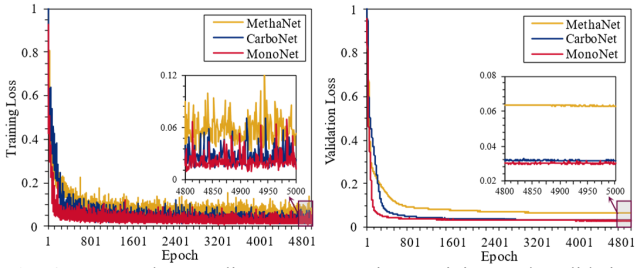


**Fig. 3** Loss Value Decline Curves During Training and Validation Iterations for MethaNet, CarboNet, and MonoNet Models

Model training utilized supervised learning methods, with actual measured emission component concentrations as labels. The AdamW optimization algorithm was chosen to optimize model performance and efficiently search for optimal solutions in a multidimensional parameter space. This algorithm considered both learning rate and weight decay, effectively balancing training speed and model accuracy (Loshchilov et al., 2017). MSE served as the loss function, providing a quantitative measure of the difference between predicted and actual measured values. Cross-validation mechanisms were introduced to reduce the risk of model overfitting and assess generalization ability (Bishop, 2006; Kohavi, 1995). Furthermore, the application of Dropout and L2 regularization further improved the model's prediction accuracy on unknown data (Ng, 2004; Srivastava et al., 2014). Early Stopping was also employed during model training to prevent overtraining and preserve the optimal model state (Prechelt, 2002). This strategy determined the stopping point for training by monitoring performance on the validation set. Main evaluation metrics for model performance included MSE, Root Mean Square Error (RMSE), and $R^2$ scores, aiding in assessing and optimizing model prediction accuracy. Results for the MethaNet, CarboNet, and MonoNet models are presented in Fig. 4.

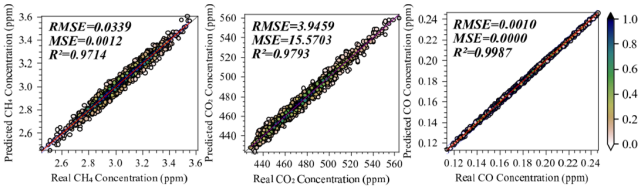


**Fig. 4** Evaluation Results for MethaNet, CarboNet, and MonoNet Models

Through the carefully designed and trained MethaNet, CarboNet, and MonoNet models, strong technical support was provided for the precise calibration of emission concentration baselines. These models not only processed complex input data but also accurately predicted emission component concentrations at specific geographic locations. The models' design, training strategies, and performance evaluation collectively ensured the accuracy and reliability of the predictions, laying a solid foundation for subsequent conversion to carbon equivalent emissions.

## PIXEL-LEVEL CARBON EMISSION CONVERSION

## Precise Calculation of Carbon Equivalent Emissions

In converting pixel-level $CO_2$, CO, and $CH_4$ concentration data into carbon equivalent emissions for the South Marsh Island Area and Eugene Island Area, a method based on physicochemical principles was applied. This approach builds on results from atmospheric concentration calibration and accounts for the effects of environmental conditions on gas behavior (Draxler et al., 1998).

The conversion process starts with an improved model of the ideal gas law to calculate the mass emissions of each gas. This model incorporates gas behavior under specific environmental conditions, with the calculation formula based on the gas's volume fraction, density under standard conditions, and molecular weight (Brock et al., 2008). For $CO_2$, $CH_4$, and CO, their mass emissions calculation formula is shown in Eq. 1.

$$m_{\text{gas}} = P_{\text{gas}} \times V \times \frac{MW_{\text{gks}}}{RT} \quad (1)$$

In this equation, $m_{\text{gas}}$ denotes the mass of the gas, $P_{\text{gas}}$ is the partial pressure of the gas, $V$ represents the volume of the gas, $MW_{\text{gas}}$ is the molecular weight of the gas, $R$ stands for the ideal gas constant, and $T$ is the absolute temperature. Eq. 1 enables the precise calculation of gas mass under specific temperature and pressure, laying the groundwork for subsequent carbon equivalent calculations.

Next, this model further employs the GWP to convert mass emissions into equivalent $CO_2$ emissions. The GWP values follow IPCC standards, with $CO_2$ set to a GWP of 1, while $CH_4$ and CO are assigned GWPs of 28 and 1.9, respectively, based on their global warming impact relative to $CO_2$ (Pachauri et al., 2014). The formula for calculating carbon equivalent emissions is presented in Eq. 2.

$$E_{\text{cqCO2}} = m_{\text{CO2}} + m_{\text{CH4}} \times GWP_{\text{CH4}} + m_{\text{CO}} \times GWP_{\text{CO}} \quad (2)$$

In this equation, $E_{eq\text{CO}_2}$ represents the equivalent $CO_2$ emissions, $m_{\text{CO}_2}$, $m_{\text{CH}_4}$, and $m_{\text{CO}}$ are the mass emissions of these gases, and $GWP_{\text{CH}_4}$ and $GWP_{\text{CO}}$ are the global warming potential values for methane and carbon monoxide, respectively.

This method converted the emissions of different gases into a unified measure of carbon equivalent emissions, providing accurate data support for comprehensive analysis and evaluation of carbon emissions in the region.

## Spatial Analysis and Emission Imaging Generation

Following the prediction of emission component concentrations by the model, a sequence of spatial processing steps was undertaken to transform these data into detailed geospatial distribution maps for the South Marsh Island Area and Eugene Island Area. This process involved coordinate system conversion, data merging, image processing, and the final generation of spatial distribution maps, aimed at accurately depicting the distribution of emission components in specific geographical locations.

Utilizing the previously trained MethaNet, CarboNet, and MonoNet, predictions were made for emission component concentrations at each geographic location point within 73 images spanning from March 6, 2021, to February 24, 2023. The model's input data included geographic location coordinates and reflectance values from 12 bands extracted from remote sensing images. For each data point, the models predicted concentrations of $CH_4$, $CO_2$, and CO, creating a dataset with a large number of predicted values. Prior to prediction, the input data underwent a standardization process identical to that used during training to ensure data consistency and accuracy.

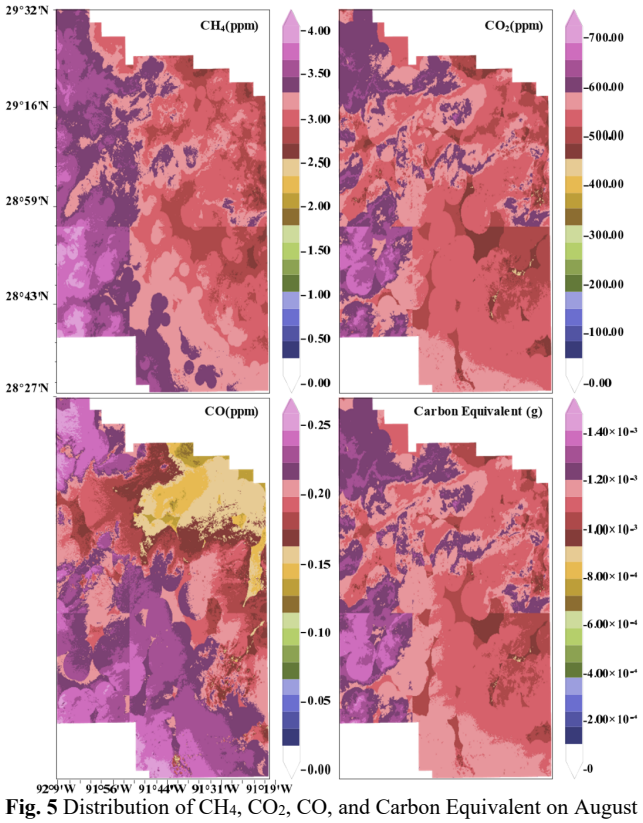


**Fig. 5** Distribution of $CH_4$, $CO_2$, CO, and Carbon Equivalent on August 18, 2022, for the South Marsh Island Area and Eugene Island Area

The emission component concentration data predicted by MethaNet, CarboNet, and MonoNet first underwent a coordinate system transformation, converting original latitude and longitude coordinates to the UTM15N coordinate system more suitable for spatial analysis. This step ensured precise spatial correspondence of the data and compatibility for subsequent GIS processing (Raju, 2006). Further, the predicted data were allocated to virtual sampling points created every 11 pixels, followed by a data merging and spatial expansion process. In this process, data from each sampling point were expanded by 5 pixels to ensure the visual continuity and completeness of the final geospatial distribution map (Longley, 2005). Through the application of GIS technology, further processing was conducted on the merged and expanded data. This included precise trimming of the generated geospatial distribution maps using shp files to ensure complete alignment with the boundaries of the South Marsh Island Area and Eugene Island Area. This step was crucial for generating atmospheric concentration distribution maps that were both spatially accurate and geographically consistent with the study area.

Fig. 5 vividly displays the specific distribution of $CH_4$, $CO_2$, CO, and carbon equivalent emissions in the South Marsh Island Area and Eugene Island Area. These images visually reflect the concentration variations of different emission components across regions, providing important insights into the distribution characteristics of emission components within these areas.

By predicting the spatiotemporal concentrations of emission components and subsequent spatial processing, an accurate benchmark for carbon equivalent emissions was established, offering detailed spatial distribution maps of emission component concentrations for the aforementioned areas. This process not only provided key technical support for the benchmarking of carbon equivalent emissions but also laid a solid foundation for high-precision, pixel-level carbon equivalent

emissions prediction using the ViT-UNet.

## APPLICATION OF VIT-UNET IN CARBON EQUIVALENT EMISSION PREDICTION

### Introduction to the Vit-UNet Network Architecture

The ViT-UNet model stands as the core innovation of this research, specifically designed and developed for precise pixel-level prediction of carbon equivalent emissions. This model ingeniously merges the advantages of the ViT with the U-Net architecture, aiming for efficient feature extraction and spatial information reconstruction, particularly addressing the complex challenges in remote sensing image analysis. By integrating ViT's powerful feature extraction capability with U-Net's excellent performance in spatial information reconstruction, ViT-UNet emerges as an ideal choice for processing hyperspectral data and parsing spatial complexities in remote sensing images. The effectiveness of this method draws inspiration from Dosovitskiy et al.'s (Dosovitskiy et al., 2020) research on ViT and Ronneberger et al.'s (Ronneberger et al., 2015) pioneering work on the U-Net architecture. As shown in Fig. 6, ViT-UNet comprises three main parts: the input transformation layer, the ViT encoder, and a customized decoder (He et al., 2016).

**Input Transformation Layer:** Considering the multispectral nature of remote sensing images, this layer converts multispectral inputs into a 3-channel format compatible with the ViT model through 1x1 convolution. This transformation not only preserves the rich information of remote sensing data but also ensures compatibility with the ViT module.

**ViT Encoder:** The ViT-UNet utilizes a pretrained ViT-B/16 model as its core encoder. Leveraging the self-attention mechanism and multilayer perceptrons, it captures image details and features from local to global levels. This feature enables ViT-UNet to precisely analyze and predict carbon equivalent emissions, providing deep and content-rich feature representations to support subsequent pixel-level predictions.

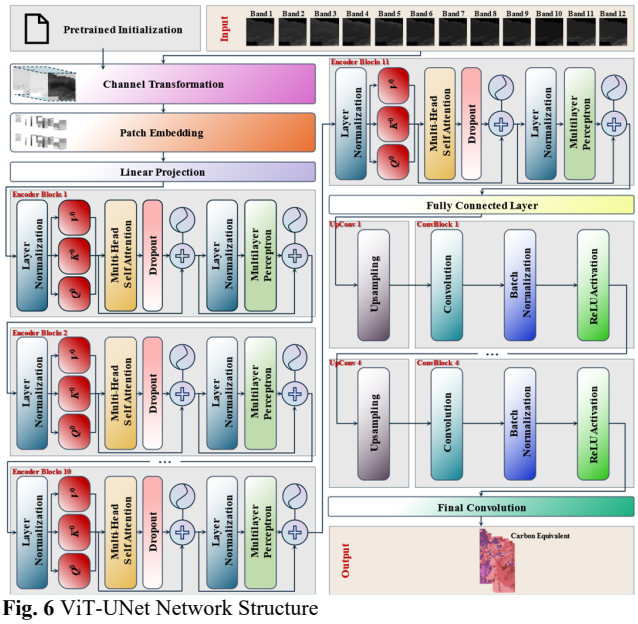


**Fig. 6** ViT-UNet Network Structure

**Customized Decoder:** Unlike traditional U-Net, the decoder of ViT-UNet is specially designed to process the output from the ViT encoder. It gradually reconstructs the spatial resolution of images through a series of upsampling and convolution layers while retaining key features transmitted by the encoder. This design allows ViT-UNet to accurately restore local details while maintaining the global context of images, providing high-precision pixel-level predictions for carbon equivalent emissions.

### Network Training Strategy and Implementation

#### *Input Adaptation of Remote Sensing Image Data*

In building the training framework for ViT-UNet, precise preprocessing and data adaptation of remote sensing images are crucial. The multispectral information contained in the hyperspectral remote sensing images used in this study provides rich terrestrial and atmospheric data for the model. However, these data in their original format are not directly applicable to the ViT-UNet model's processing. Therefore, a series of specially designed preprocessing steps were applied to ensure compatibility between the data and the model while preserving the integrity of the original information to the greatest extent. Given the input requirements of the ViT, each band of the remote sensing images was uniformly scaled to a resolution of 224x224 pixels. This step utilized an efficient bilinear interpolation algorithm to minimize potential information loss during resampling. Bilinear interpolation has been proven to be an effective means of resizing while maintaining image quality, especially important in the processing of hyperspectral remote sensing images (Hu et al., 2015). The next step involved integrating the resampled bands into a multi-channel input tensor. This process involved converting the multispectral data into a 3-channel format through a 1x1 convolution layer, thereby adapting to the input standards of the ViT. This not only served as an effective method for data dimension conversion but also a strategy to preserve data integrity and richness of information (Lin et al., 2013). The 1x1 convolution played a significant role here, not just in converting data formats but also in ensuring the maximum retention of information from the original hyperspectral bands to the 3-channel representation. Furthermore, to enhance the efficiency of data processing and the accuracy of model training, normalization was implemented to ensure uniformity and comparability among different bands. Normalization is a common data preprocessing method in deep learning, helping to improve the stability and convergence speed of model training (Ioffe et al., 2015).

Through these meticulously designed processing workflows, hyperspectral remote sensing images were effectively converted into a format suitable for deep learning models, ensuring the richness and integrity of the data. This laid a solid foundation for subsequent model training and precise pixel-level prediction of carbon equivalent emissions.

#### *Training Process and Optimization Strategies*

In the training strategy for the ViT-UNet model, key optimization strategies were employed to enhance the model's performance and generalization ability. The details and implementation of these strategies significantly impact the model's final prediction accuracy.

An adaptive learning rate adjustment strategy is commonly used in deep learning. This strategy employs learning rate annealing techniques to dynamically adjust the learning rate based on loss changes during training. Initially, the model uses a higher learning rate to explore the parameter space quickly. As training progresses, the learning rate gradually decreases to help the model finely search the parameter space and converge to the optimal solution. This strategy significantly improves training efficiency and the final performance of the model (Smith, 2017).Data augmentation continues to play a crucial role in the training strategy of ViT-UNet. Beyond basic image processing techniques like rotation, scaling, and flipping, more complex methods

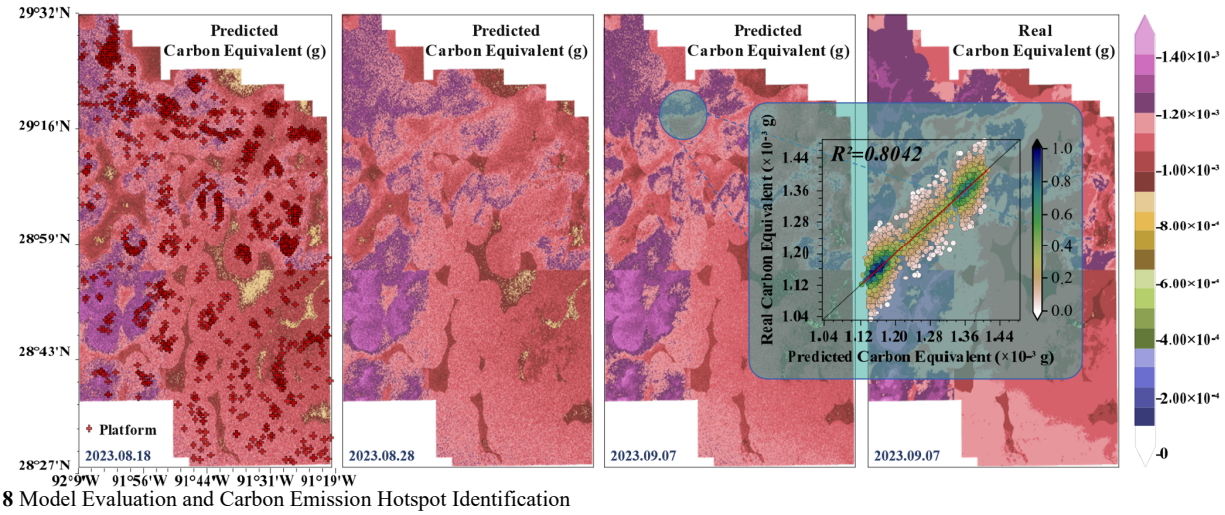


**Fig. 8** Model Evaluation and Carbon Emission Hotspot Identification

such as random noise injection and color space transformations were used. These advanced data augmentation techniques further enhance the model's adaptability to various conditions and changes in remote sensing images. For example, random noise injection can simulate real-world sensor noise, while color space transformations help the model learn to handle images under different lighting conditions. The application of these techniques improves the model's robustness and generalization ability when facing the complexity and variability of the real world (Shorten et al., 2019).

**Table 3** Training Hyperparameters for ViT-UNet

| Hyperparameter | Description | Value |
|---|---|---|
| learning_rate | Learning rate for ViT-UNet model. | 0.0001 |
| optimizer | Algorithm used for model optimization. | AdamW |
| batch_size | Number of samples in each training batch. | 32 |
| num_epochs | Number of iterations over the entire dataset for training. | 100 |
| weight_decay | Regularization term to reduce model overfitting. | 0.01 |
| dropout_rate | Dropout rate applied in ViT-UNet to reduce overfitting. | 0.5 |
| early_stopping_patience | Patience parameter for early stopping mechanism to prevent overfitting. | 10 |
| learning_rate_scheduler | Scheduler for adjusting the learning rate. | StepLR |
| step_size | Step size for learning rate adjustment. | 10 |
| gamma | Decay factor for learning rate adjustment. | 0.1 |

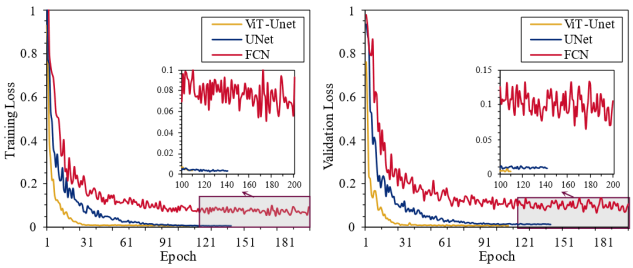


**Fig. 7** Loss Value Decline Curves During Training and Validation Iterations for ViT-UNet, U-Net, and FCN Models

The ViT-UNet model adopts MSE as the primary loss function. In the task of pixel-level prediction of carbon equivalent emissions, MSE provides an intuitive way to assess the accuracy of the model in predicting the concentration of carbon equivalent emissions for each pixel, which is crucial for the objectives of this study. Specific training parameters are detailed in Table 3, with U-Net and Fully Convolutional Networks (FCN) serving as baseline models for concurrent training with the ViT-UNet model. The loss values during the training and validation process for ViT-UNet, U-Net, and FCN models are shown in Fig. 7.

In comparisons among the three models, the ViT-UNet model demonstrated significant advantages in both training and validation loss values, recording the lowest loss value at 0.0037. This outstanding performance can be attributed to the advanced structure of the ViT-UNet model and the effectiveness of optimization strategies, especially the application of adaptive learning rate adjustment strategies, enabling more efficient exploration in the parameter space and rapid convergence to the optimal solution. Compared to ViT-UNet, the U-Net model recorded a lowest loss value of 0.0045, while the FCN model had a significantly higher lowest loss value of 0.0510. Moreover, the introduction of a pretrained Transformer model improved the ability of the ViT-UNet model to handle complex spatial relationships, resulting in faster declines in training and validation loss values compared to U-Net and FCN models.

## Model Performance Evaluation and Carbon Emission Hotspot Identification

Within the framework of developing the ViT-UNet model, the model underwent a comprehensive performance evaluation. This assessment aimed to verify its effectiveness and accuracy in pixel-level prediction of carbon equivalent emissions. The evaluation strategy combined both quantitative and qualitative analyses to ensure a deep understanding of the model's performance.

The qualitative analysis included a visual comparison between the model's output and actual ground-measured data. This comparison revealed the model's capability to accurately reconstruct carbon equivalent emission distribution maps for specific areas, providing intuitive feedback for model optimization. Notably, by juxtaposing the model's predicted carbon equivalent emission images against oil and gas platform locations, carbon emission hotspots were effectively identified. Further, by analyzing differences between model predictions and actual carbon equivalent emission maps, prediction biases in specific bands or

areas were identified, allowing for fine-tuning of model parameters.

Quantitative evaluation primarily relied on the $R^2$ score, a key metric quantifying the correlation between model predictions and actual values (Fawcett, 2006). This metric is crucial for assessing the accuracy of pixel-level carbon equivalent emission predictions. The analysis of the model's output images involved a specially designed remapping process, enabling pixel-level quantitative assessment of carbon equivalent emissions. Relying on the "tab20b" color mapping and its normalization parameters, this process accurately converted image color values back to original carbon equivalent emission data. The application of color mapping and normalization in the forward mapping process turned numerical data into color representations in images, while the remapping process ensured that color values could be precisely converted back to their corresponding numerical ranges, directly supporting the model's ability for accurate predictions and quantification of carbon emissions.

As depicted in Fig. 8, qualitative analysis showed that areas surrounding oil and gas platforms, especially regions with dense platform aggregations, exhibited significantly higher carbon equivalent emissions than other areas. This observation intuitively reflected the model's ability to identify and depict carbon emission hotspots. By comparing high-resolution images output by the ViT-UNet model with actual calibrated emission concentration images, the model demonstrated its proficiency in capturing the distribution of carbon emissions within the area. High emission zones near oil and gas platforms were distinctly identified in the images, confirming the model's advantage in maintaining image quality and spatial resolution.

Further reinforcing these findings, quantitative evaluation maintained high $R^2$ scores when comparing carbon equivalent emissions represented by specific pixel points in predicted images against actual images, indicating high accuracy in model predictions. Additionally, the model predicted carbon equivalent emission images for specific dates in 2023, revealing a slight upward trend in carbon emissions over time. Subtle variations in pixel colors illustrated this trend, further confirming the model's effectiveness in capturing short-term changes in carbon emissions.

This remapping method not only enabled the ViT-UNet model to accurately depict the spatial distribution of carbon emissions visually but also to achieve precise pixel-level carbon equivalent emission data extraction. Introducing this approach elevated the model's evaluation beyond mere qualitative visual comparisons to a quantitative analysis phase based on statistical metrics such as the $R^2$ score. Therefore, the application of the ViT-UNet model in carbon emission monitoring has not only showcased its image generation capabilities but also highlighted its significant advantage in providing precise, quantifiable carbon emission data.

## CONCLUSION

This study dedicated itself to developing and implementing an advanced framework based on ViT-UNet, focusing on the precise quantification and in-depth analysis of carbon equivalent emissions around offshore platforms. By meticulously processing remote sensing images and corresponding $CH_4$, $CO_2$, and CO monitoring data from 2021 to 2023 for the Gulf of Mexico's South Marsh Island Area and Eugene Island Area, and integrating the model's innovative design with optimized training strategies, this research successfully captured and analyzed the spatial distribution of carbon emissions in the target areas, demonstrating significant capabilities in ensuring prediction accuracy and reliability. This progress not only provides the petroleum industry with rich knowledge and tools for implementing strategic carbon management but also contributes substantively to global efforts in reducing carbon emissions caused by hydrocarbons, showcasing the potential and value of deep learning technology in environmental monitoring.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 52204017) and the National Key Research and Development Program of China (Grant No. 2022YFC28061004).

## REFERENCES

Bishop, C (2006). "Pattern recognition and machine learning," Springer google schola, 2, 531-537.

Brock, C, Sullivan, A, Peltier, R, Weber, R, Wollny, A, De Gouw, J, Middlebrook, A, Atlas, E, Stohl, A and Trainer, M (2008). "Sources of particulate matter in the northeastern United States in summer: 2. Evolution of chemical and microphysical properties," Journal of Geophysical Research: Atmospheres, 113.

Burgués, J and Marco, S (2023) Air Quality Networks: Data Analysis, Calibration & Data FusionSpringer,

Campbell, JB and Wynne, RH (2011) Introduction to remote sensing. Guilford press,

Castleman, KR (1996) Digital image processing. Prentice Hall Press,

Crosson, E (2008). "A cavity ring-down analyzer for measuring atmospheric levels of methane, carbon dioxide, and water vapor," Applied Physics B, 92, 403-408.

Dosovitskiy, A, Beyer, L, Kolesnikov, A, Weissenborn, D, Zhai, X, Unterthiner, T, Dehghani, M, Minderer, M, Heigold, G and Gelly, S (2020). "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929.

Draxler, RR and Hess, G (1998). "An overview of the HYSPLIT_4 modelling system for trajectories," Australian meteorological magazine, 47, 295-308.

Duman, Z, Mao, X, Cai, B, Zhang, Q, Chen, Y, Gao, Y and Guo, Z (2023). "Exploring the spatiotemporal pattern evolution of carbon emissions and air pollution in Chinese cities," Journal of Environmental Management, 345, 118870.

Elstohy, R and Ali, EM (2023). "A flash flood detected area using classification-based image processing for sentinel-2 satellites data: A case study of Zafaraana Road at Red Sea," The Egyptian Journal of Remote Sensing and Space Science, 26, 807-814.

Erkkilä, A, Tenkanen, M, Tsuruta, A, Rautiainen, K and Aalto, T (2023). "Environmental and Seasonal Variability of High Latitude Methane Emissions Based on Earth Observation Data and Atmospheric Inverse Modelling," Remote Sensing, 15, 5719.

Fawcett, T (2006). "An introduction to ROC analysis," Pattern recognition letters, 27, 861-874.

Goodfellow, I, Bengio, Y and Courville, A (2016) Deep learning. MIT press,

Gorroño, J, Guanter, L, Graf, LV and Gascon, F (2023). "A software tool for the estimation of uncertainties and spectral error correlation in Sentinel-2 Level-2A data products."

He, K, Zhang, X, Ren, S and Sun, J (2016). "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

Hjellbrekke, A-G and Solberg, S (2022) Ozone measurements 2020. NILU,

Hu, F, Xia, G-S, Hu, J and Zhang, L (2015). "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," Remote Sensing, 7, 14680-14707.

Iglewicz, B and Hoaglin, DC (1993) Volume 16: how to detect and handle outliers. Quality Press,

Ioffe, S and Szegedy, C (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," International conference on machine learning, pmlr, 448-456.

Juselius, J (2023). "Methane emissions mapping: analysing modern

techniques for accurate assessment and monitoring."
Kohavi, R (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," Ijcai, Montreal, Canada, 1137-1145.
Lee, H, Calvin, K, Dasgupta, D, Krinner, G, Mukherji, A, Thorne, P, Trisos, C, Romero, J, Aldunce, P and Barret, K (2023). "IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland."
Lennon, R (2002). "Remote sensing digital image analysis: An introduction," United States: Esa/Esrin.
Liang, R, Zhang, Y, Chen, W, Zhang, P, Liu, J, Chen, C, Mao, H, Shen, G, Qu, Z and Chen, Z (2023). "East Asian methane emissions inferred from high-resolution inversions of GOSAT and TROPOMI observations: a comparative and evaluative analysis," Atmospheric Chemistry and Physics, 23, 8039-8057.
Lin, M, Chen, Q and Yan, S (2013). "Network in network," arXiv preprint arXiv:1312.4400.
Liu, Z, Deng, Z, Davis, S and Ciais, P (2023). "Monitoring global carbon emissions in 2022," Nature Reviews Earth & Environment, 4, 205-206.
Longley, P (2005) Geographic information systems and science. John Wiley & Sons,
Loshchilov, I and Hutter, F (2017). "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101.
Lowe, DG (2004). "Distinctive image features from scale-invariant keypoints," International journal of computer vision, 60, 91-110.
Ng, AY (2004). "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," Proceedings of the twenty-first international conference on Machine learning, 78.
Pachauri, RK, Allen, MR, Barros, VR, Broome, J, Cramer, W, Christ, R, Church, JA, Clarke, L, Dahe, Q and Dasgupta, P (2014) Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Ipcc,
Prechelt, L (2002) Neural Networks: Tricks of the tradeSpringer,
Raju, P (2006). "Fundamentals of geographical information system," Satellite Remote Sensing and GIS Applications in Agricultural Meteorology, 103.
Ronneberger, O, Fischer, P and Brox, T (2015). "U-net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 234-241.
Shi, T, Han, G, Ma, X, Pei, Z, Chen, W, Liu, J, Zhang, X, Li, S and Gong, W (2023). "Quantifying strong point sources emissions of $CO_2$ using spaceborne LiDAR: Method development and potential analysis," Energy Conversion and Management, 292, 117346.
Shorten, C and Khoshgoftaar, TM (2019). "A survey on image data augmentation for deep learning," Journal of big data, 6, 1-48.
Smith, LN (2017). "Cyclical learning rates for training neural networks," 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, 464-472.
Srivastava, N, Hinton, G, Krizhevsky, A, Sutskever, I and Salakhutdinov, R (2014). "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, 15, 1929-1958.
Su, M, Shi, Y, Yang, Y and Guo, W (2023). "Impacts of different biomass burning emission inventories: Simulations of atmospheric $CO_2$ concentrations based on GEOS-Chem," Science of The Total Environment, 876, 162825.
Wu, C-Y, Zhang, X-Y, Guo, L-F, Zhong, J-T, Wang, D-Y, Miao, C-H, Gao, X and Zhang, X-L (2023). "An inversion model based on GEOS-Chem for estimating global and China's terrestrial carbon fluxes in 2019," Advances in Climate Change Research, 14, 49-61.
Zhao, S, Liu, M, Tao, M, Zhou, W, Lu, X, Xiong, Y, Li, F and Wang, Q (2023). "The role of satellite remote sensing in mitigating and adapting to global climate change," Science of The Total Environment, 166820.