*Sensors and Materials*, Vol. 37, No. 1 (2025) 193–205
MYU Tokyo



S & M 3900

# Mamba-based Multibranch State Space Iterative Fusion Algorithm for Multisource Power Grid Survey Data

Aiyuan Zhang,[1] Jingguo Lv,[1*] Jiyong Zhang,[2] Xiaohu Sun,[2]
Chunhui Zhao,[2] Changjiang Yang,[3] and Junjie Sun[4]

[1]Beijing University of Civil Engineering and Architecture, Beijing 102616, China
[2]State Grid Economic and Technical Research Institute Co., Ltd., Beijing 102200, China
[3]Central Southern China Electric Power Design Institute Co., Ltd. of China Power Engineering Consulting Group, Wuhan 430071, China
[4]Economic Research Institute of State Grid ZheJiang Electric Power Company, Hangzhou 310020, China

(Received July 26, 2024; accepted January 20, 2025)

***Keywords:*** multisource remote sensing data fusion, power grid surveying, Mamba, iterative attention mechanism

The effective integration of multisource survey data for power grids benefits designers by providing comprehensive and accurate analyses of the terrain and landforms surrounding the survey area. In this study, inspired by the Mamba concept, we propose an iterative attentional feature fusion Mamba (iAFF-FMA) framework that constructs a multibranch state space for iterative fusion, reducing differences between data modalities and enhancing feature interaction within the same modality. Experiments conducted with actual engineering data from ultra-high-voltage direct current (UHVDC) transmission lines demonstrate the iAFF-FMA framework's superiority over six common fusion methods. This offers a novel technical approach to the integration of power grid survey data.

## 1. Introduction

The rapid evolution of sensor technology has endowed the power grid surveying sector with access to a diverse array of data sources, including satellite panchromatic imagery, multispectral imagery, Digital Elevation Model (DEM), and geotechnical and hydrometeorological datasets. Despite their individual merits, these data sources alone are insufficient to fully represent the terrain and landforms of the surveyed area. The integration of these diverse datasets is essential for gaining a holistic and profound understanding of the topography, thereby enhancing the reliability of decision-making for power grid planning and maintenance. This integrated approach is pivotal for advancing the field of power grid surveying. Understanding and analysis of the terrain provide more reliable decision support for power grid planning and maintenance.

Multisource data fusion methods are broadly categorized into five types: pixel-level,[(1)] feature-level,[(2–5)] decision-level,[(6–8)] model-based methods,[(9–11)] and hybrid methods. Hybrid

*Corresponding author: e-mail: lvjingguo@bucea.edu.cn
https://doi.org/10.18494/SAM5257

ISSN 0914-4935 

fusion methods primarily include direct fusion,[12] attention-based fusion,[13] and multistage fusion. Multistage fusion, which combines multiscale and attention mechanisms, captures features at different levels and is conducive to fully integrating multiscale features. Common methods include Pathformer,[14] IFT,[15] CDDFuse,[16] and SwinFusion.[17] However, these methods often require high computational resources and time owing to the complexity of attention mechanisms, especially for large-scale data fusion. The emerging Mamba model of 2023, with its selective mechanism and hardware-aware efficient design, is more suitable for large-scale data fusion. However, Mamba is mostly applied to object detection, and its application in image fusion is still rare, with only a few fusion models like MambaDFuse[18] and FusionMamba[19] existing.

Power grid survey data significantly differ from other data types in their imaging methods and formats, often originating from specialized remote sensing technologies such as high-resolution satellite and aerial imagery, and LiDAR technology that provides a detailed three-dimensional perspective on terrain. The data may include multispectral imaging, extending beyond the visible light to include infrared and ultraviolet spectra, thus offering additional dimensions for the analysis of terrain and vegetation. Moreover, power grid surveys encompass geological and meteorological data, which often come in nonimage formats, adding to the complexity of data fusion. The datasets are typically vast, with rich spatial and temporal dimensions, necessitating fusion algorithms capable of managing large-scale and high-dimensional data with precise spatial and temporal alignment. To address these challenges, in this paper, we introduce an iterative fusion framework based on the Mamba model (iAFF-FMA), which performs deep iterative attentional feature fusion within a hidden state space, aiming to enhance the quality and efficiency of data fusion for power grid surveys. The main contributions of this paper are as follows.

(1) To fully exploit the characteristics of multisource data, a differentiated feature extraction module is proposed, tailoring extraction strategies for terrain remote sensing, multispectral imagery, DEM, and so forth, effectively addressing the issue of incomplete feature extraction.
(2) To enhance the efficiency and accuracy of multisource power grid survey data fusion, a Mamba-based iAFF-FMA framework is designed, integrating visual state space and iterative attention mechanisms and significantly improving the efficiency and quality of fusion.
(3) To compensate for the loss of detail in feature fusion, iAFF is introduced, where an iterative attention fusion Mamba block is designed in the state space to address the simplistic addition or parameterized fusion of features in previous fusion Mamba approaches.

## 2. Related Work

### 2.1 State space models (SSMs)

SSMs are mathematical models used to describe and predict system states. They generate outputs $y(t) \in R$ by passing a 1D input sequence $x(t) \in R$ through intermediate hidden states $h(t) \in R^N$. Mathematically, SSM is often formulated as a system of linear ordinary differential equations (ODEs).

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

Here, system behavior is defined by a set of parameters, including the state transition matrix $A \in R^{N\times N}$, observation matrix $B$, $C \in R^{N\times 1}$, and the noise impact matrix $D \in R$. $h(t)$ represents the latent state at any given time $t$.

The matrices $A$ and $B$ are discretized using zero-order hold with a time scale parameter $\Delta$. This discretization process is illustrated as follows:

$$\begin{aligned} \overline{A} &= e^{\Delta A}, \\ \overline{B} &= (\Delta A)^{-1}\left(e^{\Delta A} - I\right)\Delta B, \\ \overline{C} &= C. \end{aligned} \tag{2}$$

After discretization, the SSM is computed by a global convolution with structured convolutional kernels.

### 2.2 Mamba

In SLAM, SSMs depict estimated states with observation and prediction components. Mamba's SSM is inspired by these traditional models and has broadened its application from NLP to computer vision since its inception.

VMamba introduces the SS2D, an innovative four-directional scanning algorithm that enhances the MAMBA-based visual backbone, surpassing the Swin Transformer in performance for object detection, segmentation, and tracking. The SS2D's strength is its comprehensive capture of image context through scanning in horizontal, vertical, and diagonal directions. It optimizes computational efficiency and feature representation, showing robust adaptability and stability across visual data of different resolutions and complexities. Moreover, the synergy between SS2D and the MAMBA model significantly boosts the processing of two-dimensional visual data, enabling precise control and effective integration of information in visual tasks, which minimizes information loss. Despite Mamba's global modeling capabilities and linear complexity, its use in image fusion via SSM remains limited.

## 3. Methods

In this paper, we present a fusion model for multisource data in power grid surveying, addressing the challenges of diverse data types and large volumes. Drawing on deep learning's feature extraction capabilities and inspired by FusionMamba, we introduce a model that maintains linear complexity and global perception while optimizing computation with dynamic weight allocation.

The iAFF-FMA model fuses features in a hidden state space, using an iterative attention mechanism to dynamically integrate feature weights based on correlations, culminating in a comprehensive feature representation $F_f$, as shown in Fig. 1.

### 3.1 Data preprocessing

Before fusion research on power grid survey data, preprocessing is essential to standardize data within a consistent spatiotemporal framework. This involves the following.

Time synchronization is achieved by setting a common reference point for all data sources, using the panchromatic terrain imagery from January 20, 2024, to ensure temporal consistency. A temporal backdrop is established, covering the period from January 10 to 30, 2024, with daily resolution to align data points in time.

Spatial registration aligns diverse data sources to a common resolution and geocoding standard in the WGS 84 system. Imagery and DEM data are resampled to 0.5 m/pixel resolution using bilinear interpolation. Location data from exploration and meteorological reports are georeferenced for map overlay. The UTM projection is used for consistent spatial alignment.

### 3.2 Multisource data feature extraction

(1) Terrain Panchromatic Image Feature Extraction

In this study, we employed LightM-UNet to extract critical features from high-resolution panchromatic images, this is essential for detailing terrain features in power grid surveys. As shown in Fig. 2. The process starts with an input image $X_P \in R^{C\times H\times W}$, where LightM-UNet uses depthwise separable convolution (DWConv) layers to initially capture basic features.

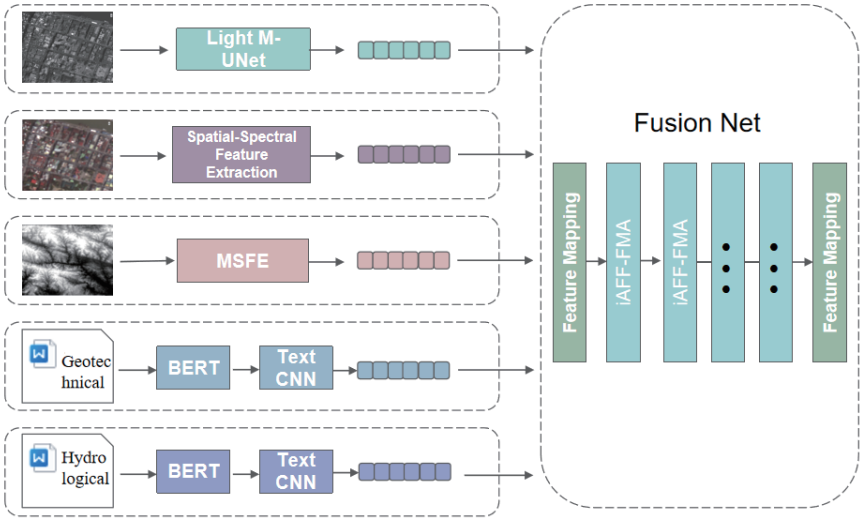


Fig. 1. (Color online) iAFF-FMA multisource survey data fusion network flowchart.

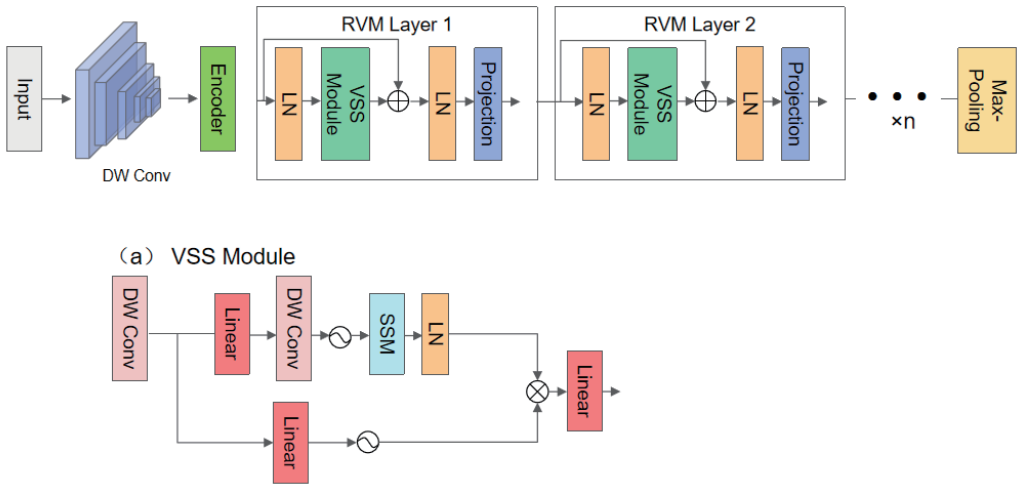


Fig. 2. (Color online) Light M-UNet feature extraction module.

(2) Feature Extraction of Tower-site Multispectral Imagery

We used group convolution on multispectral imagery to enhance the detail of spectral and spatial features in power grid surveys.

First, the tower-site multispectral image $X_M$ is fed into several spectral feature extraction modules to determine the size $R$ of the convolution kernel along the spectral dimension. Then, for each layer $i$ in the convolutional neural network and each feature map $j$ within that layer, the dot product of the kernel with the input feature map is calculated over the local region in the spectral dimension $z$ at every spatial position ($x$, $y$), yielding the output value $F_{ij}^{xyz}$. Ultimately, these operations effectively extract and emphasize spectral features within multispectral images.

Additionally, to ensure that spatial information is not mixed with spectral features, we employed group convolution instead of standard 2D convolution on the spatial dimension. Group convolution initially appeared in AlexNet to address the hardware limitations of the time, as shown in Fig. 3, resulting in compact multispectral image features for tower sites $F_M$.

(3) DEM Data Feature Extraction

Elevation data from DEM is essential for power grid surveys, detailing the terrain's slope, aspect, and contours. We use multiscale dilated convolution to adeptly capture this data's depth.

In our study, we introduce a three-branch convolutional module, each with kernels of varying sizes (1, 3, 5) for scale-specific feature capture, as shown in Fig. 4. For the kernel size of 1, we used average pooling to smooth minor noise and preserve the general terrain trends. The kernel size of 3 employed K-Max pooling to highlight key changes in terrain, such as ridges or valleys. Kernels of size 5 utilized max pooling to emphasize local extreme features of the terrain, such as steep slopes or obstacles. These features converge in a fully connected

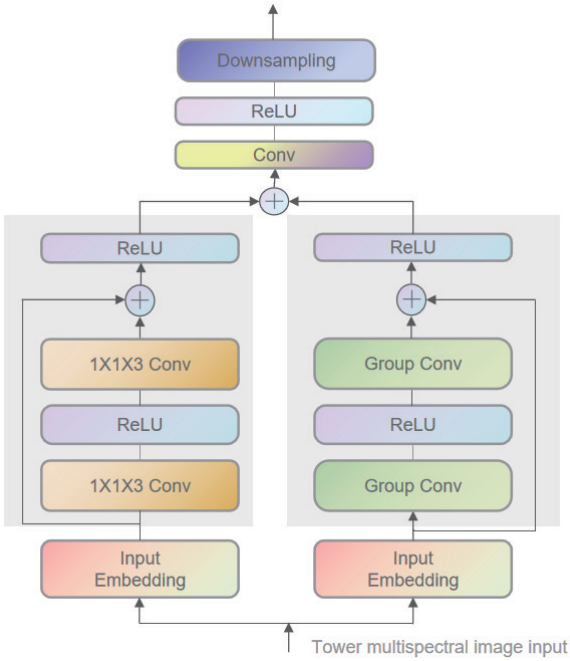


Fig. 3. (Color online) Spatial-spectral dual-branch feature extraction module.

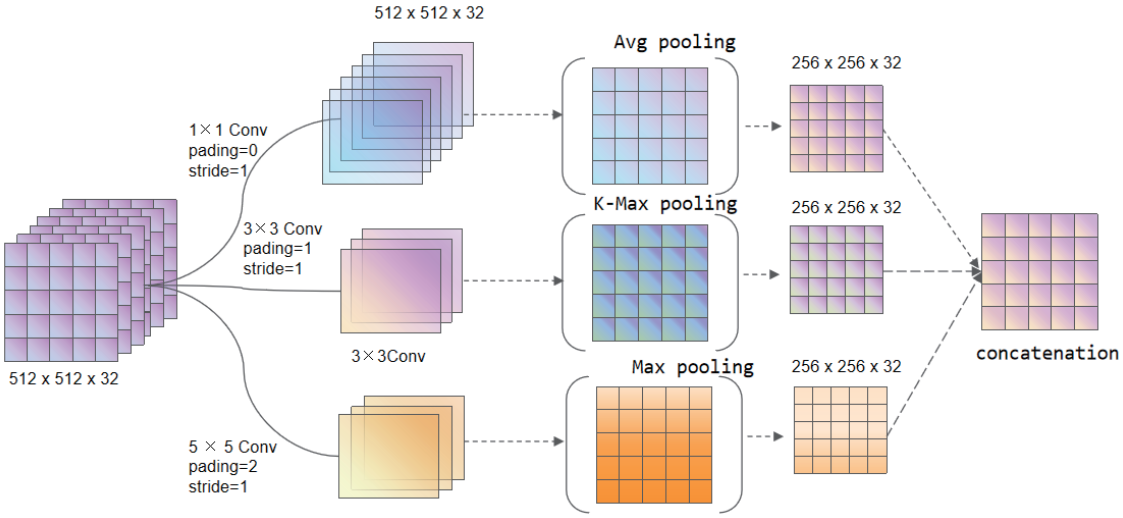


Fig. 4. (Color online) Improved multiscale feature extraction module for DEM data.

layer, generating a feature sequence $F_D$ that encompasses a wealth of scale information, providing a comprehensive description of terrain features for power grid surveys.

(4) Geotechnical Exploration and Hydrometeorological Data Feature Extraction

Geotechnical and hydrometeorological data, often textual, are analyzed using the TextCNN model in this study. It extracts key details from descriptions of geological conditions for

transmission tower foundations. The text is tokenized and then transformed into word embeddings by BERT before being processed by TextCNN. Features are captured through convolutional kernels of varying sizes and enhanced by ReLU activation and Max pooling, ultimately yielding geotechnical features $F_R$ and hydrometeorological features $F_H$.

### 3.3 Multibranch state space iAFF-FMA fusion

In this paper, we introduced a channel shuffle module to enhance the interaction between multimodal feature channels. It segments input feature maps into groups, recombines them, and thus boosts information exchange and feature integration.

Specifically, the panchromatic terrain image $F_P \in R^{N\times C\times H\times W}$ is divided into $G$ groups on the basis of $C$ channels, with each group containing $C/G$ channels. This results in $G$ groups of feature maps $F_P^1$, $F_P^2$, …, $F_P^G \in R^{N\times C/G\times H\times W}$. Subsequently, a channel shuffle is performed; for each group $F_P^i$, the feature maps are rearranged along the channel dimension $C/G$ to obtain $\hat{F}_P^i$. Then, all the rearranged groups $\hat{F}_P^i$ are merged back along the channel dimension to the original number of channels $C$, generating features $T_P$, $T_M$, $T_D$, $T_R$, $T_H \in R^{N\times C\times H\times W}$ that incorporate channel interaction information.

To enhance cross-modal feature association and complementarity, we established a hidden state space based on Mamba. We propse to project features from five modalities onto the hidden state space and utilize a gating mechanism to construct the transition of the hidden states. Furthermore, the iAFF method is employed to achieve deep cross-modal feature fusion.

After obtaining the individual channel interaction features $T_P$, $T_M$, $T_D$, $T_R$, and $T_H$, they are initially projected into the hidden state space through a VSS block without gating.

We optimized the model by substituting depthwise convolutions with grouped convolutions, which divide input channels into groups for independent kernel processing. The grouped convolution outputs are ReLU-activated, normalized, and then input into SS2D to expand feature maps in four directions, enhancing contextual comprehension.

Subsequently, by projecting $\tilde{y}_P$, $\tilde{y}_M$, $\tilde{y}_D$, $\tilde{y}_R$, and $\tilde{y}_H$, gating parameters $Z_P$, $Z_M$, $Z_D$, $Z_R$, and $Z_H$ are obtained. These parameters play a crucial role in the gating mechanism, which controls the flow of information through the network.

Then, to fully leverage the cross-branch information complementarity through dual attention, the gated outputs $Z_P$, $Z_M$, $Z_D$, $Z_R$, and $Z_H$ are used to adjust $y_P$, $y_M$, $y_D$, $y_R$, and $y_H$ , achieving the fusion of hidden state features into $\tilde{y}_P$, $\tilde{y}_M$, $\tilde{y}_D$, $\tilde{y}_R$, and $\tilde{y}_H$. This results in the hidden state features that represent the interaction of features post-fusion.

Traditional feature fusion methods typically involve mere addition or concatenation of features without considering the suitability of fusing features from different modalities. In this study, by introducing an iAFF block into the state space, a multiscale channel attention mechanism is utilized to better integrate multimodal features that vary in scale and semantic inconsistency.

From the previous steps, features from five modalities are obtained, and the features of these five modalities, denoted as $\tilde{y}_P$, $\tilde{y}_M$, $\tilde{y}_D$, $\tilde{y}_R$, and $\tilde{y}_H$ , are subjected to iterative iAFF integration. Initially, each iteration step comprises two attention modules: one for the features of the current

modality, $Att(\tilde{y}_P)$, and the other for the features of another modality, $Att(\tilde{y}_M)$. In the first iteration, weighted representations for the features of each modality are computed using the gating parameters $Z_P$, $Z_M$, $Z_D$, $Z_R$, and $Z_H$ and the attention modules:

$$\begin{aligned} Att(\tilde{y}_P) &= Z_{y_P} \cdot Attention(\tilde{y}_P)\, Att(\tilde{y}_M) + Z_{y_P} \cdot Attention(\tilde{y}_M)\, Att(\tilde{y}_P), \\ Att(\tilde{y}_M) &= Z_{y_M} \cdot Attention(\tilde{y}_M)\, Att(\tilde{y}_P) + Z_{y_M} \cdot Attention(\tilde{y}_P)\, Att(\tilde{y}_M). \end{aligned} \tag{3}$$

The attention mechanism assigns different weights to each part of the features, highlighting the important ones while suppressing the less important ones. Subsequently, the weighted feature representations are iteratively fused:

$$F_{P,M}^{i} = \alpha\left(Att(\tilde{y}_P) + Att(\tilde{y}_M)\right) \otimes Att(\tilde{y}_P) + \left(1 - \alpha\left(Att(\tilde{y}_P) + Att(\tilde{y}_M)\right)\right) \otimes Att(\tilde{y}_M), \tag{4}$$

where $F_{P,M}^{i} \in R^{C \times H \times W}$. The + symbol represents the initial feature integration. ⊗ signifies element-wise multiplication.

In the subsequent iterations, the output from the previous iteration is used as the input for the current attention module, and the formula is applied repeatedly. Each iteration generates new features $F_{P,M}^{i}$, where $i$ denotes the iteration number. After the iterative process is completed, the output from the last iteration, $F_{P,M}$, serves as the final integrated feature representation.

As shown in Fig. 5. Ultimately, $F_{P,M,D,R,H}$ is projected back to the original space to obtain the multimodal fused features.

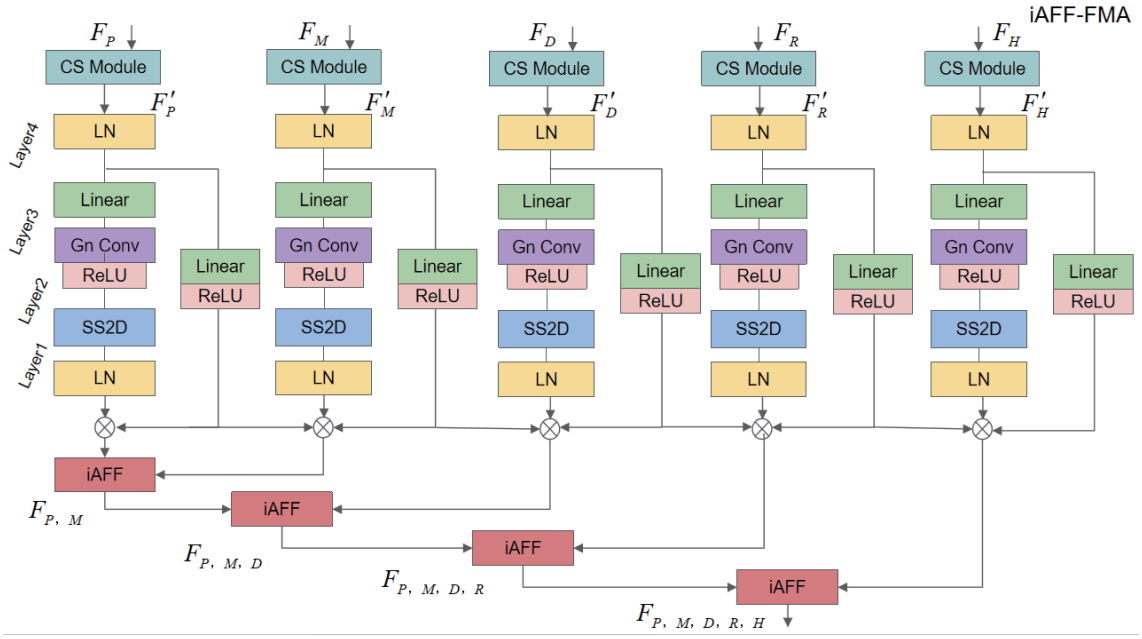


Fig. 5. (Color online) iAFF-FMA feature fusion schematic diagram.

## 4. Experiments

### 4.1 Experimental setup

To assess our proposed fusion method, we compared it with standard techniques such as SDNet, FHM, DeFusion, DIDFuse, CDDFuse, and Super Fusion. The experimental dataset, derived from the 2023 Ningxia section of the Ningxia-Hunan ±800 kV UHV DC transmission line project, is meticulously segmented into training and testing sets. The training set encompasses a substantial 80%, totaling around 200000 panchromatic terrain images, 100000 multispectral images, 80000 DEM samples, 60000 geotechnical samples, and 40000 hydrometeorological samples. The testing set constitutes the remaining 20%, with approximately 50000 panchromatic images, 25000 multispectral images, 20000 DEM samples, and 10000 samples each for geotechnical and hydrometeorological categories. This comprehensive dataset facilitates rigorous validation of the proposed fusion methodologies within the power grid surveying domain. The experiments were conducted on a server equipped with AMD Ryzen 7 PRO 5845, NVIDIA T1000 GPU, and 16 GB of RAM using TensorFlow, taking into account the area's challenging terrain and climate.

The data comprised GeoEye-1 satellite panchromatic images (0.5 m/pixel), drone-captured multispectral images (0.3 m/pixel), and LiDAR-derived DEM data (2 m/pixel). Fusion effectiveness was gauged by *RMSE*, *UIQI*, *DD*, and *mAP* at IoU 0.50, with *mAP* evaluated across a range from 0.50 to 0.95 in 0.05 increments. Higher metric values signify better model performance.

(1) *RMSE* quantifies the discrepancy between predicted and actual values. It is calculated as

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}\,, \tag{5}$$

where $y_i$ is the actual value, $\hat{y}_i$ is the predicted value, and m is the number of samples. A lower *RMSE* indicates a higher model accuracy.

(2) *UIQI* assesses image quality by considering luminance, contrast, and structure. The formula is

$$UIQI = \frac{1}{k_1 k_2}\sum_{l=1}^{k_1}\sum_{k=1}^{k_2}\frac{I_{max,k,l} - I_{min,k,l}}{I_{max,k,l} + I_{min,k,l}}\log\left(\frac{I_{max,k,l} - I_{min,k,l}}{I_{max,k,l} + I_{min,k,l}}\right). \tag{6}$$

Here, $I_{max,k,l}$ and $I_{min,k,l}$ represent the maximum and minimum values of each color component in the RGB space, and $k_1$ and $k_2$ are the dimensions of the image blocks.

(3) *DD* is a measure of the difference between two images and is often used to assess similarity. The specific calculation depends on the context, but it generally involves pixelwise differences. The formula for *DD* is context-dependent and thus not provided here. A lower *DD* indicates greater image similarity.

(4) *mAP* is a performance metric in object detection, being a measure of the average precision across categories. It is calculated as

$$mAP = \frac{1}{k}\sum_{k=1}^{K} AP_k \ , \tag{7}$$

where $K$ is the number of categories, and $AP_k$ is the $AP$ for category $kk$. $AP$ is the area under the precision–recall curve, or AUC, for each category. $mAP$ is computed at various IoU thresholds (from 0.50 to 0.95 with increments of 0.05) and then averaged to provide a comprehensive performance evaluation.

### 4.2 Ablation experiment

Here, we introduce an iAFF module within our Mamba-based power grid survey data fusion framework to enhance the fusion within the hidden state space. Ablation studies were conducted to test the iAFF module's impact by varying its components. As shown in Table 1, results indicated a 2.2% drop in $mAP_{50}$ and a 0.8% drop in $mAP$ without the iAFF block, highlighting its role in reducing modality differences and aligning feature representations, which is crucial for effective deep fusion.

### 4.3 Comparative experiment

In this experiment, we meticulously preprocessed multisource power grid survey data, including panchromatic terrain images, multispectral tower-site images, and DEM data, to remove noise and retain essential information. The data were cropped and split into training and testing sets at an 80:20 ratio, following specific spatial resolution guidelines.

In the experiment, we integrated digitized data from geotechnical and hydrometeorological reports as additional features, enriching the model's environmental and geological insights. The fusion performance was assessed using metrics such as $mAP$ and $mAP_{50}$. Our iAFF-FMA network outperformed other methods, increasing $mAP$ by 0.9% and $mAP_{50}$ by 1.0%. Although there was a noted increase in distortion, likely due to spatial resolution differences in the preprocessed data, the overall fusion performance of iAFF-FMA was superior.

To validate the effectiveness of the methods presented in this paper, as shown in Table 2, we have selected a variety of comparative algorithms, including three component substitution (CS) methods (HIS,[20] BDSD,[21] and PRACS[22]), two multiresolution analysis (MRA) methods (MTF-GLP[23] and Indusion[24]), one variational optimization (VO) algorithm (GND[25]), and six

Table 1
The iAFF module has a significant impact on the fusion model.

| Methods | $mAP_{50}$ | $mAP_{70}$ | $mAP$ | Param | Time (ms) |
|---|---|---|---|---|---|
| iAFF-FMA | 85.6 | 46.5 | 48.9 | 296.2M | 82 |
| Removing iAFF | 83.4 | 45.3 | 48.1 | 278.5M | 76 |

Table 2
Fusion experiment performance metrics.

| Comparison Methods | *RMSE* | *UIQI* | *DD* | *mAP* | $mAP_{50}$ |
|---|---|---|---|---|---|
| IHS | 13.54 | 0.24 | 7.42 | 52.6 | 65.4 |
| BDSD | 15.26 | 0.16 | 8.64 | 50.8 | 69.1 |
| PRACS | 11.92 | 0.19 | 9.36 | 52.6 | 70.9 |
| MTF-GLP | 13.61 | 0.27 | 8.47 | 53.4 | 73.5 |
| Indusion | 14.85 | 0.34 | 8.69 | 52.9 | 72.4 |
| GND | 10.46 | 0.49 | 7.06 | 51.7 | 68.6 |
| SDNet | 9.12 | 0.53 | 6.54 | 53.2 | 79.6 |
| FHM | 12.56 | 0.27 | 9.21 | 52.8 | 80.1 |
| DeFusion | 7.84 | 0.49 | 7.32 | 54.1 | 79.2 |
| DIDFuse | 10.27 | 0.52 | 5.46 | 52.5 | 81.6 |
| CDDFuse | 10.69 | 0.65 | 4.95 | 55.0 | 82.3 |
| Super Fusion | 8.37 | 0.48 | 5.07 | 56.7 | 81.9 |
| Ours | 7.05↑ | 0.71↑ | 5.95↓ | 57.6↑ | 83.3↑ |

deep-learning-based algorithms (SDNet,[26] FHM,[27] DeFusion,[28] DIDFuse,[29] CDDFuse,[30] and Super Fusion[31]).

## 5. Conclusions

We introduced the iAFF-FMA framework, which harnesses the Mamba model to amalgamate a spectrum of power grid survey data, encompassing panchromatic and multispectral imagery, digital elevation models, and geotechnical–hydrometeorological datasets. This innovative approach addresses the shortcomings inherent in conventional fusion methodologies tailored for grid survey applications.

The ablation studies conducted underscore the indispensable function of the iAFF module within the fusion process, mitigating disparities between modalities and augmenting feature congruence. Comparative analyses demonstrate the iAFF-FMA framework's superior performance, marked by enhanced metric outcomes and a more nuanced understanding of the environment. While these strides are commendable, there exists a spectrum of opportunities for further refinement. This includes the exploration of more efficient feature extraction techniques to escalate algorithmic efficacy and the contemplation of a broader array of data and scenarios to bolster the model's resilience.

Looking ahead, future endeavors may delve into the model's efficacy across a variety of power grid settings and the investigation of the potential application of the iAFF-FMA framework to disparate power grid data fusion. Moreover, we aim to scrutinize the model's responsiveness and constraints when confronted with unforeseen data patterns, thereby enriching our understanding of its adaptability and the scope of its enhancement.

## Acknowledgments

This study was supported by the Science and Technology Project of State Grid Corporation of China: Research and Application of Multi-source Collaborative and Dynamic Data Sharing Technology for Power Grid Engineering Survey Data (5700-202356317A- 1-1-ZN).

## References

1 J. Wan, X. Wan, L. Sun, H. Sheng, S. Liu, B. Zou, and Q. Wang: J. Oceanol. Limnol. **41** (2023) 865. https://doi.org/10.1007/s00343-022-1324-x.
2 K. Liu, G. Feng, X. Jiang, W. Zhou, Z. Tian, R. Zhao, and K. Bi: Sustainability **15** (2023) 9405. https://doi.org/10.3390/su15129405.
3 Z. Wang, X. Gao, J. Yang, Q. Yan, and Y. Zhang: Multimedia Syst. **28** (2023) 1. https://doi.org/10.1007/s00530-022-00906-w.
4 C. He, P. Xu, X. Pei, Q. Wang, Y. Yue, and C. Han: Accid. Anal.Prevent. **199** (2024) 107511. https://doi.org/10.1016/j.aap.2024.107511.
5 N. Zhang, H. Wu, H. Zhu, Y. Deng, and X. Han: Agriculture **12** (2022) 2014. https://doi.org/10.3390/agriculture12122014.
6 R. Das and T. Singh: ACM Trans. Asian Low-Resour. Lang. Inf. Process. **22** (2023) 1. https://doi.org/10.1145/3584861.
7 W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. Morency, and S. Poria: Proc. 2021 Int. Conf. Multimodal Interaction. (ICMI,2021) 6–15.
8 B. Yang, L. Wu, J. Zhu, b. Shao, X. Lin, and T. Liu: IEEE/ACM Trans. Audio, Speech, Language Process **30** (2022) 2015. https://doi.org/10.1109/TASLP.2022.3178204.
9 Y. Li, T. Qi, Z. Ma, D. Quan, and Q. Miao: IEEE Trans. Neural Networks Learn. Syst. **11** (2023) 1. https://doi.org/10.1109/TNNLS.2023.3295811.
10 H. Zhang, J. Koh, J. Baldridge, H. Lee, and Y. Yang: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (CVPR 2021) 833–842
11 X. Zhang, L. He, J. Chen, B. Wang, Y. Wang, and Y. Zhou: Sensors **23** (2023) 8732. https://doi.org/10.3390/s23218732.
12 Y. Tian, Y. Zhang, Y. Fu, and C. Liu: Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition. (IEEE,2020) 3360–3369.
13 F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge: Remote Sens. **12** (2020) 2207. https://doi.org/10.3390/rs12142207.
14 P. Chen, Y. Zhang, Y. Cheng, Y. Shu, Y. Wang, Q. Wen, B. Yang, and C. Guo: 12th Int. Conf. Learning Representations. (ICLR 2024) 1–10.
15 H. Li, X. Wu, and T. Durrani: IEEE Tran. Instrument. Measure. **69** (2020) 9645. https://doi.org/10.1109/TIM.2020.3005230.
16 J. Ruan, J. Li, and S. Xiang: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2024) 158–167.
17 Y. Liu, X. Chen, R. Ward, and Z. Wang: IEEE Signal Process Lett. **26** (2019) 485. https://doi.org/10.1109/LSP.2019.2895749.
18 Z. Li, H. Pan, K. Zhang, Y. Wang, and F. Yu: Proceedings of the IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR 2024) 81–85.
19 X. Xie, Y. Cui, C. Leong, T. Tam, X. Zhang, X. Zheng, and Z. Yu: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (CVPR 2024) 94-98.
20 W. Carper, T. Lillesand, and R. Kiefer: Photogramm. Eng. Remote Sens. **56** (1990) 457. https://doi.org/10.48550/arXiv.2404.09498
21 J. Choi, K. Yu, and Y. Kim: IEEE Trans. Geosci. Remote Sens. **49** (2011) 295. https://doi.org/10.1109/TGRS.2010.2051674
22 A. Garzelli, F. Nencini, and L. Capobianco: IEEE Trans. Geosci. Remote Sens. **46** (2008) 228. https://doi.org/10.1109/TGRS.2007.907604
23 L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce: IEEE Trans. Geosci. Remote Sens. **45** (2007) 3012. https://doi.org/10.1109/TGRS.2007.904923

24 M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert: IEEE Trans. Geosci. Remote Sens. **5** (2008) 98. https://doi.org/10.1109/LGRS.2007.909934
25 F. Ye, Y. Guo, and P. Zhuang: Signal Process. Image Commun. **74** (2019) 322. https://doi.org/10.1016/j.image.2019.03.004
26 H. Zhang and J. Ma: Int. J. Comput. Vision. **129** (2021) 2761. https://doi.org/10.1007/s11263-021-01501-8
27 R. Safe'i, R. Andrian, T. Maryono, S. Tapasya, and F. Gandadipoera: IOP Conf. Ser. **1352** (2024) 12049. https://doi.org/10.1088/1755-1315/1352/1/012049.
28 Y. Sun, B. Cao, P. Zhu, and Q. Hu: Proceedings of the 30th ACM Int. Conf. Multimedia. (ACM, 2022) 4003–4011.
29 Z. Zhao, S. Xu, C. Zhang, J. Liu,J. Zhang, and P. Li: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (CVPR 2020) 970–976.
30 Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. C: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. (CVPR 2023) 5906–5916.
31 L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma: IEEE/CAA J. Autom. Sin. **9** (2022) 2121. https://doi.org/10.1109/JAS.2022.106082.