Peer-to-Peer Networking and Applications (2025) 18:262
https://doi.org/10.1007/s12083-025-02079-5

RESEARCH

AE-MCDD: Attention-enhanced multiple component defects detection for UAV-assisted powerline inspection

Jiehao Li[1] · Manjia Liu[2] · Haitao Peng[3] · Longlong Liu[1] · Xiaomin Zheng[1] · Chen Yi[2] · Guozi Liu[2] · Jieyu Zhou[3] · Feng Lyu[3]

Received: 30 April 2025 / Accepted: 14 July 2025

Abstract
Transmission lines are critical infrastructure for power delivery, yet their exposure to harsh environmental conditions accelerates component degradation, leading to defects that threaten grid reliability. While UAV-assisted inspections enable efficient defect identification, existing automated detection models face persistent challenges including severe class imbalance and complex environmental interference. To address these challenges, we first analyze a real-world 5,300-image aerial dataset, demonstrating severe class imbalance and diverse environmental noise in detail. We then propose a three-pronged solution: (1) a data augmentation pipeline integrating random occlusion and mirroring to enhance rare defect samples; (2) the Attention-Enhanced Multiple Component Defects Detection (AE-MCDD) model, combining HgNetV2 for local feature extraction, a Hybrid Attention Transformer (HAT) module for global context modeling, and C2F modules with skip connections for multi-scale feature fusion; and (3) a focal-loss-optimized multi-task loss function to handle class imbalance. Extensive experiments on our real-world dataset demonstrate that the proposed AE-MCDD model achieves a 0.719 mAP_{50}, outperforming baseline methods in both common and rare defect detection.

Keywords Power line inspection · UAV-assisted inspection · Intelligent inspection · Defect detection

1 Introduction

Transmission lines are critical infrastructure in the power industry, serving as the backbone for electricity delivery. Their operational integrity directly influences the safety and stability of power grids [1]. However, these lines are frequently subjected to harsh environmental conditions—including extreme temperatures, humidity, high winds, sandstorms, and lightning strikes—which accelerate component degradation. Such conditions can lead to defects such as insulation aging, corrosion, mechanical loosening, and structural cracks [2]. Without timely detection and intervention, these defects may escalate into equipment failures or trigger cascading grid outages [3]. Consequently, the identification of defects in transmission line components is imperative to ensure the reliable and secure operation of power systems.

Conventional manual inspection methods require labor-intensive operations, exposing workers to substantial safety hazards in extreme outdoor environments. The deployment of Unmanned Aerial Vehicle (UAVs) equipped with intelligent models enables automated and efficient edge intelligent computing [4–6], such as defect identification in transmission lines by processing images captured during power line inspection [7].

However, significant challenges persist in the development of intelligent models for UAV-based multiple component defect detection [8], necessitating further research and technological advancements. First, the dataset exhibits severe class imbalance, where frequently deployed hardware elements (e.g., binding wires, insulator pins) dominate the sample distribution, while critical but uncommon components (e.g., insulation covers, vibration dampers) suffer from acute data scarcity. Second, the extreme operating environments

✉ Jieyu Zhou
zhoujieyu@csu.edu.cn

1 State Grid Hubei Transmission & Transformation Engineering CO, LTD, Wuhan 430050, Hubei, China

2 State Grid Hubei Electric Power Research Institute, Wuhan 430077, Hubei, China

3 School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China

Published online: 09 August 2025

compound detection complexity, as most transmission infrastructure spans mountainous terrain where dense vegetative backgrounds and occlusions substantially impair computer vision performance. Third, the collaborative recognition of multi-defect targets introduces new challenges for accurate model detection.

In recent years, while significant progress has been made in UAV-based transmission line defect detection technology, critical challenges such as sample imbalance and complex environmental interference persist. Existing research primarily focuses on optimizing the detection of single defect types [9]: For insulator defect detection, Liu et al. proposed the CIA-YOLO model to effectively address cross-weather domain discrepancies [10], while Sun et al. achieved precise localization of burst defects using ISNet [11]. Regarding damper defects, Huang et al. employed spatial relationship analysis [12], and Zhang et al. improved detection via DSA-Net [13]. However, these methods fail to consider the practical demand for multi-defect collaborative recognition. In data augmentation, Zhang et al.'s HC-ViT model enhanced feature representation using unlabeled samples, but its application was limited to transmission lines with relatively simple backgrounds [14]. Additionally, Yu et al. [15] and Yang et al. [16] developed lightweight adaptations of YOLO-series models, achieving a balance between inference speed and detection accuracy. Thus, detecting multiple defects in power line images—with imbalanced data and complex backgrounds—remains an unsolved challenge.

To address these challenges, we investigate data augmentation techniques and multi-target detection models. First, we conduct an extensive quantitative analysis of a real-world dataset comprising 5,300 high-resolution aerial images captured under various environmental conditions, containing 17,413 annotated defect instances across 19 distinct defect categories. This analysis reveal significant class imbalance issues, with common defects like Irregular Binding Wire appearing in about 42% of instances while critical but rare defects such as Longitudinal and Transverse Crack represented less than 1% of instances. Second, to mitigate this data scarcity problem, we implement a comprehensive data augmentation pipeline incorporating multiple complementary strategies including random occlusion, mirroring, etc. Third, we design and implement the novel attention-enhanced defect detection (AE-MCDD) model, which features a sophisticated architecture combining: (1) a backbone design integrating HgNetV2 for local feature extraction with a HAT module for global context modeling, (2) the cascaded C2F (Cross-stage Feature Fusion) modules in the neck network with skip connections to enhance multi-scale feature representation, and (3) a multi-task detection head with separate prediction branches for defect detection and bounding box regression, optimized through focal loss for imbalanced data. Extensive experimental validation on our real-world dataset demonstrate that the proposed AE-MCDD model achieves superior performance metrics.

The key contributions of this work can be summarized as follows:

- We analyzed the sample distribution in a real-world power line inspection dataset containing numerous images, identified severe sample imbalance issues, and employed a series of data augmentation algorithms to address this problem.
- We designed an attention-enhanced multiple component defects detection model comprising three key components - backbone, neck, and detection head - to address the challenge of multi-defect collaborative recognition in complex backgrounds.
- We implemented the proposed model and conducted extensive experimental evaluations on real inspection datasets, comparing our approach with multiple baseline methods across various defect types to demonstrate the superior performance of our model in terms of detection accuracy and robustness.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents data analysis and motivation. Section 4 provides an overview of AE-MCDD. Section 5 details the design of AE-MCDD. Section 6 presents performance evaluations, and Section 7 concludes this paper.

2 Related works

2.1 General-purpose object detection frameworks

Object detection, which involves the localization and classification of target objects in images or video streams, has undergone remarkable progress over the past decade. From the early developments of R-CNN [17], Faster R-CNN [18], and FPN [19], to the emergence of the YOLO series [20], detection accuracy and computational efficiency have significantly improved. Based on the internal detection strategy of the architecture, object detection methods can be broadly classified into three categories: anchor-based, anchor-free, and Transformer-based approaches.

Anchor-based detectors generate candidate regions using predefined anchor boxes and then perform classification and regression tasks. For instance, SSD [21] adopts multi-scale default boxes across different output layers and performs predictions on each feature map. RetinaNet [22] further enhances this paradigm by introducing a focal loss function to address class imbalance, leading to improved detection

performance. In contrast, anchor-free detectors eliminate the use of predefined anchor boxes. These models predict object locations by identifying keypoints (e.g., corners or center points) on a regular grid, followed by estimating object dimensions. Representative models include CornerNet [23], which detects objects by regressing their corner points, and FCOS [24], which directly predicts the center and size of the target object. Transformer-based methods have emerged as a new frontier in object detection. These models, exemplified by DETR [25], utilize the Transformer's encoder-decoder framework to model global dependencies via multi-head attention mechanisms and feed-forward networks, offering a fully end-to-end detection pipeline.

2.2 Customized detection strategies for UAV scenarios

With the increasing use of UAVs in computer vision applications [26, 27], object detection and tracking have entered a new stage of development. Drones offer exceptional mobility and flexibility, enabling geographically unrestricted and timely data collection [28, 29]. However, applying traditional object detection methods to UAV-based imagery introduces unique challenges due to constraints such as low-altitude imaging, limited payload, and the complexity of outdoor environments.Specifically, UAV platforms often encounter non-standard viewing angles, intricate backgrounds, variable flight heights, and significant scale and orientation variations. These conditions contribute to frequent object occlusion, overlap, and deformation, making accurate detection particularly difficult—especially for small objects [30].

Recently, various enhancements have been proposed. Xia et al. [31] provided a systematic overview of the obstacles associated with UAV optical imaging and proposed tailored data processing strategies. Chen et al. [32] introduced ResNeXt-d, an extension of ResNeXt [33] integrated with dilated convolutions to effectively enlarge the receptive field and improve multi-scale object detection. Focusing on small object perception, Li et al. [34] proposed a perceptual GAN that leverages the structural correlation between large and small objects to generate super-resolved representations for enhancing small object features. Similarly, Hu et al. [35] identified the distortion effects caused by traditional pooling operations and proposed a context-aware ROI pooling method to maintain spatial integrity and improve feature representation for small targets. In addition, further enhancements to ResNeXt-d [33] involving multiscale feature fusion have demonstrated improved sensitivity and robustness in detecting fine-grained details, particularly in UAV-based aerial imagery.

In summary, while state-of-the-art object detection models continue to advance, effectively deploying these methods in UAV remote sensing scenarios requires adaptation to context-specific limitations. Addressing issues such as small object detection, scale variation, and complex backgrounds remains a central research focus. The convergence of deep learning and UAV technology presents promising opportunities, calling for continued innovation in model architecture, data processing, and task-specific optimization.

3 Data analysis and motivation

3.1 Data description

In this study, we establish a comprehensive benchmark dataset specifically designed for UAV-assisted power line defect detection. The dataset was systematically collected during routine inspection operations conducted from October 15 to December 30, 2023, covering approximately 30.5 kilometers of 10kV transmission lines in a challenging mountainous region of Hubei Province, China. Our raw dataset originally comprises 8,837 aerial images of 10kV utility poles, captured using UAVs over approximately 30.5 kilometers of transmission lines, including 5 main circuits and 21 branch lines. These images were collected under real-world conditions and include some unannotated or defect-free samples. After filtering and annotation, we curated a high-quality dataset of approximately 5,300 images with a resolution of 4000×3000 pixels, which were used for model training and evaluation. The dataset contains annotations for 19 typical overhead transmission line defects, as illustrated in Table 1, such as Tower Top Damage, Irregular Binding Wire, Flange Rod Rust, etc, comprehensively covering common defect types encountered during inspections. With a total of 17,413 precisely annotated defect instances, the dataset provides substantial sample support for subsequent algorithm development and evaluation.

3.2 Data analysis and challenges

Complex environmental background Power transmission lines are predominantly deployed in remote mountainous regions where image acquisition faces significant challenges due to complex background interference, including highly variable textures (e.g., dense vegetation and rugged terrain), dynamic lighting conditions (e.g., shadows and glare), and subtle defect characteristics such as small-scale missing bolts or visually similar mild corrosion. As shown in Fig. 1, three representative defects are particularly challenging: TTD, manifested as structural degradation at power tower summits (Fig. 1(a)) that compromises load-bearing capacity; IBW, characterized by non-standard conductor-insulator connections (Fig. 1(b)) that may induce conductor slippage; FRR caused by prolonged weather exposure (Fig. 1(c)) that deteriorates protective coatings. These defects frequently appear against cluttered vegetation backgrounds featuring

Table 1 Defect and image counts

Defect	Defect Count	Image Count
Irregular Binding Wire (IBW)	7299	3462
Insurance Pin Falling off (IPF)	3671	1425
Insulator Cover Falling off (ICF)	2765	929
Tower Top Damage (TTD)	1089	1043
Missing Insulator Cover (MIC)	688	299
Insulator Contamination (IC)	652	284
Clamp Rust (CR)	270	109
Loose Fixing (LF)	230	210
Flange Rod Rust (FRR)	159	152
Rust Severe On Connector Fitting Ball (RSC)	138	75
Insulator Damage (ID)	121	124
Longitudinal and Transverse Cracks (LTC)	112	109
Insulator Sleeve Damage (ISD)	92	70
Foreign Objects and Nests (FON)	64	64
Conductor Not Bound (CNB)	31	26
Debris Accumulation at Tower Base (DATB)	19	19
Crossarm Bending, Tilt, Deformation (CBTD)	10	10
Vacuum Switch bushing Missing insulation cover (VSM)	2	2
Foreign Objects on Conductor (FOC)	1	1

intricate textures that mimic defects, geometric occlusions from overlapping branches, non-uniform illumination-all of which significantly increase the complexity of automated defect detection in real-world inspection scenarios.

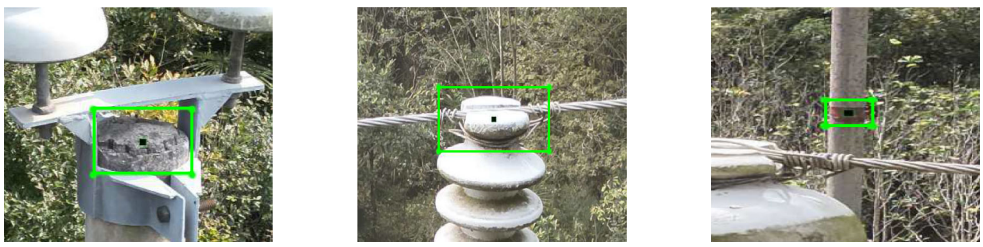
Imbalance defect samples In Fig. 2, we conducted statistical analysis on both the image count (total number of images containing each defect category) and instance count (total occurrences of each defect in the dataset) across different defect categories. The results reveal that while a positive correlation exists between image and instance counts, both exhibit a pronounced long-tail distribution pattern. This distribution stems from two primary factors: (1) Power component usage frequency variation - Significant differences exist in component deployment frequency across transmission lines (e.g., bolts and insulator strings are repeatedly used throughout the line, while tower poles appear only once per UAV photography point); (2) Power component failure probability variation - Distinct differences in damage susceptibility among components (e.g., tower poles, despite single-point installation, show higher defect frequency due to corrosion and environmental exposure, whereas crossarm, despite multiple installations, demonstrate lower defect rates due to structural robustness). This data imbalance poses substantial challenges for model training [36, 37], as models tend to overfocus on high-frequency defects, leading to degraded recognition performance for low-frequency defects compared to high-frequency ones.

Moreover, during actual inspections, UAVs must maintain a safe distance from power facilities, which introduces two key challenges: the safety distance between the UAV and power components reduces the relative pixel size of small defects such as missing bolts, while varying shooting distances further lead to significant scale variations of the same defect across images. These factors substantially increase small-target detection difficulty and impose stricter robustness demands on detection models.

4 Overview of AE-MCDD

To address the aforementioned challenges, we proposed the AE-MCDD, whose workflow is presented in Fig. 3, encompassing data collection, data augmentation, and HAT-enhanced defect inference. First of all, during data collection phase, we need to deploy UAV to the vicinity of power towers to capture aerial images, after which a subset of these images are manually annotated by human experts to identify defects.

Subsequently, in the data augmentation phase, we apply a series of random transformations techniques to the original images to extend the quantity and diversity of the collected



(a) Tower Top Damage (b) Irregular Binding Wire (c) Flange Rod Rust

Fig. 1 Illustration of the defect type

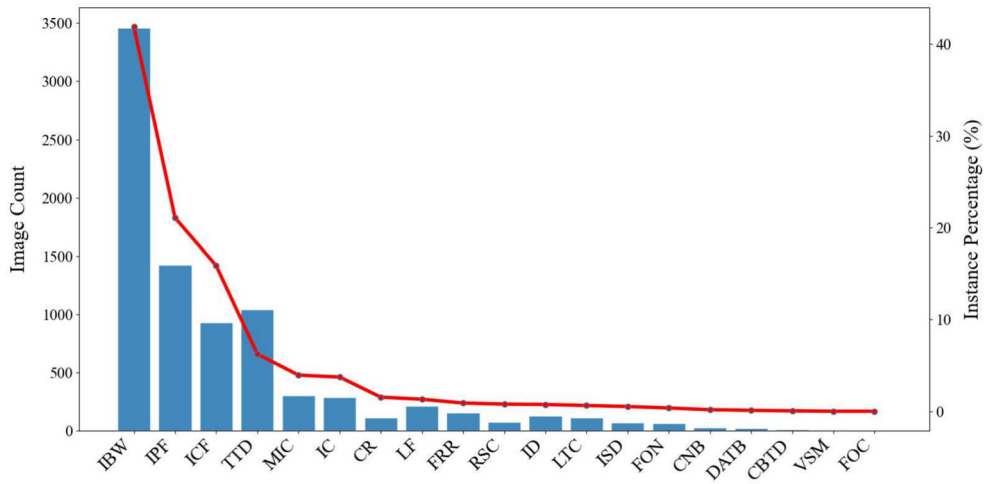


Fig. 2 Category distribution of defect samples. The bar chart shows the number of images containing each defect category, while the red line indicates the corresponding proportion of instances in the dataset

image dataset [38]. These transformations include mirroring, random occlusion, adjusting brightness, and modifying channel histograms (CH) [37]. By leveraging these image augmentation techniques, the issue of imbalanced images samples is effectively mitigated, and the model's feature extraction capabilities are significantly improved.

Finally, we implemented a defect reasoning model based on HAT, which improves the overall performance by inheriting multiple advanced modules while maintaining the real-time detection advantage. Specifically, the model first introduces the HAT module in the input stage to activate and enhance key information; then uses HgnetV2 [39] as the backbone, and realizes the deep abstraction of multi-scale features through HGStem and multiple HGBlock modules; in the neck part, the C2F module is used to efficiently integrate features at different levels; lastly, the detection head completes target positioning and classification through a multi-layer structure. Overall, the model not only focuses on feature enhancement and fusion, but also takes into account the trade-off between detection efficiency and computing resource conflicts [40], providing an efficient and accurate solution for defect detection tasks in power line inspections.

5 Design of AE-MCDD

5.1 Data enhancement

In multiple component defect detection for power line components, there is often a significant class imbalance due to

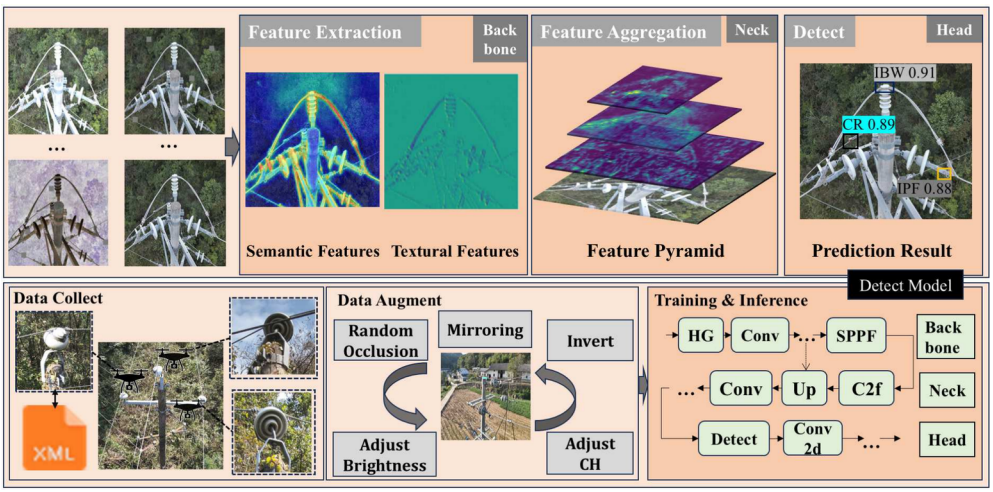


Fig. 3 Design overview of AE-MCDD

the disparity in the number of different components, such as insulator pins and transmission towers. Additionally, the high costs of data collection and annotation typically result in relatively small datasets for inspection images [41]. To overcome these challenges, data augmentation techniques are employed to increase sample diversity and expand the dataset size [42, 43]. The applied augmentations include adjusting CH, mirroring, random occlusion, adjusting brightness and inverting, as shown in Fig. 4.

Adjust CH Adjusting CH involves reordering or modifying the number of color channels in an image to improve the model's adaptability to different data distributions, as shown in Fig. 4b. This transformation can make the target segmentation more obvious and achieve edge sharpening, thus enhancing the model's color sensitivity in the feature extraction process. The transformation is defined as:

$$I_{\text{out}} = \text{Interp}\left(\frac{\sum_{k=0}^{I_{\text{in}}} H'_{i,j}(k)}{\sum_{k=0}^{L-1} H'_{i,j}(k)} \times (L-1)\right) \tag{1}$$

where $H'_{i,j}(k) = \min(H_{i,j}(k), C)$, clipped histogram at grid cell (i, j). $H_{i,j}(k)$ represents the original histogram in the local region, C represents the clip limit for contrast normalization, and L represents the number of intensity levels (e.g., 256 for an 8-bit image).

Random occlusion Random occlusion simulates real-world scenarios where objects may be partially obstructed by other elements in the scene to help the model to learn features that are invariant to partial obstructions, as depicted in Fig. 4c. This is typically implemented by masking random rectangular regions with a constant value (e.g., black or mean intensity). The transformation can be described as:

$$I_{\text{out}}(x, y) = \begin{cases} I_{\text{in}}(x, y), & \text{if } (x, y) \notin M \\ V, & \text{if } (x, y) \in M \end{cases} \tag{2}$$

where M represents the occlusion mask region, and V represents the occlusion pixel value.

Adjust brightness Adjusting brightness scales the pixel intensities uniformly to simulate diverse lighting conditions, ensuring model robustness under varying illumination, as shown in Fig. 4d. By modifying brightness levels, this augmentation helps the model generalize to real-world scenarios where lighting may be inconsistent, such as shadows, overexposure, or low-light environments. The transformation is given by:

$$I_{\text{out}} = M \cdot \exp\left(\gamma \log \frac{I_{\text{in}}}{M}\right) \tag{3}$$

where $I_{\text{in}}, I_{\text{out}} \in [0, M]$, intensity values in the image, M represents the maximum pixel value, γ is the gamma correction factor.

Mirroring Mirroring applies a spatial transformation by flipping the image along the vertical axis (left-right axis) or the horizontal axis (up-down axis), as shown in Fig. 4e. This transformation enables the model to learn invariant representations across different viewpoints, expand the dataset, and enhance the generalization and robustness of the model. The transformation is expressed as:

$$I_{\text{out}} = I_{\text{in}} \cdot P_x \tag{4}$$

(a) original (b) Adjust CH (c) Random Occlusion

(d) Adjust Brightness (e) Mirroring (f) Invert

Fig. 4 Data augmentation method

where P_x represents a $W \times W$ horizontal flip matrix:

$$P_x = \begin{bmatrix} 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}$$

This matrix reverses the order of elements when multiplied by a W-dimensional vector. I_{in} and I_{out} are $H \times W$ matrices representing input and output images.

Invert Inverting reverses the intensity of each pixel, which helps to highlight specific structural features and enhance contrast under low visibility conditions, making the image a negative film, enhancing the feature extraction ability of the model, as shown in Fig. 4f. The transformation is defined as:

$$I_{\text{out}} = M \cdot \left(1 - \frac{I_{\text{in}}}{M}\right) \tag{5}$$

where $I_{\text{in}}, I_{\text{out}} \in [0, M]$, representing the intensity values in the image, M is the maximum pixel value.

In general, the comprehensive enhancement strategy adopted in our study involves applying one to three randomly selected transformations to an original image, thereby artificially expanding the dataset and significantly increasing the diversity of samples in the dataset without losing key feature information. Figure 4 shows the successful application of our proposed image enhancement scheme [44]. This methodical enhancement scheme not only enriches the dataset but also improves the generalization ability of the model, ultimately resulting in a more robust and accurate defect recognition system.

5.2 Defect detection inference

Our proposed defect detection model incorporates the advanced feature extraction module and the fusion mechanisms for inspections of power infrastructure. The architecture consists of three main components: the backbone, neck and head, optimized for efficient and accurate defect identification with the aim of contributing to a refined multiscale feature extraction, fusion and detection accuracy that enhances defect detection across varying conditions [45, 46], as Fig. 5 shown.

5.2.1 Backbone

The backbone serves as the feature extraction module, consisting of an Attention-Enhanced Stem followed by multiple HGBlocks organized into our hierarchical stages. The Stem module first processes the input image through convolutional layers, extracting initial feature representations. The HAT module refines these features by emphasizing critical spatial and channel-wise information, allowing the model to better capture defect-relevant details, particularly in challenging scenarios such as occlusion or low contrast. As the features propagate through the backbone, each HGBlock applies depth-wise convolutions and lightweight transformations to maintain computational efficiency while extracting deep semantic features. The multistage structure ensures that the network produces feature maps at different scales, each capturing defects at various levels of granularity. Such

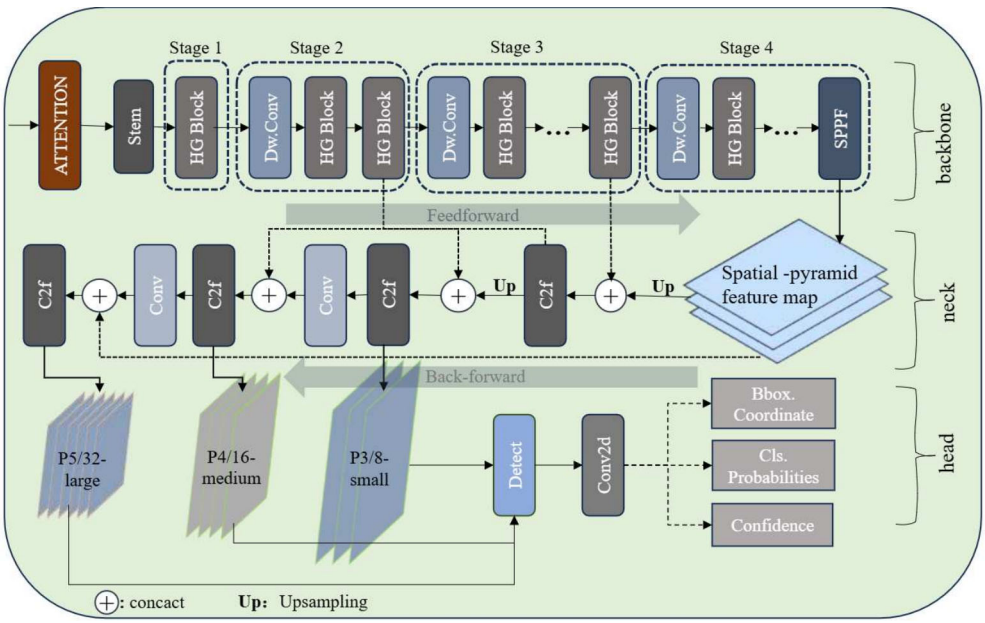


Fig. 5 The diagram of network structure

structured flow of feature extraction ensures that both fine-grained and high-level semantic information is preserved, crucial for detecting small or subtle defects.

In detailed, we integrated a HAT module into the backbone to address the inherent challenges in defect detection, particularly the need for both fine-grained feature extraction and global context awareness. Traditional convolutional architectures often struggle to capture long-range dependencies and intricate spatial relationships, which are crucial for identifying subtle anomalies in power infrastructure. By incorporating channel attention, window-based self-attention, and overlapping cross-attention mechanisms, our approach effectively refines feature representations, allowing the network to selectively emphasize defect-relevant regions while suppressing background noise.

At the core of this enhancement, the channel attention mechanism operates to recalibrate feature responses across different channels, ensuring that discriminative patterns are accentuated. This is achieved by leveraging a learnable weighting scheme applied to aggregated channel-wise activations. By adaptively adjusting feature importance, this mechanism enhances the network's ability to focus on structural characteristics indicative of defects, while reducing interference from less relevant regions, formally defined as:

$$\mathbf{C}_{\text{out}} = \sigma(\mathbf{W}_c \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{X} + \mathbf{W}_2 \cdot \mathbf{X})) \tag{6}$$

where $\mathbf{X}$ is the input feature map, and $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_c$ are learnable weights, σ denotes an activation function.

Then, a key limitation of standard self-attention mechanisms is their computational inefficiency when applied across entire feature maps, particularly in high-resolution images. To mitigate this, HAT employ a window-based self-attention mechanism, which segments the feature map into non-overlapping local windows, computing self-attention independently within each window. This partitioning significantly reduces computational complexity while ensuring that local structural details—such as texture variations, fine cracks, and material inconsistencies—are effectively captured with less input.Given an input feature map partitioned into $\mathbf{M}$ windows, attention within each local window is computed as:

$$\mathbf{A} = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V} \tag{7}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively, d_k is the dimension of the key, and $\mathbf{B}$ is the relative position bias.

Lastly, there is a extend in the self-attention mechanism by incorporating overlapping cross-attention to bridge the gap between local and global contextual understanding, allowing adjacent windows to exchange information. Instead of treating each window as an independent processing unit, this mechanism ensures continuity in feature representation across neighboring regions, preventing potential spatial fragmentation. The attention computation is defined as:

$$\mathbf{O}_{\text{CA}} = \text{SoftMax}\left(\frac{\mathbf{Q}_{\text{O}}\mathbf{K}_{\text{O}}^T}{\sqrt{d_k}} + \mathbf{B}_{\text{O}}\right)\mathbf{V}_{\text{O}} \tag{8}$$

where $\mathbf{Q}_{\text{O}}$, $\mathbf{K}_{\text{O}}$, and $\mathbf{V}_{\text{O}}$ are the query, key, and value matrices from overlapping windows, and $\mathbf{B}_{\text{O}}$ is the relative position bias for overlapping windows.

In order to more intuitively demonstrate the benefits this mechanism brings to the model. We visualize the different performance of the model for feature extraction with and without HAT. As shown in the Fig. 6, when there is HAT module, the model pays more attention to the details of the tower body and components, and the tower and the background are more clearly separated.

In a word, channel attention emphasizes important features across channels, while window-based self-attention captures local patterns within image segments [47]. The intermediary attention module bridges the gap by promoting interactions between adjacent windows, effectively integrating local and global information [48]. This hybrid approach enables HAT to achieve superior performance in various image restoration applications by leveraging the complementary strengths of these attention mechanisms.

5.2.2 Neck

The neck adopts a C2F-based fusion mechanism, which helps to efficiently aggregate features across multiple scales. This component uses upsampling and convolutional layers to merge the outputs from different backbone stages, such as the feature maps of the three scales s, m and l, combining spatial pyramid enhanced feature maps with multiscale representations. The fusion of fine details in early layers and semantic features in deeper layers ensures that defects of different sizes are effectively highlighted, thereby enhancing the robustness of the model in actual inspection tasks.

5.2.3 Head

The part of detection head processes the three multiscale feature maps from the neck, ensuring precise defect localization and classification at different resolutions. At each scale, a single detection head is used. Each head contains Conv2D layer, responsible for refining features before passing them to the final prediction layers, which handle bounding box regression, confidence estimation, and classification probability

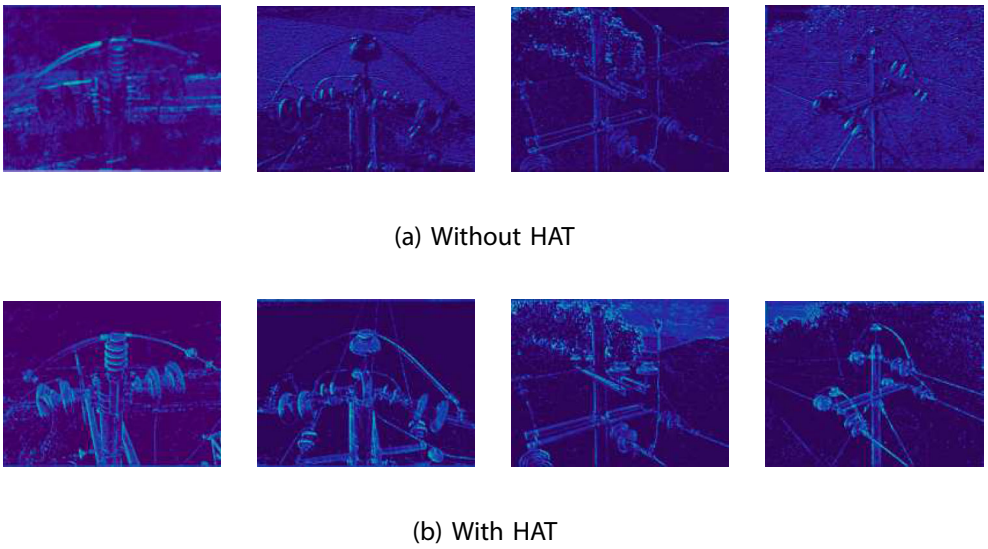


Fig. 6 Feature map comparison. The first row displays the feature maps extracted by the baseline backbone without HAT modules, while the second row shows the corresponding feature maps obtained using the HAT-enhanced backbone for the same set of input images

prediction. In detailed, the small-scale feature map focuses on detecting fine-grained defects such as cracks and insulation wear, while the medium- and large-scale feature maps are optimized for identifying larger structural defects. To enhance detection accuracy, the final feature maps undergo additional Conv2D processing, which refines spatial consistency and feature resolution. This structured multiscale detection approach ensures a balance between computational efficiency and accurate defect identification, making it well suited for power infrastructure defect [45] inspections.

The overall design of the model facilitates a smooth transition from feature extraction to detection. By integrating attention-driven feature enhancement, hierarchical multiscale extraction, and adaptive feature fusion, this model effectively captures and processes defect-related information across different scales. The synergy between the backbone, neck, and head ensures a balanced trade-off between computational efficiency and detection accuracy, making the model highly suitable for UAV-based power infrastructure inspections.

5.3 Loss function

Additionally, we adopt Unified-IoU-Focal-inv (UIoU-Focal-inv) loss to replace the conventional CIoU loss in bounding box regression. CIoU enhances localization by incorporating geometric factors such as overlap area, center distance, and aspect ratio into the loss function. However, it treats all predictions equally regardless of their confidence scores, which may lead to suboptimal gradient allocation particularly in dense scenes or tasks requiring precise localization. Compared to CIoU, UIoU-Focal-inv enhances the detection quality not by adding geometric penalties but by dynamically adjusting the importance of samples during training based on both their localization quality and prediction confidence. This leads to a better trade-off between convergence speed and detection accuracy, especially under high IoU thresholds. Formally, the formula for Unified IoU Focal Loss-inv is given by:

$$L_{\text{UIoU-Focal-inv}} = L_{\text{IoU-based}}\left(\hat{P}, \hat{G}\right) \times c^{\gamma}, \tag{9}$$

where c is the confidence score, γ is a hyperparameter, and ratio can be adjusted (e.g., by a cosine schedule) to shift the model's focus from lower-quality to higher-quality boxes.

For clarity, Fig. 7 provides a schematic comparison between standard Focal Loss, and the inverse Focal weighting strategy in UIoU-Focal-inv, illustrating how each method distributes gradient focus across predictions with varying confidence levels. The core strategy of standard Focal Loss lies in enhancing the optimization weight for low-confidence samples. This design is based on the fundamental assumption that low-confidence predictions often correspond to challenging samples, and increasing their learning weight helps improve the model's discriminative capabilities. However, in scenarios with highly dense objects and severe bounding box overlap, this strategy encounters an inherent paradox: low-confidence prediction boxes are predominantly generated by multi-object overlapping (e.g., invalid detection boxes covering multiple insulators), where IoU values approach zero and representations become ambiguous. Persistently intensifying the optimization weight on such samples essentially guides

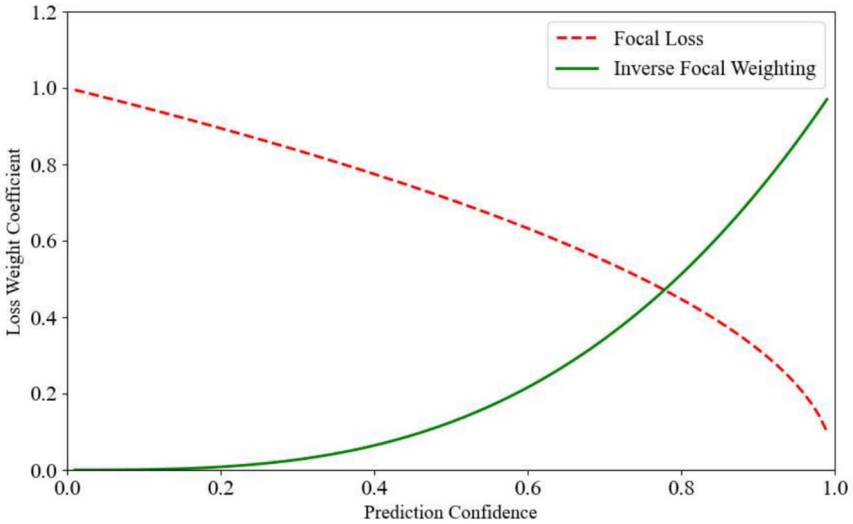


Fig. 7 Divergent weighting strategies: standard focal loss vs. inverse focal loss

the model to fit a noise distribution, amplifying false detection risks and impeding the development of high-precision localization capabilities.

In contrast, the inverse weighting strategy adopted by UIoU-Focal-inv achieves targeted breakthroughs in dense scenes through its paradigm-shifting design. This method systematically reduces the gradient weight of low-confidence samples and concentrates optimization resources on high-confidence predictions. The inherent advantage of this approach stems from capturing the underlying pattern of dense environments: in regions with extensive bounding box overlap, high-confidence detection boxes often correspond to critical samples of effectively separated target entities (e.g., results that precisely frame single defect). These samples possess scarcity and high value in densely stacked scenarios, and reinforcing their weights is equivalent to establishing an accurate localization reference system within complex spatial relationships. Thus, the mechanism demonstrates significant value in dense power equipment inspection scenarios.

6 Performance evaluation

6.1 Experimental setup

In this paper, our experiments were conducted on a server equipped with eight NVIDIA TITAN V 12GB GPUs. The software environment includes Python 3.9.7 and PyTorch 2.2.0. For the training process, we adapted the following hyperparameters: a batch size of 8, an input images resolution of 4000 x 3000 pixels, and the SGD optimizer with a learning rate of 0.01 and momentum of 0.937. Our dataset was partitioned into training (3,180 images), validation (1,060 images), and test sets (1,060 images) to ensure rigorous evaluation.

6.2 Metrics

In this work, we use several common performance metrics to evaluate our model, such as precision (P), recall (R) and F1-score, which are defined as follows:

- **Precision (Pre):** measures the accuracy of positive predictions, representing the ratio of correctly identified objects to the total number of identified objects, which is formally defined as:

$$Pre = \frac{TP}{TP + FP}, \tag{10}$$

  where TP denotes the number of correctly predicted positive samples, and FP indicates indicates the number of incorrectly predicted positive samples.
- **Recall (R):** evaluates the model's sensitivity in detecting all positive samples, defined as the ratio of correctly identified positives to all actual positives in the dataset:

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

  where TP are correctly predicted positive samples and FN are the positive samples incorrectly predicted as negative.
- **F1-score:** is the harmonic mean of the precision and recall, providing a balanced evaluation metrics for imbalanced datasets:

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{12}$$

- **CoCo-Metrics:**On COCO, the average precision is evaluated over several IoU values. We chose AP.50-95 computed in the range 0.50-0.95 of IoU thresholds and AP.75 computed in the 0.75 of IoU thresholds. Meanwhile, the target is divided into small, medium and large data according to the size or area of the bounding box of the object. The AP calculated by group is called APm, APl and APs. Similar to AP, AR (Average Recall) for instance segmentation considers both detection and segmentation. It measures how well the model recalls instances of objects at various IoU thresholds. A prediction is considered a true positive if its IoU with the ground truth exceeds a certain threshold. We chose the AR.50-95 computed in the range 0.50-0.95 of IoU thresholds.

6.3 Baselines

To benchmark our approach against existing methods, we implemented several widely-recognized baseline models for comparison. These reference models are briefly described below:

- **YOLOv8** [20]: leverages depth-wise separable convolutions and adaptive feature fusion to maintain competitive accuracy while reducing computational overhead.
- **DETR (Detection Transformer)** [25]: is a end-to-end Transformer-based object detection framework, and demonstrates remarkable versatility beyond conventional detection tasks.
- **YOLOX** [49]: establishes a state-of-the-art balance between inference latency and detection precision, representing the most recent advancement in the YOLO architectural series.
- **Real-Time Multi-Objection Detection (RTMDet)** [50]: achieves robust detection capabilities while maintaining high efficiency, specifically designed for real-time multi-object detection.
- **Yolo-SwinTransformer (YS)** [51]: is a fusion of the YOLO object detection framework and the Swin Transformer architecture. This combination enhances detection accuracy by leveraging Swin Transformer's hierarchical feature extraction and window-based attention mechanism, which improves the model's ability to capture both local and global features.

6.4 Overall performance

In this section, we first demonstrate the model's convergence, then comprehensively evaluate its performance across various defect categories and metrics compared to the aforementioned baseline models. In Fig. 8, we present the variations of the bounding box loss, classification loss, and confidence loss during the training process. The blue lines represent the loss values at each epoch, while the orange points denote the smoothed loss values. It can be observed that the bounding box loss, classification loss and confidence loss all gradually decrease as the number of epochs increases. After 130 epochs, the loss values begin to stabilize, indicating that the model has achieved good convergence.

In Fig. 9, we present several performance metrics comparison between our model and baseline models in the dataset. Our model demonstrates significant improvements across all metrics, achieving 0.719 mAP.50 (+107.2-191.3% over baselines), 0.721 F1-score (+79.6-130.4%), 0.739 precision (+57.2-126.7%), and 0.703 recall (+77.5-170.4%). These advancements stem from three key architectural innovations: 1) A dual-path feature fusion mechanism combining spatial feature map and channel-wise HAT attention (reducing false positives on tiny size), addressing YOLOX's poor recall (0.367) and DETR's precision limitations (0.470); 2) Data augmentation that resolves class imbalance issues evident in Yolo-SwinTransformer's severe underperformance (0.247 mAP.50); 3) A multi-scale feature aggregation module boosting small insulator breakage detection, particularly overcoming YOLOv8's localization errors (0.289 recall). The precision-recall tradeoff optimization (F1 gain of 112.6% over YOLOv8) validates our training strategy employing focal loss for hard samples and UIoU-aware confidence calibration.

6.5 Performance of defects

To conduct a detailed comparison of our model's performance on each defect type, we first present the mAP.50

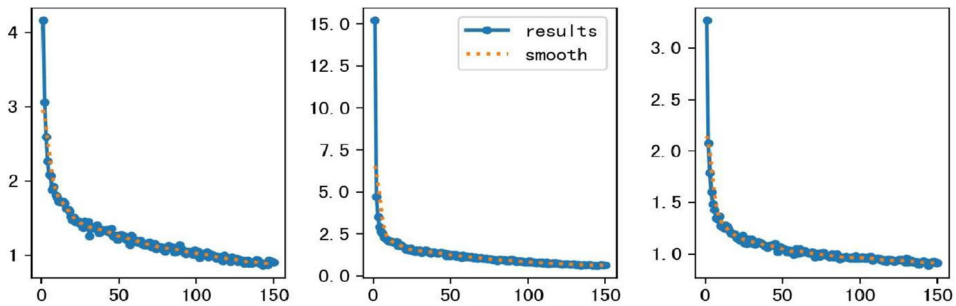


Fig. 8 Loss in training phase. The three subplots illustrate the evolution of bounding box regression loss, classification loss, and confidence loss over the course of training. All three loss curves demonstrate a consistent downward trend and stabilize, indicating effective model convergence across multiple learning targets

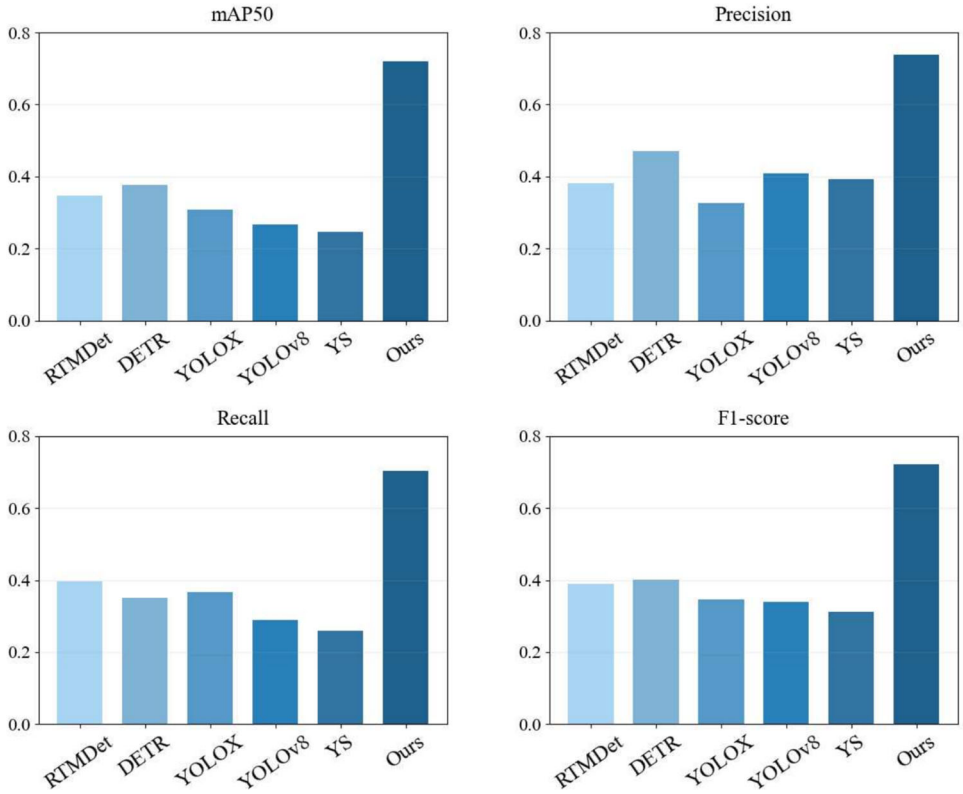


Fig. 9 Metrics comparison

comparison between our proposed model and baseline models across multiple defects, followed by demonstrating various accuracy metrics of our model for different defect categories. In Fig. 10, we selected eight important defects with significantly different sample sizes and showed the mAP.50 performance of our model versus five baseline models on these defects. The results indicate that, for defects with larger sample sizes like IBW and TDD, the model performance is generally better than for other defect types with fewer samples, especially FON. Our proposed model, along with YOLOv8 and Yolo-SwinTransformer, shows significantly better recognition performance than the other three baseline models for most defect types. Our model consistently achieves either the best or second-best defect detection accuracy. For defects with sample instances accounting for less than 2% of total data, such as LF, FRR, and FON, our model significantly outperforms other baseline models. This improvement benefits from our data augmentation approach which effectively supplements the relevant instance data.

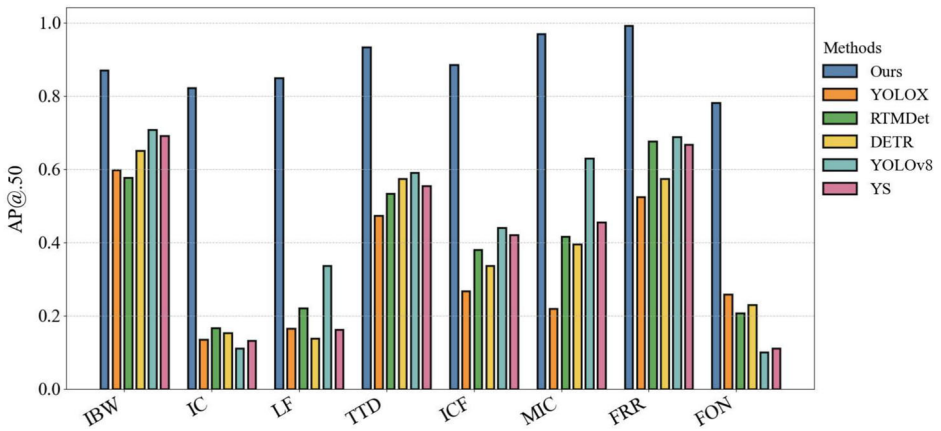


Fig. 10 Performance of each defects comparison

Table 2 Class-wise results of our model

Defect Type	AP@.50:.95	AP@.75	APl	AR@.50:.95
IBW	0.7512	0.8769	0.8562	0.3759
IPF	0.4056	0.3889	0.9000	0.1943
ICF	0.6453	0.7805	0.7398	0.2543
TTD	0.7148	0.8350	0.8356	0.7457
MIC	0.7275	0.8310	0.7658	0.3423
IC	0.5810	0.6604	0.6554	0.3179
CR	0.5685	0.6959	0.6277	0.3304
LF	0.7203	0.7964	0.8036	0.7904
FRR	0.7313	0.8692	0.7637	0.7480
RSC	0.3981	0.3990	0.5102	0.2556
LTC	0.6800	0.8337	0.6800	0.6962
ISD	0.3126	0.2650	0.5020	0.2800
FON	0.6357	0.7822	0.6357	0.6500
CNB	0.3382	0.1683	0.1515	0.3667
DATB	0.5590	0.6824	0.5590	0.5667
CBTD	1.0000	1.0000	1.0000	1.0000

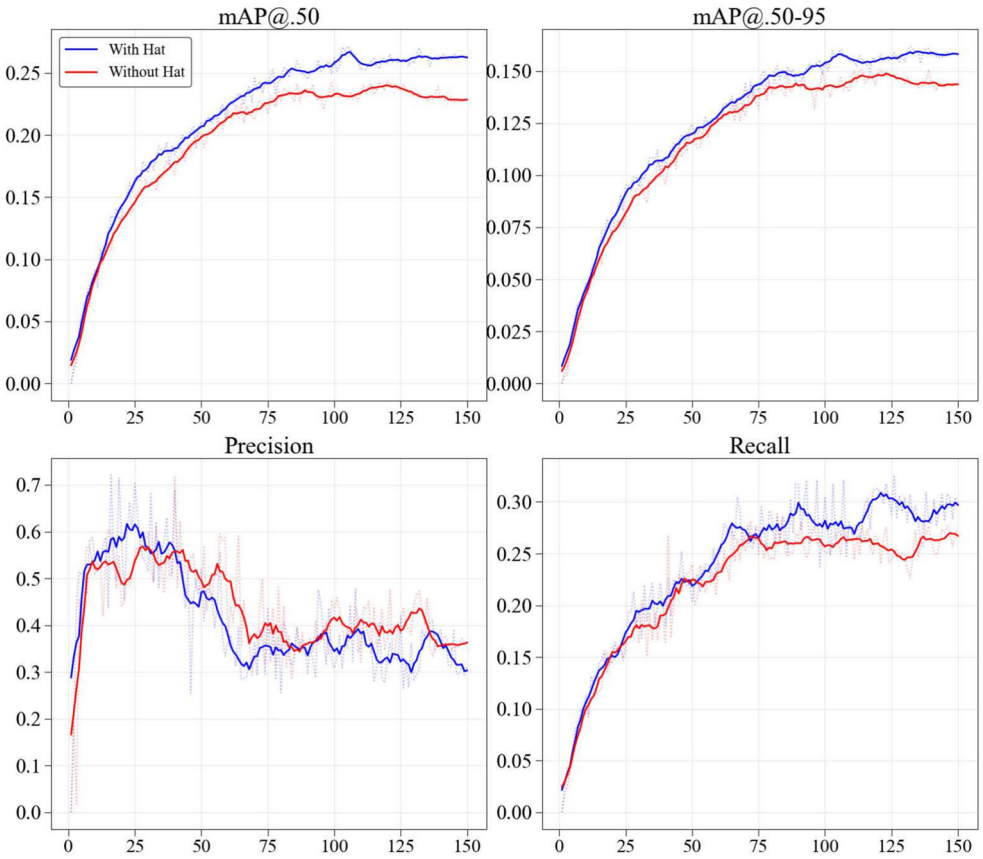


Fig. 11 Comparison vs. HAT. Each subplot demonstrates training dynamics under two configurations: *With HAT* and *Without HAT*. While the blue curves demonstrate clear advantages in mAP and recall metrics throughout the training phase, precision exhibits a trade-off in later stages. Shaded regions indicate epoch-to-epoch metric variability

For the test set as Table 2, our model continues to perform exceptionally well. For IBW, it achieves an AP of 0.751 and an AR of 0.376, indicating strong detection capabilities despite potential variations in defect presentation. In detecting TTD, the model attains an AP of 0.715 and an AR of 0.746, showcasing its effectiveness in identifying significant structural issues with high precision. For FRR, it reaches an AP of 0.731 and an AR of 0.748, demonstrating its ability to detect both subtle and severe forms of corrosion with remarkable accuracy. The model also shows robust performance in identifying LF with an AP of 0.720 and an AR of 0.790, as well as ICF with an AP of 0.645. These results indicate its versatility in detecting various types of mechanical and electrical defects, even when the defects are less pronounced or occur in challenging conditions.

Overall, our model shows remarkable consistency and excellence across different evaluation phases and defect categories. The high AP and AR scores across multiple categories in both validation and test sets demonstrate its strong generalizability and reliability in practical defect detection scenarios.

6.6 Ablation experiments

In this section, we verify the effectiveness of adding the HAT module to the Head and the data augmentation method used in this study.

By conducting an ablation study in the original dataset without data enhancement, we found obvious differences by comparing the results of the two experiments. After adding the HAT module, the overall mAP.50 and the recall rate of the model increased by approximately 2-5 percentage points. Thus, we believe that the HAT module plays a positive role in the stable convergence of the model as Fig. 11 shows.

Besides, the Fig. 12 presents a comparative analysis of detection performance using two different bounding box regression losses: the baseline CIoU loss and UIoU-Focal-inv loss. The evaluation was conducted on the original dataset

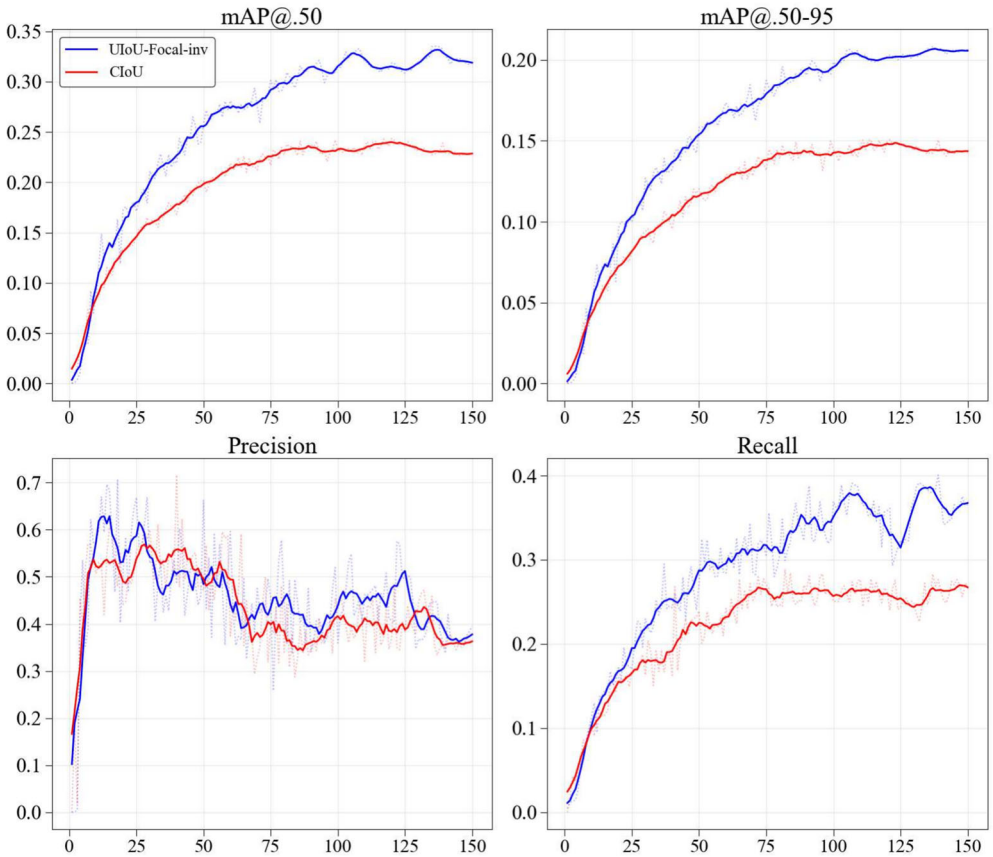


Fig. 12 Comparison vs. loss function. Each subplot illustrates the training performance under two loss functions: UIoU-Focal-inv and CIoU baseline. The metrics shown include mAP, Precision and Recall. Shaded regions around the curves indicate metric fluctuations during training

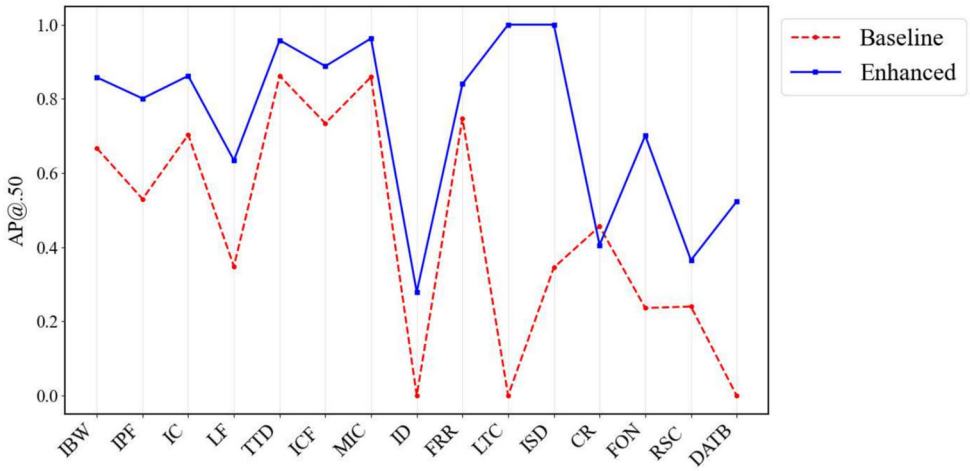


Fig. 13 Comparison vs. enhance

without data augmentation. The results clearly show that replacing CIoU with UIoU-Focal-inv leads to consistent improvements in mAP@50, mAP@50–95, precision, and recall throughout the training process. Notably, the UIoU-Focal-inv loss facilitates more stable convergence and higher final accuracy, validating its effectiveness in improving both localization quality and classification confidence.

Furthermore, our data augmentation method provides the model with more diverse training samples, enhancing the model's recognition ability for most categories. Generally speaking, our designed model achieves notable improvements across the most defect types as Fig. 13. Conduct detailed analysis from the perspective of defect area, defects such as TTD likely cover a significant portion of the image. The relatively high baseline and improved 0.9 suggest that these defects have distinct visual features that are easier for the model to detect even without enhancement. This may also be due to the fact that the HAT module enhances the model's ability to extract detailed texture features, emphasizing the importance of the HAT module in capturing edge cases and paying attention to changes. For defects with smaller area or less pronounced features as LF and ID, the baseline performance is lower (0.348 for LF and nearly 0 for ID). With data enhancement, there is a notable improvement (0.634 for LF and 0.279 for ID). This indicates that data augmentation helps model better detect when the affected size is small. From the perspective of the total number of defect, our method has a greater average increase in defects with fewer samples while an average 1-2 point increased in defects with abundant samples, suggesting that synthetic augmentation might be crucial in bridging the sample scarcity and teaching the model the variability in appearance for these rare defects.

Overall, we confirms that incorporating data augmentation and HAT not only compensates for class imbalance and variability in defect appearance but also significantly improves the performance of defect detection models in practical, real-world scenarios.

7 Conclusion

We center on refining the original model for enhanced object detection performance across specific industrial applications. Our study structure around three key objectives: evaluating our hybrid model's effectiveness, optimizing model parameters, and understanding the impact of training dynamics on performance metrics.

Through extensive experimentation, we observe that increasing batch and input sizes leads to higher GPU utilization but reduces total training time. This trade-off results in improved recall rates but a slight decrease in precision. Additionally, reducing the initial learning rate extends model convergence epochs but stabilizes the training process, mitigating overfitting risks and bolstering model robustness. Conversely, increasing training epochs can enhance accuracy but may compromise generalization.

In summary, we present an effective solution to the persistent challenges of defect detection in UAV-Assisted power line multiple component inspections, including class imbalance, complex backgrounds, and small object localization. Rather than relying solely on architectural depth, our approach combines attention-driven feature refinement and targeted data augmentation to enhance model robustness and generalization. Beyond improvements in detection accuracy, our work underscores the importance of integrating task-specific priors—such as hybrid attention mechanisms and augmentation informed by defect characteristics—into model design. The consistent gains across both common and rare defect categories validate the practicality of our approach in real inspection scenarios.

Moving forward, our findings offer valuable insights for scaling intelligent inspection systems across diverse infrastructure domains. Future research will focus on lightweight deployment and domain-adaptive learning to further bridge the gap between academic models and real-world applications.

Acknowledgements Not applicable.

Author Contributions JHL and MJL conceived the study and designed the model; JHL, MJL, LLL, ZMZ, CY, and GZL collected and analyzed the data; HTP implemented the model and conducted the experiments; JHL, MJL, JYZ, and FL wrote the manuscript draft. All authors discussed and interpreted the results, critically revised the manuscript, and approved the final version.

Funding Not applicable.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Ethical Approval This article does not contain any studies with human participants.

Compliance with Ethical Standards The authors declare compliance with ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Competing Interests The authors declare no competing interests.

References

1. Geng H, Liu L, Li R (2018) Synchronization and reactive current support of pmsg-based wind farm during severe grid fault. IEEE Trans Sustain Energy 9(4):1596–1604
2. Yuan S, Wu H, An Z, Chen Y, Sun A (2023) Design of power transmission line condition monitoring and analysis system based on improved particle swarm optimization. In: 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), pp 67–71
3. Tian X, Wang X, Liu Z, Peng B, Liu Y (2024) Overview of uav-based overhead power line defect detection based on deep learning. In: 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), pp 474–478
4. Duan S, Wang D, Ren J, Lyu F, Zhang Y, Wu H, Shen X (2022) Distributed artificial intelligence empowered by end-edge-cloud computing: A survey. IEEE Commun Surv Tutor 25(1):591–624
5. Duan S, Lyu F, Zhang J, Lu H, Yang P, Wu H, Zhang Y, Shen X (2025) Moco: Urban user mobile contact detection based on cellular signaling trace. IEEE Trans Mobile Comput
6. Lyu F, Ren J, Cheng N, Yang P, Li M, Zhang Y, Shen XS (2020) Lead: Large-scale edge cache deployment based on spatio-temporal wifi traffic statistics. IEEE Trans Mobile Comput 20(8):2607–2623
7. Yolo-drone (2024) A scale-aware detector for drone vision. Chinese J Electron 33(4):1034–1045
8. Zhang T, Xiong Y, Jiang S, Dan P, Gui G (2024) Small target disease detection based on yolov5 framework for intelligent bridges. Peer-to-Peer Netw Appl 17(5):2651–2660
9. Ning Y, Xiang L, Hongyuan J, Xinna S, Ping S, Aidong C (2024) Insulator defect detection in complex scenarios based on cascaded networks with lightweight attention mechanism. Peer-to-Peer Netw Appl 17(4):2123–2136
10. Liu Y, Huang X, Liu D (2024) Weather-domain transfer-based attention yolo for multi-domain insulator defect detection and classification in uav images. Entropy 26(2):136
11. Sun S, Chen C, Yang B, Yan Z, Wang Z, He Y, Wu S, Li L, Fu J (2024) Id-det: Insulator burst defect detection from uav inspection imagery of power transmission facilities. Drones (2504-446X) 8(7)
12. Huang X, Wu Y, Zhang Y, Li B (2022) Structural defect detection technology of transmission line damper based on uav image. IEEE Trans Instrument Measure 72:1–14
13. Zhang Y, Li B, Shang J, Huang X, Zhai P, Geng C (2023) Dsa-net: An attention-guided network for real-time defect detection of transmission line dampers applied to uav inspections. IEEE Trans Instrument Measure 73:1–22
14. Zhang K, Zhou R, Wang J, Xiao Y, Guo X, Shi C (2024) Transmission line component defect detection based on uav patrol images: A self-supervised hc-vit method. IEEE Trans Syst, Man, Cybern: Syst
15. Yu Q, Liu A, Yang X, Diao W (2024) An improved lightweight deep learning model and implementation for track fastener defect detection with unmanned aerial vehicles. Electronics 13(9):1781
16. Yang Z, Xu Z, Wang Y (2022) Bidirection-fusion-yolov3: An improved method for insulator defect detection using uav image. IEEE Trans Instrument Measure 71:1–8
17. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 580–587
18. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28
19. Lin T-Y, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ (2016) Feature pyramid networks for object detection. 2017 IEEE Conf Comput Vis Pattern Recogn (CVPR), 936–944
20. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 779–788
21. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October, 2016, Proceedings, Part I 14, pp 21–37. Springer
22. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2980–2988
23. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 734–750
24. Tian Z, Shen C, Chen H, He T (2019) Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9627–9636
25. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision, pp 213–229. Springer
26. Sujaritha M, Sujatha R (2021) Smart Drone with Open CV to Clean the Railway Track, pp 131–140
27. Huang M, Li H, Zhou Y, Ma T, Su J, Zhou H (2024) A uav aided lightweight target information collection and detection approach. Peer-to-Peer Netw Appl 17(3):1667–1681
28. Bajaj A, Philips B, Lyons E, Westbrook D, Zink M (2020) Determining and communicating weather risk in the new drone economy. In: 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), pp 1–6

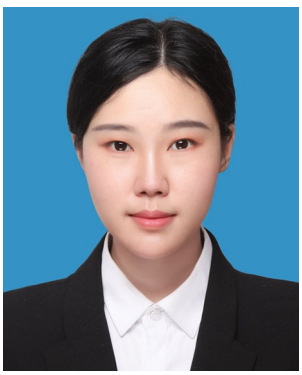
29. Shahbazi M, Théau J, PM.: Recent applications of unmanned aerial imagery in natural resource management. GIScience & Remote Sens 51(4):339–365
30. Liu Y, Zhao J, Jiang G, et al (2024) Fusion of multi-scale and context for small target detection algorithm of unmanned aerial vehicle rescue. Chinese J Internet of Things 08(03):146–156
31. Xiang T-Z, Xia G-S, Zhang L (2019) Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects. IEEE Geosci Remote Sens Magaz 7(3):29–63
32. Chen C, Gong W, Chen Y, Li W (2019) Object detection in remote sensing images based on a scene-contextual feature pyramid network. Remote Sens 11(3)
33. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1492–1500
34. Li J, Liang X, Wei Y, Xu T, Feng J, Yan S (2017) Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1222–1230
35. Hu X, Xu X, Xiao Y, Chen H, He S, Qin J, Heng P-A (2018) Sinet: A scale-insensitive convolutional neural network for fast vehicle detection. IEEE Trans Intell Transport Syst 20(3):1010–1019
36. Im J, Kasahara JYL, Maruyama H, Asama H, Yamashita A (2024) Blend autoaugment: Automatic data augmentation for image classification using linear blending. IEEE Access 12:68770–68784
37. Shin Y, Palakonda V, Yun S, Kim I-M, Kim S-G, Park S-M, Kang J-M (2024) Randmixaugment: A novel unified technique for region- and image-level data augmentations. IEEE Access 12:8187–8197
38. Takahashi R, Matsubara T, Uehara K (2020) Data augmentation using random image cropping and patching for deep cnns. IEEE Trans Circ Syst Video Technol 30(9):2917–2931
39. Zheng Q, Luo Z, Guo M, Wang X, Wu R, Meng Q, Dong G (2025) Hgo-yolo: Advancing anomaly behavior detection with hierarchical features and lightweight optimized detection. arXiv:2503.07371
40. Xing Y, Jiang J, Xiang J, Yan E, Song Y, Mo D (2023) Lightcdnet: Lightweight change detection network based on vhr images. IEEE Geosci Remote Sens Lett 20:1–5
41. Zhang X, Wang Z, Liu D, Lin Q, Ling Q (2021) Deep adversarial data augmentation for extremely low data regimes. IEEE Trans Circ Syst Video Technol 31(1):15–28
42. Alin AY, Kusrini Kusrini Yuana KA (2023) Data augmentation method on drone object detection with yolov5 algorithm. In: 2023 Eighth International Conference on Informatics and Computing (ICIC), pp 1–6
43. Lu H, Lyu F, Ren J, Wu H, Zhou C, Liu Z, Zhang Y, Shen X (2024) Code+: Fast and accurate inference for compact distributed iot data collection. IEEE Trans Parallel Distrib Syst
44. You Q, Wan C, Sun J, Shen J, Ye H, Yu Q (2019) Fundus image enhancement method based on cyclegan. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 4500–4503
45. Shuai Y, Lin Z, Chen W, Shenghuai W, Yu T (2024) Sf-yolo: An evolutionary deep neural network for gear end surface defect detection. IEEE Sensors J 24(13):21762–21775
46. Liu Y, Wang J, Xiao L, Liu C, Wu Z, Xu Y (2025) Foregroundness-aware task disentanglement and self-paced curriculum learning for domain adaptive object detection. IEEE Trans Neural Netw Learn Syst 36(1):369–380
47. Liu W, Ma L, Wang J, xsChen H (2019) Detection of multiclass objects in optical remote sensing images. IEEE Geosci Remote Sens Lett 16(5):791–795
48. Song Z, Zhang Y, Liu Y, Yang K, Sun M (2022) Msfyolo: Feature fusion-based detection for small objects. IEEE Latin America Trans 20(5):823–830
49. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: Exceeding yolo series in 2021. arXiv:2107.08430
50. Lyu C, Zhang W, Huang H, Zhou Y, Wang Y, Liu Y, Zhang S, Chen K (2022) Rtmdet: An empirical study of designing real-time object detectors. arXiv:2212.07784
51. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 9992–10002

Jiehao Li Master's student, graduated from Zhejiang University of Technology in 2016, majoring in environmental science and engineering, senior engineer, deputy manager of the transmission and inspection branch of State Grid Hubei Power Transmission and Transformation Engineering Co., Ltd., research direction: overhead transmission line operation, mainly engaged in ultra-ultra-high voltage transmission line operation and maintenance maintenance, UAV inspection business management.

Manjia Liu received the M.S. degree in energy system from the University of New South Wales, Sydney, Australia, in 2014. Since 2015, she has been employed by State Grid Hubei Electric Power Research Institute as a electrical engineer. She has been primarily engaged in research on power big data analysis and the application of AI in the power system. Her work includes the development of electric vehicle charging devices and study on operational strategies, power load analysis and forecasting, and the study of drone technology for power line inspections.

Haitao Peng is currently a Master student in the School of Computer Science and Engineering at Central South University. His research interests include mobile computing and computer vision.

Longlong Liu Member of the Communist Party of China, engineer, "transmission digital expert" and "five-star work leader" of Hubei Electric Power Company, mainly engaged in the operation and maintenance of ultra-high voltage transmission lines, and research on the direction of digital and intelligent transformation.



Xiaomin Zheng In 2017, he graduated from Three Gorges University in China, majoring in electrical engineering and automation, engineer. "Transmission Digital Expert" of Hubei Electric Power Company, research direction: intelligent inspection, transmission UAV inspection business management.

Chen Yi received B.Sc. in Software Engineering from Northeastern University, China and M.Sc. in Intelligent Systems from the Institute of Systems Science, National University of Singapore. His research interest is computer vision algorithms and their real world applications.

Guozi Liu received B.Sc. in Automation from Tsinghua University, China and M.Sc. in Computer Science from Carnegie Mellon University, USA. His research interest is Machine Learning and its real world applications.

Jieyu Zhou is currently a Ph.D. student in the School of Computer Science and Engineering at Central South University, China. His research interests include mobile computing and data mining.

Feng Lyu received the Ph.D. degree in Department of Computer Science and Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2018. From 2018 to 2020, he was a Postdoctoral Fellow with the Broadband Communications Research Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Full Professor with the School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include mobile networks, beyond 5G networks, big data measurement and application design, and cloud/edge computing. He has published over 100 scientific articles in leading journals and top conferences, including Proceedings of the IEEE, IEEE JSAC, IEEE/ACM ToN, IEEE TMC, IEEE TPDS, and IEEE INFOCOM, VLDB, ACM SenSys, etc. His research results have been highly visible, with 9 publications being ranked as ESI Highly Cited Papers. He is the recipient of the best paper award of IEEE ICC 2019, IEEE/CAA Journal of Automatica Sinica "Norbert Wiener Review Award" in 2020, Outstanding Paper Award from Chinese Journal on Internet of Things in 2021, IEEE Technical Committee on Hyper-Intelligence (IEEE HITC) 2023 Early Career Researcher, and 2024-2025 IEEE ComSoc Distinguished Lecturer. He currently serves as associate editors for IEEE Internet of Things Journal, IEEE Systems Journal, and Peer-to-Peer Networking and Applications, and serves as TPC members for many international conferences.