

ORIGINAL RESEARCH

Open access

Recurrent Attention for Efficient Insulator Defect Detection in Power Transmission Systems Using YOLOv11

Nokuthula Dlamini^{1*}, Siphso Zulu¹

Abstract

Ensuring reliable inspection of high-voltage transmission lines remains vital to maintaining smart grids. Nevertheless, human-based inspections are time-consuming, expensive, and dangerous, and even modern detection models frequently fail to detect small insulator flaws concealed within complex outdoor environments. This study proposes YOLOv11-RAH, a compact network architecture designed specifically for edge devices in smart inspection setups. It integrates the YOLOv11 backbone with a recurrent attention head that progressively enhances multi-scale feature representations via depthwise gated convolutions and terminates refinement upon reaching convergence, thereby boosting localization accuracy with almost no increase in inference time. Robustness to changes in perspective, lighting conditions, and object scale is further strengthened by a four-phase data augmentation protocol that includes geometric distortions, photometric adjustments, Mosaic blending, and random cropping. Evaluation was conducted on a merged insulator dataset compiled from two refined public benchmarks plus a fresh private UAV-collected corpus, capturing varied real-world outdoor scenarios and defect patterns. Broad comparative experiments against alternative detection approaches reveal steady performance gains from the proposed architecture. Detailed ablation analyses further verify the value of the novel attention mechanism and the complete augmentation strategy. In particular, YOLOv11-RAH obtains an approximate 5 percentage point gain in mAP compared with the standard YOLOv11n model on the aggregated public sets PTL-AI Furnas and CPLID, as well as on the private UAV collection AIDD, illustrating excellent cross-source generalization. Overall, these outcomes demonstrate that recurrent attention paired with purpose-built augmentation constitutes a practical and robust strategy for fully automated detection of insulator defects in intelligent power-line monitoring systems.

Keywords Edge intelligence network, Insulator defect detection, Power-transmission inspection, YOLOv11, Recurrent attention

*Correspondence:

Nokuthula Dlamini
dlaminin@ukzn.ac.za

¹ Department of Energy Systems Engineering, Faculty of Engineering, University of KwaZulu-Natal, Durban, South Africa

Introduction

The ongoing shift toward intelligent electrical grids and large-scale industrial IoT ecosystems has intensified demand for autonomous inspection technologies that can operate reliably across expansive, geographically dispersed infrastructure. Maintaining both operational reliability and the physical safety of power grids remains a central concern in this landscape [1]. Conventional inspection approaches still predominantly involve qualified technicians physically ascending transmission towers—a practice that demands considerable manpower, incurs substantial costs, and exposes personnel to hazardous conditions [2]. As electrical networks grow in scale and become increasingly linked to cloud-based surveillance systems, the frequency of inspections is rising while the volume of captured visual information is expanding

dramatically [3]. Such trends clearly reveal the constraints of manual methods and highlight the pressing need for automated inspection solutions that integrate seamlessly with current networked monitoring architectures.

Advances in computer vision have enabled powerful automated visual fault detection of power lines and associated components within intelligent inspection frameworks [4]. Especially noteworthy are the swift improvements in deep-learning-based object detection pipelines—including the YOLO series, SSD, and numerous R-CNN variants—which have proven highly effective at tackling sophisticated recognition challenges in distributed, connected settings [5]. This progress positions these state-of-the-art detectors as an attractive option for identifying insulator anomalies in transmission-line imagery, thereby supporting more efficient smart-grid maintenance practices.

Nevertheless, object detectors developed for insulator monitoring in networked, field-deployed contexts must contend with several unique obstacles inherent to power grid environments. Highly variable natural illumination, diverse atmospheric conditions, and cluttered backgrounds severely complicate consistent and accurate detection [6]. Furthermore, many common defect types—including fine cracks, surface deposits, and subtle displacements—cover only a handful of pixels within high-resolution aerial photographs, posing particular difficulties for resource-constrained edge devices [7]. Compounding these issues is the substantial domain shift that frequently exists between curated laboratory-style datasets and genuine field-acquired imagery; detectors pretrained on broad object-detection collections often exhibit limited adaptability when confronted with the specific visual signatures typical of real-world power transmission scenes [8].

Recent investigations have thus focused on improving pretrained object detectors specifically for insulator fault identification within intelligent inspection networks. The incorporation of attention mechanisms that emphasize important spatial and channel-wise features has demonstrated particular effectiveness for this application [9]. However, a significant number of current attention designs remain limited in their ability to perform the progressive, iterative refinements required to reveal minute defects obscured by intricate backgrounds [10, 11]. To overcome these limitations in networked inspection environments, this work introduces YOLOv11-RAH, a tailored solution for detecting defects on power transmission line insulators. The approach employs a recurrent attention process that iteratively dynamically refines feature representations, enabling the model to progressively concentrate on potentially defective regions until anomalies become distinctly identifiable. The entire architecture has been engineered with the strict computational limitations of edge devices in inspection networks in mind, thereby achieving enhanced efficiency and superior detection performance suitable for real-world deployment. The primary contributions of this research are outlined below:

1. A novel recurrent attention head (RAH) is introduced, featuring a lightweight depthwise-gated convolution unit together with an adaptive early-stopping criterion. The gated convolution operation enhances feature quality at a very low additional cost. At the same time, the early-stopping mechanism ceases iterative updates as soon as convergence is detected, ensuring precise defect localization within strict real-time constraints appropriate for edge-based systems. Complementing this, a comprehensive four-stage data augmentation sequence—encompassing geometric transformations, photometric variations, Mosaic composition, and random cropping—is incorporated to increase dataset variety and bolster model resilience under diverse deployment scenarios encountered in networked inspection.
2. A unified insulator inspection dataset has been assembled by merging two carefully preprocessed public benchmarks (PTL-AI Furnas-I [12] and CPLID [13]) with a newly collected private UAV dataset (AIDD). This combined corpus provides broad coverage of realistic outdoor settings and detailed annotations across multiple defect classes, facilitating robust benchmarking of inspection algorithms in networked contexts.
3. An in-depth experimental evaluation is performed, encompassing comparisons with multiple detection architectures (SSD, Faster R-CNN, YOLOv5n, various YOLOv11 configurations) and detailed ablation studies of the augmentation strategy, with special attention to performance on edge hardware. Under consistent experimental conditions, the proposed YOLOv11-RAH achieves 81.7% mAP on PTL-AI Furnas-I, 75.7% on CPLID, and 79.0% on AIDD—surpassing

the plain YOLOv11n baseline (without attention modules or the proposed augmentation) by roughly 5 percentage points while preserving full compatibility with resource-limited devices.

Related Work

Power-transmission inspection using an object detection network

Object detection—the task of simultaneously localizing and classifying objects in images—has represented a core problem in computer vision for many years. Over the last two decades, the discipline has evolved dramatically, progressing from conventional hand-engineered feature approaches to highly effective deep learning methodologies [14]. Leveraging these algorithmic advances, inspection of power transmission lines has shifted away from purely manual visual assessments toward more advanced techniques, notably those utilizing Unmanned Aerial Systems (UAS) [12].

Drone-assisted inspection represents a significant advancement, delivering reduced costs, improved maneuverability, and safer access to challenging or hazardous locations that are difficult for human crews [13]. Over the past few decades, object detection methods have undergone several major paradigm shifts. Initial techniques relied on manually designed features and exhaustive sliding-window searches, laying the foundation for subsequent real-time-capable systems [15]. A pivotal development occurred in 2014 with the introduction of R-CNN (Regions with Convolutional Neural Networks) by Girshick *et al.* [16], followed by Fast R-CNN in 2015 [17], which markedly enhanced accuracy while mitigating computational bottlenecks. In 2016, SSD (Single Shot MultiBox Detector) [18] improved both speed and precision by leveraging multi-scale feature maps and predefined anchor boxes.

Concurrently, the YOLO (You Only Look Once) framework, proposed by Redmon *et al.* in 2016 [19], pioneered the one-stage detection paradigm and achieved unprecedented real-time inference speeds [20]. Successive iterations from YOLOv1 through YOLOv11 have incorporated extensive architectural innovations, refined loss formulations, advanced feature aggregation techniques, and lightweight building blocks optimized for fast inference [20]. YOLOv2 through YOLOv4 integrated a wide array of performance-enhancing strategies, while YOLOv5 introduced a widely adopted, community-supported PyTorch implementation. In more recent years, a variety of YOLO derivatives and specialized lightweight models have emerged, explicitly targeting edge-device deployment [21]. Models ranging from YOLOv6 to YOLOv8 have placed strong emphasis on boosting detection precision, improving feature representation quality, and reducing computational demands [21]. At the same time, the incorporation of attention mechanisms and neural architecture search (NAS) techniques has fostered the emergence of more adaptive and intelligent object detection systems [22].

Continuing the YOLO family's progression, YOLOv11 maintains the series' focus on achieving an optimal balance between speed and precision. It accomplishes this by adopting more efficient backbone modules and compact spatial attention components, while retaining the anchor-free detection head and multi-task capability [23]. These enhancements deliver a superior accuracy-latency profile compared to prior iterations, positioning YOLOv11 as a compelling foundation for applications spanning both cloud- and edge-based deployments [24]. Drawing on these strengths, the present study selects YOLOv11 as the core detector. It explores the potential of recurrent attention integration combined with domain-specific augmentation to further improve performance in insulator defect detection.

Attention mechanisms in object detection network

Attention mechanisms have attracted considerable interest owing to their ability to adjust feature activations dynamically, enabling models to prioritize the most relevant portions of the input [25]. Within object detection frameworks, such mechanisms are particularly valuable for accentuating key features amid distracting or complex backgrounds—a factor of particular importance in power transmission fault identification tasks.

These attention modules serve as modular add-ons that can be incorporated into either the backbone or the neck components to refine feature extraction and enhance sensitivity across spatial and channel dimensions [26]. In this research, we incorporate two prominent attention mechanisms into the YOLOv11 architecture: Squeeze-and-Excitation (SE) and Coordinate Attention (CA). SE modules improve channel-level feature emphasis via global average pooling and subsequent gating operations, enabling the network to selectively amplify informative channels [26]. CA, by contrast, captures both inter-channel dependencies and accurate location encoding, rendering it highly advantageous for detection problems that demand strong spatial precision [27]. Integrating either of these into YOLOv11 holds promise for substantially raising detection reliability, particularly under demanding conditions characterized by occlusion, diminutive targets, and dense clutter.

This paper presents augmented variants of YOLOv11 that embed SE, CA, and the proposed recurrent attention head (RAH) to boost overall detection effectiveness. Theoretically, embedding attention into YOLOv11 offers several benefits, including the ability to re-prioritize feature activations by focusing on salient areas, thereby elevating the overall quality of extracted representations [28]. Attention further supports the modeling of extended-range interactions and enriched contextual cues—critical for reliable object recognition in intricate visual settings. Additionally, it sharpens feature distinctiveness, empowering the detector to more effectively differentiate between visually similar items or those obscured within noisy or partially hidden contexts [28].

Materials and Methods

To clearly present the technical underpinnings of our proposed edge-intelligence detection framework, this section provides a structured overview of the architecture along with its key novel elements. We begin by depicting the complete detection pipeline in **Figure 1**, where the conventional YOLOv11 backbone–neck structure is augmented by a lightweight recurrent attention head (RAH). This head iteratively refines features at multiple scales before final predictions. Collectively, these innovations produce a high-accuracy detector that remains computationally lightweight, thereby aligning well with the stringent requirements of real-world industrial inspection applications.

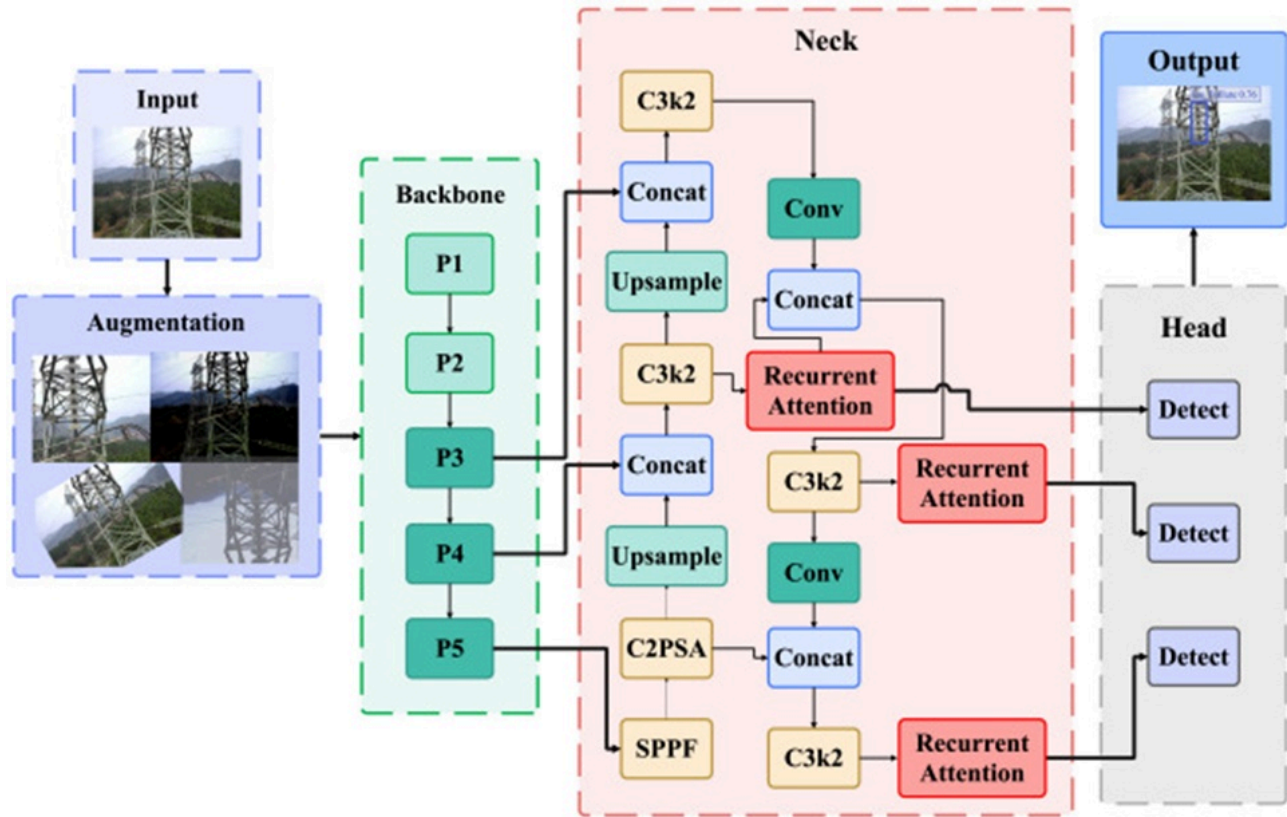


Figure 1. Integration of the recurrent attention head with the YOLOv11 architecture.

Recurrent attention head

In contemporary object detection frameworks, the quality of feature representations plays a decisive role in achieving strong performance. Although standard feed-forward networks can extract rich information from input features, a single forward pass is often insufficient to resolve ambiguities or capture subtle characteristics of target objects. To overcome this limitation, we introduce the recurrent attention head (RAH)—an iterative attention module that applies successive refinement passes to the input features, as illustrated in [Figure 2](#). The design draws inspiration from the Gated Recurrent Unit (GRU). It applies it recurrently across a sequence of steps directly on the feature map.

Let $X \in \mathbb{R}^{C \times H \times W}$ represent the feature tensor arriving from the neck. It is first projected via a 1×1 convolution to a reduced channel dimension h , rendering subsequent depthwise-separable convolutions (DW-Conv) significantly more computationally efficient. The hidden state $H^{\{t\}}$ evolves via a GRU-style gating mechanism that replaces conventional BatchNorm with per-pixel LayerNorm; this substitution ensures stable statistics even under small mini-batch sizes (typically 8–16), which are prevalent in high-resolution industrial datasets. A spatial dropout layer is included to regularize the candidate update path. Ultimately, a final 1×1 convolution projects the refined hidden state to produce the detection logits.

$$Z_t = \sigma(\text{LN}(\text{DWConv}_{3 \times 3}([H_{t-1}, X']))) , \quad (1)$$

$$R_t = \sigma(\text{LN}(\text{DWConv}_{3 \times 3}([H_{t-1}, X']))) , \quad (2)$$

$$\widetilde{\mathbf{H}}_t = \tanh(\text{DropLN}(\text{DWConv}_{3 \times 3}([\mathbf{R}_t \odot \mathbf{H}_{t-1}, \mathbf{X}']))) , \quad (3)$$

$$\mathbf{H}_t = (1 - \mathbf{Z}_t) \odot \mathbf{H}_{t-1} + \mathbf{Z}_t \odot \widetilde{\mathbf{H}}_t. \quad (4)$$

where $\mathbf{X}' \in \mathbb{R}^{h \times H \times W}$ is the channel-reduced input feature map; $\mathbf{H}_t$ denotes the hidden state after the t -th iteration, initialized as $\mathbf{H}_0 = \mathbf{X}'$; $\sigma(\cdot)$ is the sigmoid function constraining gate activations to $[0, 1]$; $\text{DWConv}_{3 \times 3}(\cdot)$ indicates a 3×3 depthwise-separable convolution; $\text{LN}(\cdot)$ stands for per-channel layer normalization applied pixel-wise; $\text{Drop}(\cdot)$ denotes spatial dropout with retention probability p ; and $\odot$ represents element-wise (Hadamard) multiplication.

Input: Feature map $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$

Output: Detection map $\mathbf{Y}_T$

```

1:  $\mathbf{X}' \leftarrow \text{Conv}_{1 \times 1}(\mathbf{X})$ 
2:  $\mathbf{H}_0 \leftarrow \mathbf{X}'$ 
3: for  $t \leftarrow 1$  to  $T$  do
4:    $\mathbf{Z}_t \leftarrow \sigma(\text{LN}(\text{DWConv}_{3 \times 3}([\mathbf{H}_{t-1}, \mathbf{X}'])))$ 
5:    $\mathbf{R}_t \leftarrow \sigma(\text{LN}(\text{DWConv}_{3 \times 3}([\mathbf{H}_{t-1}, \mathbf{X}'])))$ 
6:    $\widetilde{\mathbf{H}}_t \leftarrow \tanh(\text{DropLN}(\text{DWConv}_{3 \times 3}([\mathbf{R}_t \odot \mathbf{H}_{t-1}, \mathbf{X}'])))$ 
7:    $\mathbf{H}_t \leftarrow (1 - \mathbf{Z}_t) \odot \mathbf{H}_{t-1} + \mathbf{Z}_t \odot \widetilde{\mathbf{H}}_t$ 
8:    $\mathbf{Y}_t \leftarrow \text{Conv}_{1 \times 1}(\mathbf{H}_t)$ 
9:   if  $\text{mean}(\mathbf{Z}_t) > 0.9$  then
10:    break
11:  end if
12: end for
13: return  $\mathbf{Y}_T$ 
    
```

▷ channel down-projection
 ▷ initial hidden state
 ▷ Eq. (1)
 ▷ Eq. (2)
 ▷ Eq. (3)
 ▷ Eq. (4)
 ▷ auxiliary prediction

Algorithm 1. Recurrent attention-enhanced feature refinement

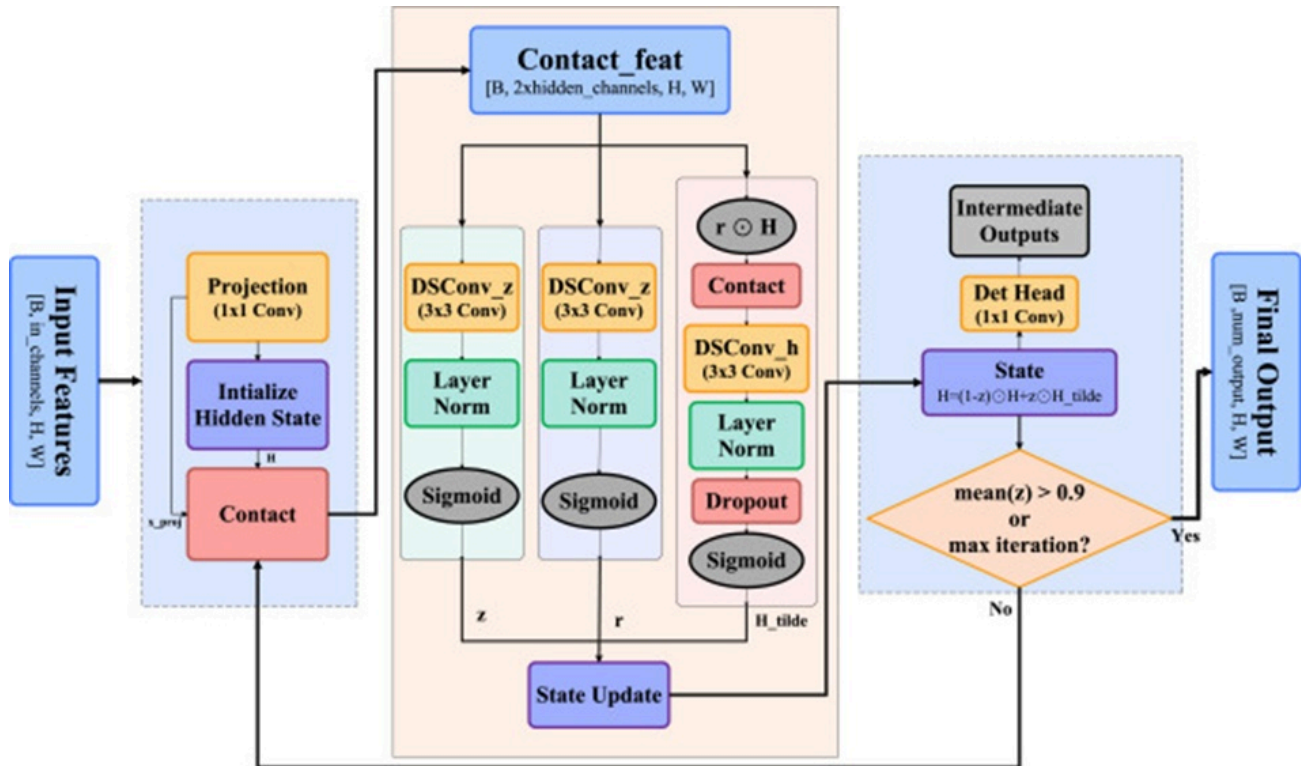


Figure 2. Architecture details of the recurrent attention head.

The module begins by projecting the incoming feature map X through a fixed 1×1 convolution to obtain a static compressed representation X' , which remains unchanged and serves as a persistent input throughout all recurrent steps. Refinement then proceeds iteratively over T timesteps ($t = 1, \dots, T$), with the hidden state H_t embodying a progressively improved version of the features at each step; the process starts with $H_0 = X'$.

The transition from H_{t-1} to H_t relies on learned gating controls that manage information retention and update. Specifically, an update gate Z_t and a reset gate R_t are computed from the concatenation of the prior hidden state H_{t-1} and the fixed input X' , following [Eqs. 1 and 2](#). Both gates employ efficient 3×3 depthwise-separable convolutions (DWConv 3×3) to incorporate spatial context with low overhead, and use layer normalization (LN) instead of batch normalization to maintain training stability across varying batch sizes.

A candidate hidden state $\tilde{H}^t$ is subsequently computed as shown in [Eq. 3](#). The reset gate selectively filters information from the previous hidden state; this filtered state is concatenated with X' and processed through a block consisting of DWConv 3×3 , LN, spatial dropout for regularization, and tanh nonlinearity.

The updated hidden state H_t is then formed by a gated linear interpolation between the previous state and the candidate, as given in [Eq. 4](#), where Z_t controls the relative contribution of retained history versus newly proposed features.

The iterative refinement process terminates when either the spatial average of the update gate surpasses a predefined threshold—i.e., $\text{mean}(Z_t) > \tau$ with $\tau = 0.9$ —or the maximum allowed iteration limit T_{max} is reached ([Algorithm 1](#), line 9). Because $H_t = (1 - Z_t) \odot H_{t-1} + Z_t \odot \tilde{H}^t$, a high mean update-gate value indicates that the majority of spatial locations have largely converged, implying that further iterations would produce only marginal changes (i.e., small $\|H_t - H_{t-1}\|_2$). The impact of different threshold values $\tau \in \{0.85, 0.90, 0.95\}$ is assessed in [Table 1](#); $\tau = 0.9$ was found to provide the most favorable compromise, delivering relatively higher mAP while requiring fewer average iterations. The strict upper bound, T_{max} , ensures predictable, bounded inference latency.

After T refinement steps, the final refined feature map H_T is forwarded to the detection head, achieving performance comparable to much deeper feed-forward architectures but with substantially greater parameter efficiency.

Table 1. Impact of the early-stopping threshold τ on convergence speed and detection accuracy (PTL-AI).

τ	Avg. iters	mAP
0.85	2.9	0.815
0.90	3.0	0.817
0.95	3.5	0.817

Depthwise-separable convolution

Depthwise-separable convolution (DSConv) decomposes a standard dense convolution into two lightweight stages—depthwise followed by pointwise—resulting in approximately a $3 \times$ reduction in FLOPs while fully maintaining the original receptive field size and output tensor dimensions. Formally, given an input tensor $U \in \mathbb{R}^{h \times h \times W}$:

$$V_{c,:,:,i} = \sum_{(i,j) \in \Omega_{3 \times 3}} k_{c,i,j}^{\text{dw}} U_{c,:,:, [* + i, * + j]}, \quad (5)$$

$$\mathbf{Y}_{c',::} = \sum_{c=1}^h k_{c',c}^{\text{pw}} \mathbf{V}_{c,::} \quad (6)$$

where $k_{c',c}^{\text{pw}}$ represents the channel-wise 3×3 kernel coefficients, $\Omega_{3 \times 3}$ indicates the local 3×3 spatial neighborhood, and $k^{\text{pw}} \in \mathbb{R}^{\text{out} \times h}$ is implemented via a 1×1 pointwise convolution.

Input: Tensor $\mathbf{U} \in \mathbb{R}^{B \times h \times H \times W}$
Output: Tensor $\mathbf{Y} \in \mathbb{R}^{B \times h_{\text{out}} \times H \times W}$

```

1: for  $c \leftarrow 1$  to  $h$  do
2:    $\mathbf{V}_{c,::} \leftarrow \text{Conv}_{3 \times 3}^{\text{depthwise}}(\mathbf{U}_{c,::})$ 
3: end for
4:  $\mathbf{Y} \leftarrow \text{Conv}_{1 \times 1}^{\text{pointwise}}(\mathbf{V})$ 
5: return  $\mathbf{Y}$ 

```

▷ Eq. (5)
▷ Eq. (6)

Algorithm 2. Depthwise-Separable Convolution

Data augmentation strategies

Data augmentation is essential for improving model robustness and generalization by synthetically increasing the diversity of the training data through carefully designed transformations applied to original samples [29]. Beyond simply increasing dataset size, these methods help mitigate overfitting, build resistance to diverse imaging conditions, and enable the detector to better handle variations in object orientation, scale, and environmental factors commonly encountered in industrial inspection settings [30]. The proposed framework incorporates four complementary augmentation techniques that maintain semantic integrity while introducing meaningful variability; illustrative examples of processed images appear in the left portion of **Figure 1**. The complete sequence of these four operations is outlined in **Algorithm 3**. Lines 1–3 perform sequential geometric alterations, line 4 introduces photometric distortions, lines 5–7 conditionally construct a Mosaic by combining randomly selected neighboring images, and line 8 executes the final random crop followed by resize before outputting the augmented sample. Each stage is described below in the same sequence.

Input: Image tensor $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$
Output: Augmented image tensor $\mathbf{I}^* \in \mathbb{R}^{C \times H \times W}$

```

1:  $\mathbf{J} \leftarrow \text{RandomRotation}(\mathbf{I})$ 
2:  $\mathbf{J} \leftarrow \text{RandomFlip}(\mathbf{J})$ 
3:  $\mathbf{J} \leftarrow \text{RandomScale}(\mathbf{J})$ 
4:  $\mathbf{K} \leftarrow \text{ColorJitter}(\mathbf{J})$ 
5: if  $\text{rand}() < p_{\text{mosaic}}$  then
6:    $\mathbf{K} \leftarrow \text{Mosaic}(\mathbf{K}, \text{rand\_images}())$ 
7: end if
8:  $\mathbf{I}^* \leftarrow \text{RandomCropResize}(\mathbf{K})$ 
9: return  $\mathbf{I}^*$ 

```

▷ geometric transforms
▷ photometric transforms

Algorithm 3. Progressive Data Augmentation Pipeline

First, geometric transformations include random rotations, horizontal flipping with a controlled probability, and uniform scaling centered around the original resolution. These operations effectively mimic a variety of viewpoints and object attitudes—particularly beneficial in UAV-based aerial inspections where camera angles vary due to changing flight trajectories and environmental disturbances such as wind [31].

Second, photometric augmentations alter properties of the color space to improve tolerance to lighting variations [29]. Random modifications to brightness and contrast, along with subtle shifts in hue, saturation, and value, replicate the effects of differing illumination conditions and seasonal changes, enabling the model to adapt more effectively to outdoor atmospheric and time-of-day variations.

Third, the Mosaic augmentation combines multiple randomly selected images into a single composite arranged in a 2×2 grid. Each constituent image is independently resized and positioned within its assigned quadrant, creating rich, multi-instance scenes that promote the learning of contextual cues and enhance the detection of small or sparsely distributed objects [32]. This method greatly increases scene complexity and diversity without compromising training throughput.

Finally, random cropping extracts sub-regions from the original image at varying relative sizes. By compelling the detector to recognize objects across a broad range of scales and locations within the frame, this technique improves performance on targets captured at different distances from the sensor [33].

Experiments and Evaluation

Details of datasets

To thoroughly assess both the effectiveness and computational efficiency of the proposed attention mechanism integrated with YOLOv11, extensive experiments were performed across three distinct datasets. These datasets represent a variety of real-world power transmission line inspection contexts and include two publicly available benchmarks and one private dataset, collectively offering rich, diverse resources for evaluating insulator defect detection performance, as summarized in Table 2.

Table 2. Overview of datasets used in experiments.

Dataset	Number of categories	Number of images	Annotations per image
PTL-AI Furnas-I	3	12 548	1.9
CPLID	2	848	1.8
AIDD (ours)	4	2058	1.2

Public datasets: PTL-AI Furnas-I and CPLID

The first benchmark employed is PTL-AI Furnas-I, which constitutes a specialized subset extracted from the publicly released PTL-AI Furnas dataset [12]. The “I” suffix specifically indicates an insulator-focused selection, retaining only the three insulator-related categories for this study. This filtered subset preserves the broad environmental diversity and imaging variability of the parent collection—such as differing lighting conditions, observation angles, and background complexity—while concentrating exclusively on insulator-related detection challenges.

The second public benchmark is the Chinese Power Line Insulator Dataset (CPLID) [13], a widely recognized resource designed for both insulator localization and defect classification. It contains annotated aerial images captured under a range of viewpoints and illumination settings. Together, these two open-access datasets provide a reliable, standardized foundation for measuring the generalization ability and robustness of the proposed approach.

Table 3. Detailed breakdown of the insulator-combined dataset.

Datasets	Categories	Number of images	Image width (px)
AIDD	Insulator_Normal	750	640–1200
	Insulator_Breakage	375	1000–8000

Insulator_Pollution	518	640–8000
Insulator_Shifting	415	640–8000

Proprietary dataset: Aerial Insulator Defect Detection (AIDD)

The third dataset is the proprietary aerial insulator defect detection dataset (AIDD), compiled from operational power grid inspections conducted in China. In addition to images of intact insulators, AIDD includes a comprehensive collection of real-field defect instances encompassing breakage, contamination (pollution), and positional displacement (shifting). A detailed category-wise distribution of the AIDD dataset is presented in Table 3. The corpus covers four insulator states: normal condition plus three distinct defect types. Image resolutions span from 640 to 8000 pixels in width, capturing a wide spectrum of acquisition scenarios and thereby strengthening the evaluation under realistic aerial imaging variability.

Although the original images exhibit significant size differences, all inputs are uniformly resized during training to maintain consistent dimensions across models. For instance, SSD-based detectors process images resized to 300 × 300 pixels [34], while YOLO-family models standardize inputs to 640 × 640 pixels [35]. This normalization step ensures compatibility among different detection pipelines and removes any bias arising from varying native resolutions. Furthermore, annotations are provided in dual formats to support diverse training pipelines: a detailed XML format with full annotation metadata and a concise TXT format tailored for efficient YOLO training. Representative samples of the various defect categories from the AIDD dataset are displayed in Figure 3, highlighting the realistic range of fault appearances observed during actual transmission line inspections.

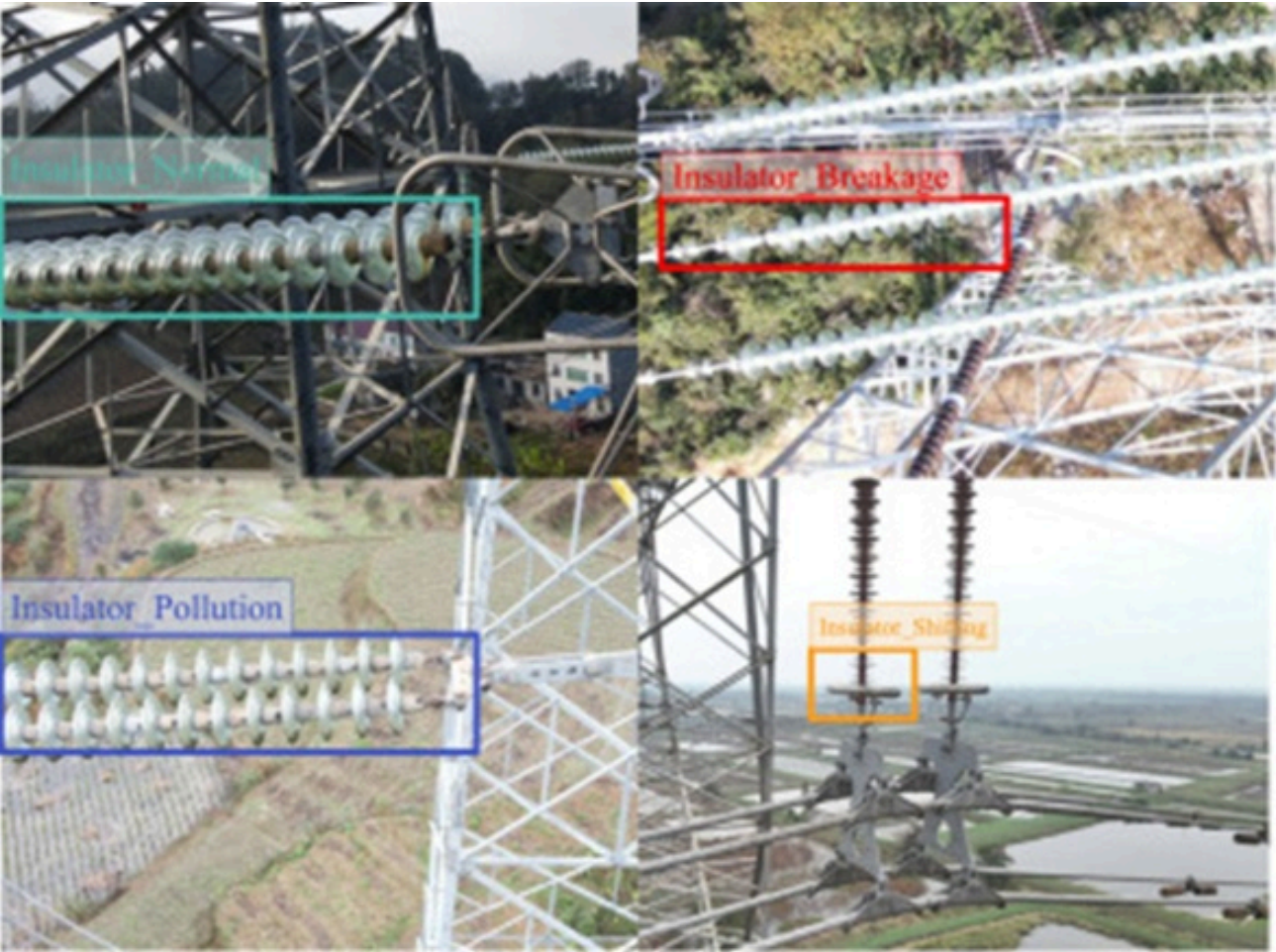


Figure 3. Representative examples of insulator defects from the proprietary dataset.

Training implementation and evaluation metrics

All experimental trials were conducted on a workstation equipped with two NVIDIA GeForce RTX 4090 graphics cards (each with 24 GB of VRAM) and 64 GB of system RAM. The operating system was Ubuntu 20.04 LTS, with the software stack comprising Python 3.9, PyTorch 1.12.0 (built with CUDA 11.6 support), and Anaconda 4.12.0 for environment and dependency management. Image processing was handled by OpenCV 4.6.0, while NumPy 1.21.0 and Matplotlib 3.5.0 were utilized for numerical operations and result visualization, respectively.

Training was performed with an initial learning rate (lr0) of 0.01, a final one-cycle learning rate (lrf) of 0.01, a batch size of 16, and 300 total epochs. All remaining hyperparameters were kept at their default settings.

For fair and reproducible evaluation, each dataset was partitioned into three disjoint subsets: 70% allocated for training, 15% for validation, and 15% reserved for testing and final performance reporting. To avoid bias and maintain class balance, the split was performed randomly with stratified sampling, ensuring that the proportions of each defect category in each subset closely matched their distributions in the original dataset.

The experimental comparison includes seven distinct model configurations to enable a thorough performance analysis. The baseline detectors comprise SSD with MobileNet backbone [34], Faster R-CNN [36], YOLOv5n [35], and YOLOv11n [24]. The enhanced versions build upon YOLOv11n by incorporating different attention modules: Squeeze-and-Excitation (SE) [37], Coordinate Attention (CA) [38], and the custom recurrent attention head (RAH) proposed in this work.

Model performance is assessed using two primary groups of metrics that together provide a complete picture of detection quality. Standard object detection indicators are employed to measure precision and recall behavior, specifically Precision, Recall, and mean Average Precision at IoU = 0.5 (mAP@0.5), all calculated in accordance with the COCO evaluation protocol. Consistent with common practice in object detection benchmarking, the reported mAP for each dataset is the average per-class Average Precision across all categories, assigning equal importance to each class regardless of the number of instances [39].

Effect of data augmentation

The augmentation pipeline consists of four complementary operations: geometric transformations, photometric adjustments, Mosaic synthesis, and random cropping. The resulting impact on dataset scale is shown in **Figure 4**. Applying the full augmentation sequence increases the effective training set size by nearly fourfold for each corpus: PTL-AI Furnas-I grows from 12,548 to 50,192 images, CPLID from 848 to 3,392 images, and AIDD from 2,058 to 8,232 images. The logarithmic scale adopted in the figure effectively illustrates the multiplication factor while preserving the semantic integrity of the original images. Consequently, the augmented datasets introduce substantially greater visual variability during training without modifying the fundamental detection objective.

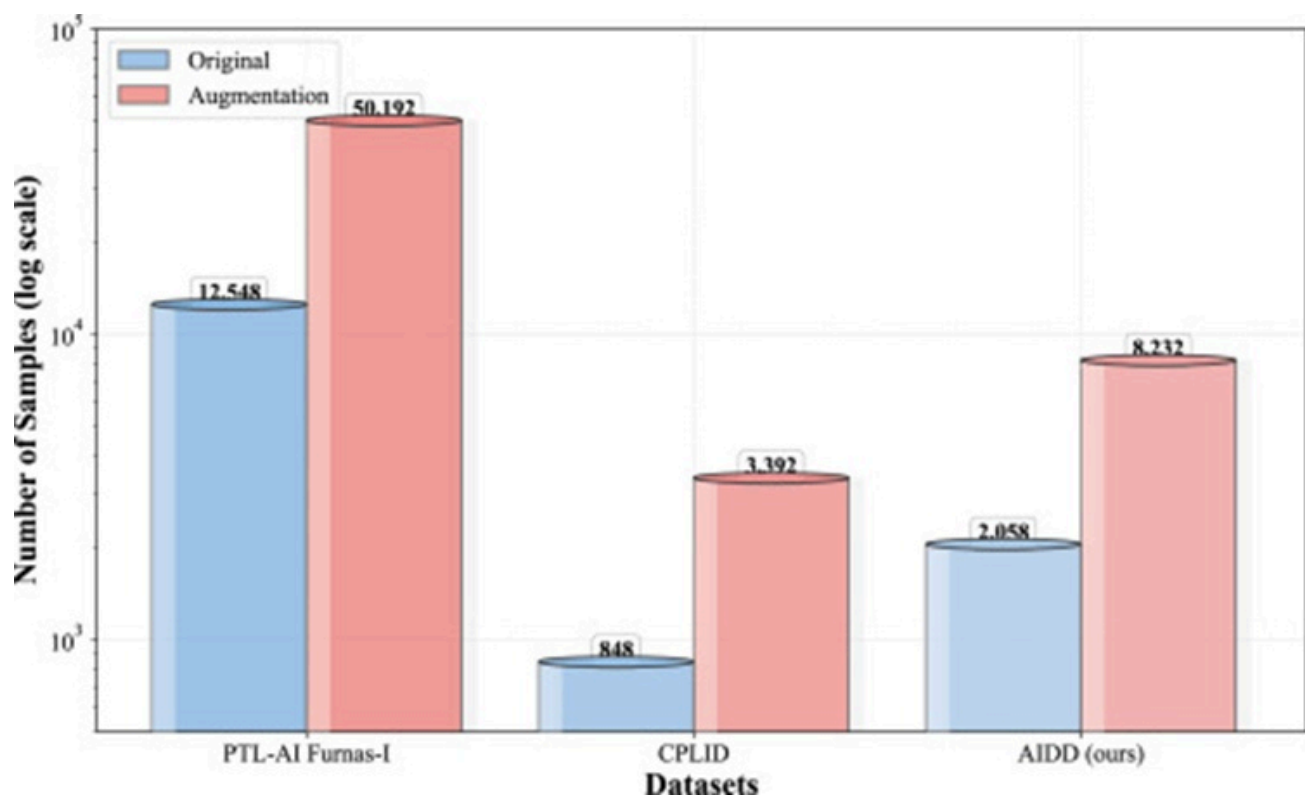


Figure 4. Dataset size before and after augmentation.

Overall performance

Table 4 presents a comprehensive performance comparison across the three datasets, reporting the highest achieved results under fully augmented training conditions. The two traditional detectors—SSD and Faster R-CNN—exhibit only moderate accuracy on all evaluated corpora. Their mAP values range from 0.629 (on CPLID) to 0.697 (on PTL-AI), with Precision confined to the interval 0.641–0.702 and Recall slightly lower at 0.618–0.683. These figures indicate that while classic two-stage architectures remain functional for aerial power-line imagery, they lack the necessary precision and adaptability required to handle highly cluttered backgrounds and variable illumination effectively.

The YOLO family consistently delivers markedly superior accuracy compared to the traditional baseline detectors. Transitioning from YOLOv5n to the more recent YOLOv11n backbone yields clear improvements in mAP: from 0.729 to 0.775 on PTL-AI Furnas-I, from 0.693 to 0.734 on CPLID, and from 0.706 to 0.753 on AIDD, accompanied by comparable increases in both precision and recall. Adding attention modules yields additional consistent improvements: the SE-integrated version achieves mAP scores of 0.785, 0.759, and 0.772, while the CA-integrated variant improves them to 0.789, 0.763, and 0.776. This stepwise improvement demonstrates that both channel-focused (SE) and spatially aware (CA) attention mechanisms improve the modeling of spatial features, particularly for small or partially obscured targets.

Table 4. Best experiment results comparison using augmented datasets, where bold values indicate the top performance in each category, and mAP is computed at an IoU threshold of 0.5.

Models	PTL-AI	Furnas-I	CPLID	AIDD
	Precision	Recall	mAP	Precision
SSD	0.668	0.642	0.655	0.641
Faster RCNN	0.702	0.683	0.697	0.675

YOLOv5n	0.745	0.724	0.729	0.698
YOLOv11n	0.792	0.768	0.775	0.748
YOLOv11n-SE	0.801	0.770	0.785	0.774
YOLOv11n-CA	0.806	0.781	0.789	0.778
YOLOv11-RAH (ours)	0.825	0.809	0.817	0.765

The proposed recurrent-attention head (YOLOv11-RAH) achieves the highest scores on the largest corpus, PTL-AI Furnas-I, recording 0.825 precision, 0.809 recall, and 0.817 mAP. On the smaller CPLID and the medium-sized AIDD datasets, its mAP values (0.757 and 0.790, respectively) still outperform all competing models, indicating that the recurrent refinement process provides the most substantial gains when ample training data are available, yet remains highly competitive even in data-constrained scenarios. Taken together, the consistent upward trend—from the plain YOLOv11n through attention-augmented versions to the recurrent variant—demonstrates that contemporary lightweight architectures paired with deliberate attention mechanisms offer the most reliable and accurate detection for UAV-captured power line inspection imagery. Qualitative detection outcomes are presented in [Figure 5](#).



Figure 5 . Detection examples from the AIDD dataset.

Ablation study

To precisely quantify the contribution of each proposed component, a detailed ablation study was conducted across all three datasets. Two primary factors were investigated: the inclusion of the Recurrent-Attention Head (RAH) module and the application (or absence) of the data augmentation pipeline (Aug.). All other training configurations and hyperparameters were held constant to ensure direct comparability of the results reported in [Table 5](#).

The results summarized in [Table 5](#) reveal three consistent patterns across the datasets. First, enabling data augmentation consistently boosts every model variant by approximately two to three percentage points in precision, recall, and mAP. This underscores the direct benefit of increased visual diversity for improving generalization. The effect is especially noticeable on CPLID, the dataset with the smallest sample count, where augmentation delivers the largest relative improvement—emphasizing how synthetic variability becomes particularly valuable under limited data availability. Moreover, the augmentation operations were specifically designed and tuned for UAV-based insulator imaging, effectively

addressing prevalent challenges such as viewpoint shifts, lighting fluctuations, and scale differences, thereby significantly enhancing overall robustness and detection reliability.

Second, incorporating the recurrent-attention head (RAH) provides an additional, nearly uniform gain over the baseline backbone in every case. Across all three datasets, the RAH-only configuration surpasses the plain backbone by roughly two percentage points in precision, recall, and mAP, confirming that the added iterative spatial refinement improves detection performance even in the absence of extra training samples.

Third, the combination of data augmentation and RAH consistently yields the best results across all metrics and datasets. Without augmentation, the RAH variant already matches or exceeds the performance of the augmented baseline model; when both are applied together, mAP reaches 0.817 on PTL-AI Furnas-I, 0.757 on CPLID, and 0.790 on AIDD, with precision and recall also attaining their maximum values.

Table 5. Ablation results on the three different datasets, where bold values indicate the top performance in each category, and mAP is computed at an IoU threshold of 0.5.

Datasets	RAH	Aug.	Precision	Recall	mAP
Empty Cell	×	×	0.771	0.752	0.766
PTL-AI Furnas-I	✓	×	0.813	0.794	0.802
Empty Cell	×	✓	0.792	0.768	0.775
Empty Cell	✓	✓	0.825	0.809	0.817
Empty Cell	×	×	0.728	0.706	0.719
CPLID	✓	×	0.740	0.714	0.727
Empty Cell	×	✓	0.748	0.715	0.734
Empty Cell	✓	✓	0.765	0.744	0.757
Empty Cell	×	×	0.755	0.735	0.745
AIDD	✓	×	0.780	0.755	0.770
Empty Cell	×	✓	0.762	0.740	0.753
Empty Cell	✓	✓	0.803	0.775	0.790

Edge hardware evaluation

To further assess the practical viability of the proposed YOLOv11-RAH approach for real-world deployment, targeted experiments were conducted comparing it against several YOLO baseline models on a widely adopted embedded GPU development platform—the NVIDIA Jetson Orin NX with 16 GB of memory. This board has a peak power draw of only 25 W, rendering it an excellent representative for simulating typical edge-device deployment scenarios in power line inspection systems [40].

All evaluated models were first converted from PyTorch to the ONNX format and then optimized into TensorRT inference engines using FP16 precision. To preserve consistency with the main experimental setup, the input resolution remained fixed at 640 × 640 pixels, and all preprocessing and postprocessing steps were kept identical. Performance was measured across four key dimensions: model parameter count (Params), computational complexity (FLOPs), detection accuracy (mAP@0.5), and per-image inference latency. Because model parameters and FLOPs are architecture-

dependent and independent of the specific dataset, these edge-deployment tests were conducted exclusively on the proprietary AIDD dataset to eliminate redundant computation while still providing representative real-world timing results.

The outcomes, detailed in [Table 6](#), demonstrate that YOLOv11-RAH achieves the most favorable trade-off between detection accuracy and computational efficiency among the compared models. Although the addition of the recurrent attention head (RAH) modestly increases the total parameter count and FLOPs relative to the plain YOLOv11n, the measured end-to-end inference latency on the Jetson Orin NX is actually shorter. This counterintuitive improvement arises from the adaptive early-stopping behavior inherent in the RAH design, which intelligently halts iterative refinement once convergence is reached, thereby avoiding unnecessary computation during attention propagation. As a result, the extra operations introduced by the recurrent module do not impose a meaningful runtime penalty; instead, they deliver more focused and discriminative features that enhance overall detection quality. All four models sustain real-time performance on the Jetson Orin NX 16 GB platform, with per-image latency remaining comfortably below 20 ms—fully satisfying the timing constraints of field-deployable inspection applications.

In conclusion, these edge-specific evaluations confirm that YOLOv11-RAH strikes an effective balance between superior detection precision and efficient inference speed. The model exhibits clear accuracy gains (in mAP@0.5) over the baseline configurations while preserving a lightweight footprint and real-time responsiveness. The reported figures for latency, parameter count, and FLOPs on the Jetson Orin NX platform collectively affirm that YOLOv11-RAH satisfies the essential requirements for practical, resource-constrained edge deployment, thereby reinforcing the real-world applicability of the proposed method in intelligent power transmission inspection networks.

Table 6. Edge deployment performance on AIDD (Jetson Orin NX 16 GB, FP16). Latency is end-to-end per image; lower values are better.

Models	Params (M)	FLOPs (G)	mAP	Latency (ms/image)
YOLOv5n	6.9	6.4	0.706	19.2
YOLOv11n	7.4	6.9	0.753	17.5
YOLOv11-RAH (ours)	7.8	7.3	0.790	16.9

Conclusion

The primary objective of this study was to investigate whether a lightweight detection architecture could consistently identify minute insulator defects in complex, real-world power transmission line images while maintaining real-time inference capabilities essential to intelligent inspection systems. By integrating the highly efficient YOLOv11 backbone with a novel recurrent attention head (RAH) and a carefully designed four-stage data augmentation pipeline, the proposed YOLOv11-RAH successfully meets these requirements. Evaluated across the unified PTL-AI Furnas, CPLID, and AIDD datasets, the model consistently achieves an approximate 5 percentage-point improvement in mAP over the baseline YOLOv11n, while retaining the low-latency characteristics of single-stage detection. This performance demonstrates both superior accuracy and strong generalization to previously unseen UAV platforms and environmental settings in networked inspection contexts.

Three key elements drive these results. First, the entropy-driven RAH iteratively refines multi-scale feature representations, markedly improving the localization of subtle defects while employing adaptive early stopping to virtually eliminate any increase in inference time. Second, the combination of geometric, photometric, Mosaic, and random-crop augmentations substantially increases the diversity of the training data, thereby enhancing the model's resilience to variations in camera viewpoint, lighting conditions, and object scale. Third, the newly assembled insulator dataset—merging carefully cleaned public benchmarks (PTL-AI Furnas-I and CPLID) with a proprietary UAV collection (AIDD)—

establishes a challenging, diverse, and reproducible evaluation framework suitable for benchmarking intelligent inspection algorithms.

This work delivers two main contributions: (1) a portable and effective recipe that combines targeted data augmentation with recurrent attention mechanisms, and (2) a unified, publicly shareable insulator inspection corpus that integrates PTL-AI Furnas-I, CPLID, and AIDD for standardized and comparable research. Furthermore, targeted evaluations on resource-constrained edge hardware have confirmed the model's practicality for real-world deployment within intelligent inspection networks.

Looking ahead, future efforts will focus on rigorously testing the model's robustness under more severe and variable weather conditions—such as dense fog, heavy rainfall, and high winds—commonly encountered in operational power transmission environments. To support this direction, we plan to expand the existing datasets by incorporating images captured under these challenging conditions and to systematically evaluate the detector's stability and adaptability within networked settings. In parallel, we intend to broaden the evaluation scope by applying the model to a much larger and more diverse collection of transmission infrastructure datasets gathered across different geographic regions, seasons, and climatic zones, while continuing to monitor and optimize performance on edge computing platforms deployed in intelligent inspection systems.

Acknowledgements

This work was supported by the Science and Technology Project of State Grid under grant number [5500-202322539A-3-2-ZN](#).

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Mar 2023

Revised: 27 Apr 2023

Accepted: 18 Jun 2023

Published online: 20 September 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Alhassan AB, Zhang X, Shen H, Xu H. Power transmission line inspection robots: a review, trends and challenges for future research. *Int J Electr Power Energy Syst*. 2020;118:105862.
<https://doi.org/10.1016/j.ijepes.2020.105862>.
- Takaya K, Ohta H, Kroumov V, Shibayama K, Nakamura H, Mizuno Y. Construction of power line inspection system using a quadrotor helicopter. *IEEJ Trans Electr Electron Eng*. 2024;19(2):280–8.
<https://doi.org/10.1002/tee.24150>.
- Liu Y, Yang X, Wen W, Xia M. Smarter grid in the 5G era: a framework integrating power internet of things with a cyber physical system. *Front Commun Netw*. 2021;2:689590.
<https://doi.org/10.3389/frcmn.2021.689590>.
- Hütten N, Alves Gomes M, Hölken F, Andricevic K, Meyes R, Meisen T. Deep learning for automated visual inspection in manufacturing and maintenance: a survey of open-access papers. *Appl Syst Innov*. 2024;7(1):11.
<https://doi.org/10.3390/asi7010011>.
- Hussain M. Comparative analysis of deep learning image detection algorithms. *J Big Data*. 2021;8(1):66.
<https://doi.org/10.1186/s40537-021-00434-w>.
- Liu C, Wu Y, Liu J, Sun Z, Xu H. Improved YOLOv3 network for insulator detection in aerial images with diverse background interference. *Electronics*. 2021;10(7):771.
<https://doi.org/10.3390/electronics10070771>.
- Liu C, Wu Y, Liu J, Sun Z, Xu H. Summary of insulator defect detection based on deep learning. *Electr Power Syst Res*. 2023;224:109688.
<https://doi.org/10.1016/j.epsr.2023.109688>.
- Xu X, Feng H, Wang S, Liu Y, Liu X, Zhang H, et al. Weather-domain transfer-based attention YOLO for multi-domain insulator defect detection and classification in UAV images. *Entropy*. 2024;26(2):136.
- Xu ZY, Tang X. Transmission line insulator defect detection algorithm based on MAP-YOLOv8. *Sci Rep*. 2025;15:10288.
<https://doi.org/10.1038/s41598-025-92445-3>.
- Wang Y, Li Z, Zhang W, Liu Y, Chen H. A high precision YOLO model for surface defect detection based on PyConv and CISBA. *Sci Rep*. 2025;15:1930.
<https://doi.org/10.1038/s41598-025-91930-z>.
- Xiang X, Liu M, Zhang S, Wei P, Chen B. Multi-scale attention and dilation network for small defect detection. *Pattern Recognit Lett*. 2023;172:82–8.
<https://doi.org/10.1016/j.patrec.2023.06.010>.
- FS de Oliveira, de Carvalho M, Campos PHT, Soares ADS, Júnior AC, Silva Quirino ACR, et al. PTL–AI Furnas Dataset: A public dataset for fault detection in power transmission lines using aerial images. In: *Proc 35th SIBGRAPI Conf Graph Patterns Images*; 2022. p. 7–12.
<https://doi.org/10.1109/SIBGRAPI55357.2022.9991806>.
- Tao X, Zhang D, Wang Z, Liu X, Zhang H, Xu D. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Trans Syst Man Cybern Syst*. 2018.
- Sun Y, Sun Z, Chen W. The evolution of object detection methods. *Eng Appl Artif Intell*. 2024;133:108458.
<https://doi.org/10.1016/j.engappai.2024.108458>.

Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055.
<https://doi.org/10.48550/arXiv.1905.05055>.

Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524.
<https://doi.org/10.48550/arXiv.1311.2524>.

Girshick R. Fast R-CNN. arXiv preprint arXiv:1504.08083.
<https://doi.org/10.48550/arXiv.1504.08083>.

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: Single shot MultiBox detector. In: Proc Eur Conf Comput Vis (ECCV). 2016:21–37.
https://doi.org/10.1007/978-3-319-46448-0_2.

Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2016:779–88.
<https://doi.org/10.1109/CVPR.2016.91>.

Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. arXiv preprint arXiv:1506.02640.
<https://doi.org/10.48550/arXiv.1506.02640>.

Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extr. 2023;5(4):1680–716.
<https://doi.org/10.3390/make5040083>.

Sah S, Kumar R, Ganji DC, Saboori E. ActNAS: generating efficient YOLO models using activation NAS. arXiv preprint arXiv:2410.10887.
<https://doi.org/10.48550/arXiv.2410.10887>.

Hidayatullah P, Syakrani N, Sholahuddin MR, Gelar T, Tubagus R. YOLOv8 to YOLO11: a comprehensive architecture in-depth comparative review. arXiv preprint arXiv:2501.13400.
<https://doi.org/10.48550/arXiv.2501.13400>.

Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.
<https://doi.org/10.48550/arXiv.2410.17725>.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv preprint arXiv:1706.03762.
<https://doi.org/10.48550/arXiv.1706.03762>.

Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: Proc IEEE Int Conf Acoust Speech Signal Process (ICASSP). 2023. p. 1–5.
<https://doi.org/10.1109/ICASSP49357.2023.10096516>.

Qiu X, Min Z. A review of attention mechanisms in computer vision. In: Proc 2023 8th Int Conf Image Vis Comput (ICIVC). 2023:577–83.
<https://doi.org/10.1109/ICIVC58118.2023.10270435>.

Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, et al. Attention mechanisms in computer vision: A survey. Comput Vis Media. 2022;8(3):331–68.
<https://doi.org/10.1007/s41095-022-0271-y>.

Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48.
<https://doi.org/10.1186/s40537-019-0197-0>.

Kim H, Kim D, Ahn P, Suh S, Cho H, Kim J. ContextMix: a context-aware data augmentation method for industrial visual inspection systems. arXiv preprint arXiv:2401.10050.
<https://doi.org/10.48550/arXiv.2401.10050>.

Tang G, Ni J, Zhao Y, Gu Y, Cao W. A survey of object detection for UAVs based on deep learning. *Remote Sens.* 2024;16(1):149.
<https://doi.org/10.3390/rs16010149>.

Zhang H, Zhang S, Zou R. Select-mosaic: Data augmentation method for dense small object scenes. (2024), preprint arXiv:2406.05412.
<https://doi.org/10.48550/arXiv.2406.05412arXiv>.

Yang R, Wang R, Deng Y, Jia X, Zhang H. Rethinking the random cropping data augmentation method used in the training of CNN-based SAR image ship detector. *Remote Sens.* 2021;13(1):34.
<https://doi.org/10.3390/rs13010034>.

Li C. High quality, fast, modular reference implementation of SSD in pytorch(2018). <https://github.com/lufficc/SSD>

Jocher G. Ultralytics YOLOv5.(2020), URL <https://github.com/ultralytics/yolov5>, (Accessed: 06 May 2025)
<https://doi.org/10.5281/zenodo.3908559>..

Yang J, Lu J, Contributors. Faster R-CNN PyTorch implementation(2017)URL <https://github.com/jwyang/faster-rcnn.pytorch>, Supports multi-GPU training, RPN, ROI pooling/align; includes demo and training scripts

Yang L, Sun S, Li Z, Shi Y, Zhang Z, Liang Y, et al. YOLO-powerse: An efficient deep learning model for UAV-based power cable fault detection. *Proceedings of the 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT;2025.* p. 1161-5.
<https://doi.org/10.1109/AINIT65432.2025.11035129>.

Wu J, Dong J, Nie W, Ye Z. A lightweight YOLOv5 optimization of coordinate attention. *Appl Sci.* 2023;13(3):1746.
<https://doi.org/10.3390/app13031746>.

Padilla R, Passos WL, Dias TL, Netto SL, da Silva EA. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics.* 2021;10(3):279.

Rey L, Bernardos AM, Dobrzycki AD, Carramiñana D, Bergesio L, Besada JA, et al. A performance analysis of you only look once models for deployment on constrained computational edge devices in drone applications. *Electronics.* 2025;14(3):638.
<https://doi.org/10.3390/electronics14030638>.