Ain Shams Engineering Journal 17 (2026) 104067

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Full length article

Input attention, squeeze and excitation, and spatial transformer of YOLO for fault detection using UAV

João Pedro Matos Carvalho [a, b, c, *], Stefano Frizzo Stefenon [d, e], Valderi Reis Quietinho Leithardt [b, f, g], Laio Oriel Seman [h], Kin-Choong Yow [e], Juan Francisco De Paz Santana [g]

[a] LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[b] COPELABS, Lusófona University, Campo Grande 376, 1749-024 Lisbon, Portugal
[c] Center of Technology and Systems (UNINOVA-CTS) and LASI, 2829-516 Caparica, Portugal
[d] Instituto Superior de Engenharia de Lisboa, ISEL, Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal
[e] Faculty of Engineering and Applied Sciences, University of Regina, Saskatchewan, S4S 0A2, Canada
[f] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026, Lisboa, Portugal
[g] Expert Systems and Applications Lab, Faculty of Science, University of Salamanca, 37008, Salamanca, Spain
[h] Department of Automation and Systems Engineering, Federal University of Santa Catarina, Florianopolis, SC, Brazil

ARTICLE INFO

Keywords:
Fault detection
Input attention
YOLO
Power grid
Squeeze and excitation
Spatial transformer

ABSTRACT

The detection of faults in insulators is important to guarantee the continuous supply of electricity. To identify faults in these components, various object detection methods based on deep learning have been explored. This paper investigates architectural enhancements to the You Only Look Once (YOLO) framework for fault detection in electrical power grid insulators. Three structural variants are proposed: the Input Attention Transformer (IAT-YOLO) for spatial feature refinement, Squeeze-and-Excitation (SAE-YOLO) modules for channel recalibration, and Spatial Transformer Networks (STN-YOLO) for geometric alignment. Experiments were conducted on a publicly available insulator dataset from Unmanned Aerial Vehicles (UAVs), comprising seven defect categories, including pollution, breakage, and flashover damage. Results demonstrate that STN-YOLO and SAE-YOLO consistently improve generalization and robustness, achieving mAP values of up to 0.995 for specific classes. The findings highlight the effectiveness of integrating attention mechanisms and spatial transformations to enhance YOLO-based detection, contributing to improved automated inspection of the power grid.

1. Introduction

Electrical insulators play a critical role in the reliable operation of power transmission systems by preventing the unwanted flow of current to the ground and ensuring system safety [1]. However, environmental factors such as pollution, weathering, and mechanical stress can cause insulators to degrade or fail over time, leading to power outages, equipment damage, and safety hazards [2]. Timely and accurate detection of faults in insulators is therefore essential for maintaining the health of the electrical infrastructure [3].

Traditional methods, such as manual visual inspections and helicopter patrols, while still in use, are increasingly being supplemented or replaced by advanced technologies, including Unmanned Aerial Vehicles (UAVs) [4], thermal imaging [5], infrared [6], light detection and ranging [7], and machine learning algorithms [8]. These innovations enable more precise fault detection [9], condition monitoring [10], and predictive maintenance [11], reducing downtime and operational costs. Digital inspection methods facilitate real-time data acquisition and remote diagnostics, which are essential for modern smart grid management [12]. The shift toward automated inspection not only improves the speed and accuracy of grid assessments but also enhances safety by minimizing human exposure to high-risk environments [13]. The continuous development and implementation of inspection technologies are indispensable for maintaining the integrity and resilience of power infrastructure in the face of growing environmental and operational challenges [14].

* Corresponding author at: LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
Email addresses: jpecarvalho@ciencias.ulisboa.pt (J.P.M. Carvalho), stefano.stefenon@isel.pt (S.F. Stefenon), valderi.leithardt@iscte-iul.pt (V.R.Q. Leithardt), laio.seman@ufsc.br (L.O. Seman), kin-choong.yow@uregina.ca (K.-C. Yow), fcofds@usal.es (J.F.D.P. Santana).

https://doi.org/10.1016/j.asej.2026.104067
Received 31 July 2025; Received in revised form 15 October 2025; Accepted 15 February 2026
Available online 24 February 2026

Most of the power grid inspection methods currently used often involve manual visual assessments or the use of specialized equipment, both of which can be time-consuming, labor-intensive, and prone to human error [15]. With the advent of computer vision and Deep Learning (DL), automated image-based inspection methods have emerged as a promising alternative [16]. In this context, the You Only Look Once (YOLO) model shows promising results for insulator fault detection [17]. A great advantage of using DL-based models is their ability to handle big data, which is necessary to address current engineering problems [18].

A challenging task in the application of DL models, especially YOLO, is the model definition, since several versions have been released [19]. Each model can have a better result depending on the database used and the specific task, since there is a trade-off between increasing performance by using larger models that can be trained with higher data non-linearities and reducing computational effort [20]. Some models show promising results for identifying faults in insulators, but they have difficulty in generalization, being adjusted for specific problems, and may have difficulties in being applied to other tasks [21].

Considering these challenges, this paper studies different model structural variations such as input attention, squeeze and excitation, and spatial transformer. Based on the variations in YOLO, the architecture has the following contributions:

- Input attention module that enhances the network's focus on critical regions of the image, thereby improving detection accuracy.
- The squeeze and excitation modules refine feature maps by reweighting channels to highlight important information while suppressing less useful signals.
- The spatial transformer modules allow dynamic adjustment of spatial features, enabling the model to handle varying orientations and scales.

This study enhances electrical grid reliability and safety by improving automated fault detection in insulators, reducing outages, equipment failures, and risks to workers. It also lowers maintenance costs, minimizes human exposure to hazards, and supports smarter, more resilient grids, ultimately boosting energy security, reducing economic losses, and promoting sustainable utility management.

Despite advances in automated fault detection, deploying artificial intelligence in safety-critical infrastructure raises concerns about reliability, data bias, and the risks of false or missed detections. It stresses the need for human oversight and ethical considerations to ensure trustworthy, accountable, and socially aligned implementation.

The remainder of this paper is as follows: Section 2 presents related works focused on the discussion of YOLO applications. Section 3 presents the compared methods. Section 4 presents the results and discussions, and finally, Section 5 presents a conclusion and suggestions for future research. For standardization purposes, Table 1 presents the acronyms used in this paper.

Table 1
List of acronyms.

Acronym	Definition
CFPNet	Centralized Feature Pyramid Network
DETR	DEtection TRansformer
DFKD	Dynamic-Focused Knowledge Distillation
DL	Deep Learning
DRR	Dilated Re-parameterized Residual
DSC	Depth Separable Convolution
EMA	Exponential Moving Average
EVCBlock	Explicit Visual Center Block
FC	Fully Connected
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
GCA	Group Collaborative Attention
GConv	Global Convolution
GELU	Gaussian Error Linear Unit
GPF	Global Pooling Fusion
HAM	Hybrid Attention Mechanisms
HWU	Haar Wavelet Upsampling
IAT	Input Attention Transformer
IoU	Intersection over Union
mAP	mean Average Precision
MSA	Multi-Scale Attention
MSFFM	Multi-Scale Feature Fusion Module
MSIF	Multi-Scale Information Fusion
ReLU	Rectified Linear Unit
SAE	Squeeze-And-Excitation
SATS	Self-Attention Transformer Structure
SAE	Squeeze-And-Excitation
STD	STandard model
STN	Spatial Transformer Network
TP	True Positive
YOLO	You Only Look Once

2. Related works

The application of DL methods for insulator defect detection has gained considerable attention in recent years, with various approaches addressing the challenges of automated power line inspection through unmanned aerial vehicles and computer vision techniques [22]. The YOLO architecture has emerged as the predominant framework for this task due to its real-time processing capabilities and balanced accuracy–speed performance [23].

Several state-of-the-art techniques have been applied to improve the quality of the electrical system [24], such as those presented by Zheng et al. [25] using a decentralized mechanism for privacy-preserving computation in smart grid, or in [26], where a federated learning-based digital twin framework is applied. In [27], the Finite Element Method (FEM) is used to develop an optimal design for power grid spacers. In [28], the FEM is applied to evaluate the design of power grid insulators. Other models, such as those presented in [29], use thermal images, which can help in fault diagnosis. In [30], the fault diagnosis is based on acoustic analysis.

Chen et al. [31] implemented a YOLOv8n-based system for hexacopter inspections, achieving 99.4% mean Average Precision (mAP)@[0.5] on a dataset of 6020 insulator images across four defect categories: normal, broken, polluted, and flashover surfaces. Their approach incorporated image augmentation techniques and real-time onboard processing using Raspberry Pi 4 hardware. Similarly, Liu et al. [32] introduced an enhanced YOLOv8 architecture incorporating a C2f-Faster-Exponential Moving Average (EMA) module that replaces the original C2f backbone component. Their model achieved 91.5% mAP@[0.5] with 5.66M parameters and 21.1 giga floating-point operations per second computational requirements, demonstrating detection speeds of 113 frames per second through the integration of FasterNet for parameter reduction and EMA-based attention mechanisms.

Attention mechanisms have been extensively explored to improve feature representation learning in insulator detection tasks. Wang et al. [33] incorporated Contextual Transformer modules into YOLOv8 backbone architecture, achieving 97.5% precision and 86.2% mAP through enhanced contextual understanding and Partial Convolution layers for computational efficiency. Zhang et al. [34] developed modified YOLO with Global Convolution (GConv) modules for spatial and channel information integration, achieving 90.9% AP with 90 frames per second detection speed. Their approach included C3-Global Pooling Fusion (GPF) and MultiScale Information Fusion (MSIF) modules for improved feature extraction in complex backgrounds.

Alternative architectural modifications have focused on feature fusion and loss function improvements. Yang et al. [35] proposed the multi-fault insulator using YOLO, replacing the C2F network with MSA-GhostBlock components and implementing ResPANet for multi-scale feature fusion, achieving 93.9% mean accuracy. Wang et al. [36] introduced an insulator fault method based on YOLOv10n, incorporating

Haar Wavelet Upsampling (HWU) and Group Collaborative Attention (GCA) mechanisms, resulting in 94.6% detection accuracy at 170.7 frames per second.

Knowledge distillation techniques have been employed to address deployment challenges on resource-constrained edge devices. Li et al. [37] proposed Dynamic-Focused Knowledge Distillation (DFKD) with dual focus weight factors and adaptive sample matching, combined with model pruning to create a new YOLO architecture for edge deployment scenarios. Zhou et al. [38] explored transformer-based approaches with AdIn-DEtection TRansformer (DETR), incorporating Gaussian saliency guidance adapters and achieving 95.4% and 96.1% AP50 accuracy with R50 and R101 backbones, respectively.

Recent works have also addressed multi-scale detection challenges and complex background scenarios. Xu et al. [39] proposed modifications to YOLOX-s incorporating Self-Attention Transformer Structures (SATS) and Fully Connected Feature Pyramid Networks (FC-FPN), achieving 82.00% average detection accuracy. Lu et al. [40] developed Dilated Re-parameterized Residual (DRR)-YOLO with dilated reparameterized residual modules and Large separable kernel attention Spatial Pyramid Pooling Fast, reaching 94.7% mAP while maintaining computational efficiency. In recent years, modified YOLO models have become increasingly prominent in this domain, including methods such as the insulator lack-global attention mechanism in YOLO proposed by Zhang et al. [41], the Explicit Visual Center Block (EVCBlock) module from Centralized Feature Pyramid Network (CFPNet) presented by Ding et al. [42], the Hybrid Attention Mechanisms (HAM), regularization, and Depth Separable Convolution (DSC) for YOLOX introduced by Li et al. [43], and YOLO11 focused on a Multi-Scale Feature Fusion Module (MSFFM) presented by Shen et al. [44].

While these approaches have demonstrated improvements in detection accuracy and computational efficiency, the integration of channel attention mechanisms at the feature pyramid level remains underexplored. Our work addresses this gap by incorporating Squeeze-and-Excitation blocks into the YOLO detection head to enhance channel-wise feature representation learning for improved insulator defect detection performance.

Table 2 summarizes the key characteristics and performance metrics of recent insulator defect detection methods, highlighting the diverse approaches and their respective contributions to the field.

3. Methodology

This paper presents a study of the structural variations of the YOLO network. The modifications are input attention, squeeze and excitation blocks, and spatial transformer modules. These modifications are explained in this section.

3.1. Input attention transformer enhanced YOLO architecture

We introduce an Input Attention Transformer (IAT) module into the YOLO architecture to enhance spatial feature representation learning through self-attention mechanisms (see Fig. 1). The modification targets the detection head's deepest feature level to improve multi-scale object detection performance by capturing long-range spatial dependencies and contextual relationships within feature maps.

3.1.1. Input attention transformer formulation

The IAT module implements a spatial self-attention mechanism that recalibrates feature maps by modeling spatial interdependencies across all spatial locations. Given an input feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width respectively, the IAT module performs a sequence of operations based on the Transformer architecture [50].

The module first projects the input features into an embedding space through a convolutional transformation:

$$\mathbf{X}_{emb} = \mathrm{Conv}_{1\times1}(\mathbf{X}) \tag{1}$$

where $\mathbf{X}_{emb} \in \mathbb{R}^{D\times H\times W}$ and D represents the embedding dimension.

The spatial feature maps are then reshaped and transposed to create sequence representations suitable for multi-head self-attention:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathrm{Reshape}(\mathbf{X}_{emb}) \in \mathbb{R}^{HW\times B\times D} \tag{2}$$

where B is the batch size, and the spatial dimensions are flattened into sequence length HW.

Multi-head self-attention is applied to capture spatial relationships:

$$\mathbf{A} = \mathrm{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{3}$$

$$\mathbf{X}_{attn} = \mathrm{LayerNorm}(\mathbf{X}_{seq} + \mathbf{A}) \tag{4}$$

where $\mathbf{X}_{seq}$ represents the input sequence and $\mathbf{A}$ contains the attention-weighted features.

A position-wise feed-forward network with Gaussian Error Linear Unit (GELU) activation enhances the representation:

$$\mathbf{F} = \mathrm{Linear}_{D\to D_{ff}}(\mathbf{X}_{attn}) \tag{5}$$

$$\mathbf{F}_{out} = \mathrm{Linear}_{D_{ff}\to D}(\mathrm{GELU}(\mathbf{F})) \tag{6}$$

$$\mathbf{X}_{ff} = \mathrm{LayerNorm}(\mathbf{X}_{attn} + \mathbf{F}_{out}) \tag{7}$$

where D_{ff} represents the feed-forward dimension.

Finally, the enhanced features are projected back to the original channel space and reshaped into a spatial format:

$$\tilde{\mathbf{X}} = \mathrm{Conv}_{1\times1}(\mathrm{Reshape}(\mathbf{X}_{ff})) \tag{8}$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$ represents the attention-enhanced feature maps.

Table 2
Summary of recent insulator defect detection methods.

Author	Base architecture	Key modifications	Max. mAP
Zhou et al. [38]	DETR	GSG-Adapter, LFO-Adapter	0.961
Xu et al. [39]	YOLOX	SATS, FC-FPN, selective kernel network	0.820
Li et al. [43]	YOLOX	HAM, Regularization, and DSC	0.913
Qiu et al. [45]	YOLOv4	Augmentation method based on GraphCut	0.973
Gao et al. [46]	YOLOv5	Transfer learning, PINet integration	0.960
Zhang et al. [34]	YOLOv5	GConv, C3-GPF, MSIF	0.909
Zhang et al. [41]	YOLOv5	EVCBlock module from CFPNet	0.942
Li et al. [47]	YOLOv7	Attention alignment multiscale adversarial domain adaptation	0.994
Chen et al. [31]	YOLOv8	Image augmentation, embedded system	0.994
Liu et al. [32]	YOLOv8	C2f-Faster-EMA, BiFPN-P, Inner-IoU	0.915
Wang et al. [33]	YOLOv8	Contextual Transformer, Partial Conv	0.862
Yang et al. [35]	YOLOv8	MSA-GhostBlock, ResPANet	0.939
Lu et al. [40]	YOLOv8	DRR module, Spatial Pyramid Pooling Fast, Inner-MPDIoU	0.947
Li et al. [48]	YOLOv9	Context anchor concat and C2 locality-aware attention	0.998
Mahapatra et al. [49]	YOLOv9	Fast gradient sign method and projected gradient descent	0.965
Wang et al. [36]	YOLOv10	HWU, GCA, Pyramid Bottleneck	0.946
Shen et al. [44]	YOLOv11	Attention mechanism and a MSFFM	0.915

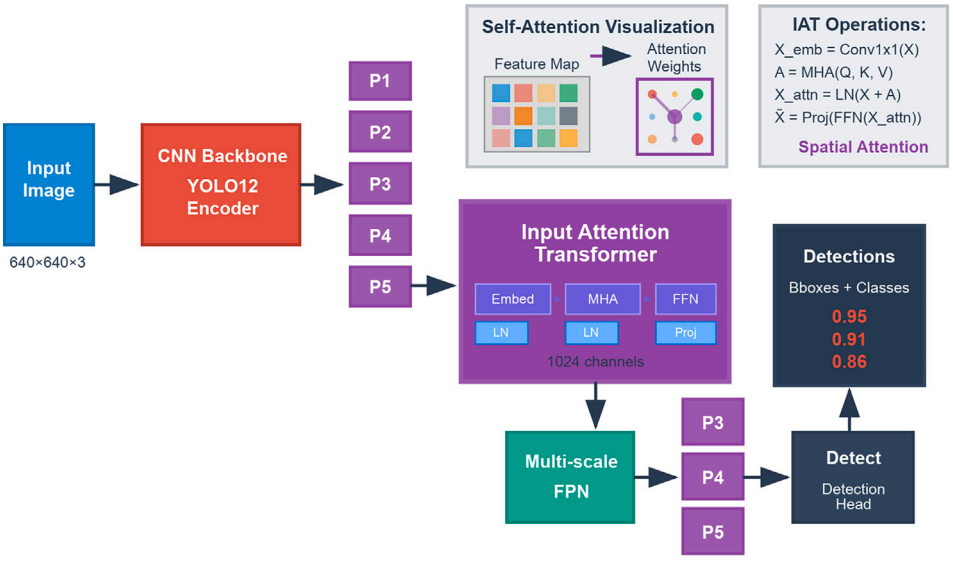

Fig. 1. YOLO with input attention transformer enhancement architecture.

3.1.2. Architectural integration

The IAT module is strategically positioned in the YOLO12 detection head immediately after the deepest backbone features (P5/32) and before the feature pyramid network upsampling operations. The module operates on feature maps with dimensionality $\mathbb{R}^{1024\times H/32\times W/32}$ extracted from the backbone network.

The modified detection head follows the sequence:

$$\mathbf{F}_{P5} = \text{Backbone}(\mathbf{I}) \tag{9}$$

$$\mathbf{F}_{IAT} = \text{InputAttentionTransformer}(\mathbf{F}_{P5}) \tag{10}$$

$$\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\} = \text{FPN}(\mathbf{F}_{IAT}) \tag{11}$$

where $\mathbf{I}$ represents the input image, $\mathbf{F}_{P5}$ denotes the deepest backbone features, $\mathbf{F}_{IAT}$ represents the attention-enhanced features, and $\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\}$ are the multi-scale features used for detection.

We configure the transformer with embedding dimension $D = 512$, single attention head ($h = 1$), and feed-forward dimension $D_{ff} = 1024$. This configuration introduces approximately $2 \times D^2 + 2 \times D \times D_{ff}$ additional parameters. The single-head attention design prioritizes computational efficiency while maintaining the capacity to capture global spatial relationships.

The IAT placement before feature pyramid upsampling ensures that learned spatial attention patterns propagate through all detection scales, potentially improving detection consistency across the multi-scale pyramid. This design allows the network to adaptively focus on spatially relevant regions that contribute most significantly to object detection across different scales and object categories, while capturing long-range dependencies that traditional convolutional operations cannot effectively model.

3.2. Squeeze-and-excitation enhanced YOLO architecture

We incorporate Squeeze-And-Excitation (SAE) blocks [51] into the YOLO architecture to improve channel-wise feature representation learning (see Fig. 2). The modification targets the feature pyramid network head to enhance multi-scale object detection performance through adaptive channel attention mechanisms.

3.2.1. Squeeze-and-excitation block formulation

The SAE block implements a channel attention mechanism that recalibrates feature maps by explicitly modeling interdependencies between channels. Given an input feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width, respectively, the SAE block performs three sequential operations [52].

The first operation aggregates global spatial information through global average pooling to generate channel-wise statistics:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{12}$$

where z_c represents the globally average-pooled feature for channel c, and $x_c(i, j)$ denotes the activation at spatial location (i, j) in channel c.

Channel-wise dependencies are then captured through a bottleneck architecture consisting of two fully connected layers:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \mathbf{z})) \tag{13}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learned parameter matrices, r is the reduction ratio, δ represents the ReLU activation function, and σ denotes the sigmoid activation function. The output $\mathbf{s} \in \mathbb{R}^C$ contains channel-wise attention weights [53].

Finally, the original feature maps are recalibrated through channel-wise multiplication:

$$\tilde{\mathbf{x}}_c = s_c \cdot \mathbf{x}_c \tag{14}$$

where $\tilde{\mathbf{x}}_c$ represents the recalibrated feature map for channel c.

3.2.2. Architectural integration

The SAE block is integrated into the YOLO12 detection head at the deepest feature level (P5/32) before multi-scale feature fusion. The block operates on feature maps with dimensionality $\mathbb{R}^{1024\times H/32\times W/32}$ extracted from the backbone network.

The modified detection head follows the sequence:

$$\mathbf{F}_{P5} = \text{Backbone}(\mathbf{I}) \tag{15}$$

$$\mathbf{F}_{SAE} = \text{SAEBlock}(\mathbf{F}_{P5}) \tag{16}$$

$$\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\} = \text{FPN}(\mathbf{F}_{SAE}) \tag{17}$$

where $\mathbf{I}$ represents the input image, $\mathbf{F}_{P5}$ denotes the deepest backbone features, $\mathbf{F}_{SAE}$ represents the attention-enhanced features, and $\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\}$ are the multi-scale features used for detection.

We set the reduction ratio $r = 1$ to maintain full channel dimensionality in the bottleneck layer, preserving the representational capacity of

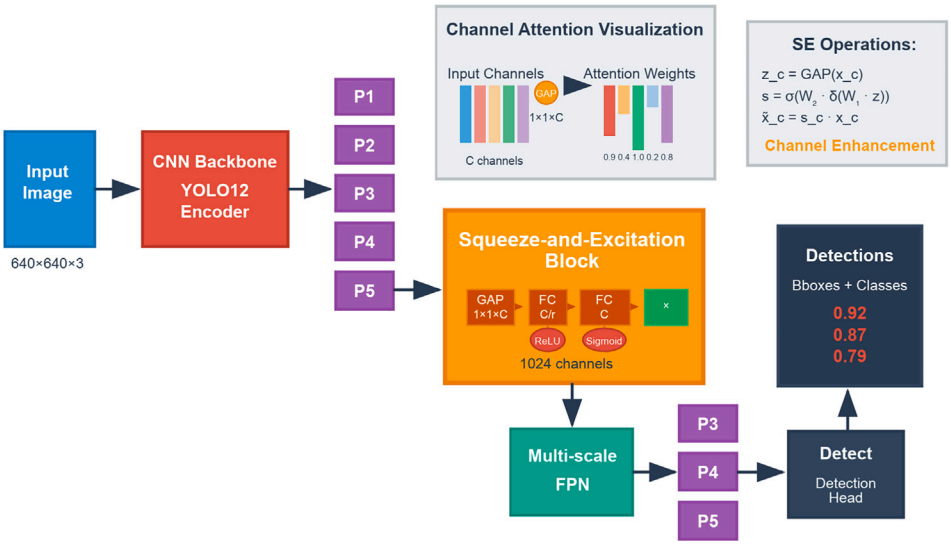

Fig. 2. YOLO with squeeze-and-excitation enhancement architecture.

the attention mechanism. This configuration introduces approximately $2 \times C^2$ additional parameters, where $C = 1024$ for the P5 feature level. The SAE block placement before feature pyramid upsampling ensures that learned channel attention weights propagate through all detection scales, potentially improving detection consistency across the multi-scale pyramid. This design allows the network to adaptively emphasize channels that contribute most significantly to object detection across different scales and object categories.

3.3. Spatial transformer network enhanced YOLO architecture

We incorporate a Spatial Transformer Network (STN) module into the YOLO architecture to improve geometric invariance and spatial feature alignment through learnable affine transformations (see Fig. 3). The modification targets the detection head's deepest feature level to enhance multi-scale object detection performance by enabling the network to actively transform feature maps to achieve better spatial alignment for object localization and classification.

3.3.1. Spatial transformer network formulation

The STN module implements a differentiable attention mechanism that performs explicit spatial transformations on feature maps by learning optimal affine transformation parameters. Given an input feature tensor $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width respectively, the STN module performs three sequential operations: localization, grid generation, and sampling [54].

The localization network first extracts spatial transformation parameters through a series of convolutional and fully connected layers:

$$\mathbf{F}_{loc} = \text{Conv}_{7\times7}(\mathbf{X}) \tag{18}$$

$$\mathbf{F}_{pool} = \text{MaxPool}_{2\times2}(\text{ReLU}(\mathbf{F}_{loc})) \tag{19}$$

$$\mathbf{F}_{adapt} = \text{AdaptiveAvgPool}_{28\times28}(\text{ReLU}(\mathbf{F}_{pool})) \tag{20}$$

The feature maps are then flattened and processed through fully connected layers to predict affine transformation parameters:

$$\mathbf{F}_{flat} = \text{Flatten}(\mathbf{F}_{adapt}) \in \mathbb{R}^{8\times28\times28} \tag{21}$$

$$\theta = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{F}_{flat} + \mathbf{b}_1) + \mathbf{b}_2 \tag{22}$$

where $\mathbf{W}_1 \in \mathbb{R}^{32\times6272}$, $\mathbf{W}_2 \in \mathbb{R}^{6\times32}$, and $\theta \in \mathbb{R}^6$ represent the affine transformation parameters.

The transformation parameters are reshaped into a 2×3 affine transformation matrix:

$$\mathbf{T}_\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \tag{23}$$

where the matrix is initialized to the identity transformation $\mathbf{T}_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ to ensure stable training initialization.

A sampling grid is generated using the predicted transformation matrix:

$$\mathbf{G} = \text{AffineGrid}(\mathbf{T}_\theta, \text{size}(\mathbf{X})) \tag{24}$$

where $\mathbf{G} \in \mathbb{R}^{H\times W\times2}$ contains the transformed coordinate mappings for each spatial location.

Finally, the transformed feature maps are obtained through differentiable bilinear sampling:

$$\tilde{\mathbf{X}} = \text{GridSample}(\mathbf{X}, \mathbf{G}) \tag{25}$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$ represents the spatially transformed feature maps [55].

3.3.2. Architectural integration

The STN module is strategically positioned in the YOLO12 detection head immediately after the deepest backbone features (P5/32) and before the feature pyramid network upsampling operations. The module operates on feature maps with dimensionality $\mathbb{R}^{1024\times H/32\times W/32}$ extracted from the backbone network.

The modified detection head follows the sequence:

$$\mathbf{F}_{P5} = \text{Backbone}(\mathbf{I}) \tag{26}$$

$$\mathbf{F}_{STN} = \text{SpatialTransformer}(\mathbf{F}_{P5}) \tag{27}$$

$$\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\} = \text{FPN}(\mathbf{F}_{STN}) \tag{28}$$

where $\mathbf{I}$ represents the input image, $\mathbf{F}_{P5}$ denotes the deepest backbone features, $\mathbf{F}_{STN}$ represents the spatially transformed features, and $\{\mathbf{F}_{P3}, \mathbf{F}_{P4}, \mathbf{F}'_{P5}\}$ are the multi-scale features used for detection.

The localization network is configured with 8 feature channels after the initial 7×7 convolution, followed by adaptive average pooling to a fixed 28×28 spatial resolution. The fully connected layers compress the spatial information from 6272 dimensions to 32 intermediate features

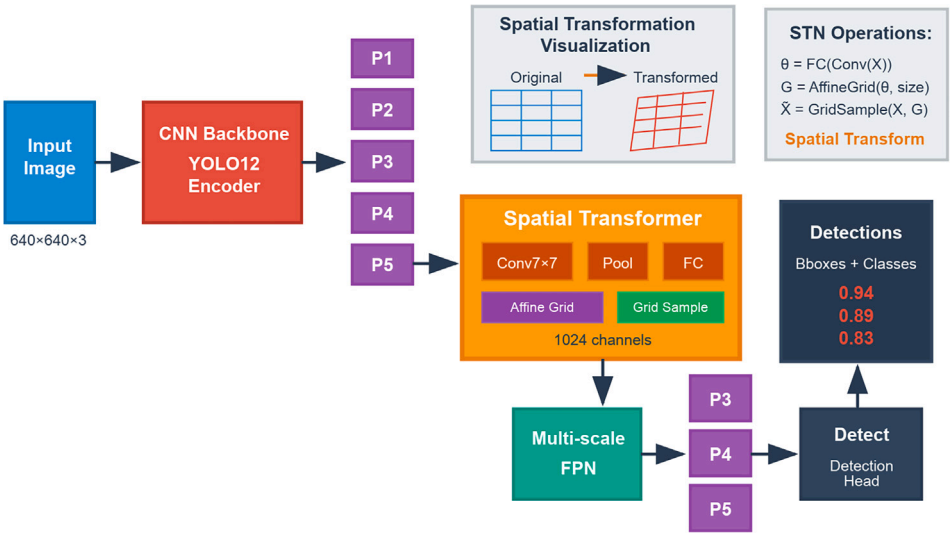

Fig. 3. YOLO with spatial transformer network enhancement architecture.

before predicting the 6 affine parameters. This configuration introduces approximately $6,272 \times 32 + 32 \times 6 = 200,896$ additional parameters for the transformation parameter prediction.

The STN placement before feature pyramid upsampling ensures that learned spatial transformations propagate through all detection scales, potentially improving detection consistency across the multi-scale pyramid. This design allows the network to actively correct for geometric distortions, scale variations, and spatial misalignments that may occur in complex detection scenarios, while maintaining differentiability throughout the transformation process for end-to-end training.

4. Results and discussion

This section first presents the dataset used for experimental validation, and then provides an in-depth analysis of detection performance across different fault classes. Next, a comparative assessment is conducted among the structural variations applied to multiple YOLO architectures and model scales, namely IAT-YOLO, SAE-YOLO, and STN-YOLO. Finally, the models are evaluated against STandard YOLO (STD-YOLO) variants in various detection scenarios.

4.1. Dataset

The dataset used in this study was sourced from the open-access Roboflow platform [56]. It consists of high-resolution annotated images capturing electrical insulators in various conditions, commonly observed during power line inspections.

The dataset encompasses seven distinct classes: two insulator columns; dirty glass surface; glass missing; contaminated polymer surface; broken disc; standard insulator reference; and flashover damage. These classes represent a comprehensive set of operational and defective states of insulators, including surface contamination, structural damage, and partial loss, thereby supporting multi-class fault detection under realistic conditions.

Each image is annotated using bounding boxes in YOLO format, enabling supervised learning for object detection tasks. The dataset was randomly divided into training (80%), validation (10%), and testing (10%) subsets. Image preprocessing included normalization, resizing to 640×640 pixels, and data augmentation strategies such as flipping, brightness adjustment, and rotation to enhance model robustness. Representative samples from the dataset are illustrated in Fig. 4, highlighting the diversity and complexity of visual patterns across classes.

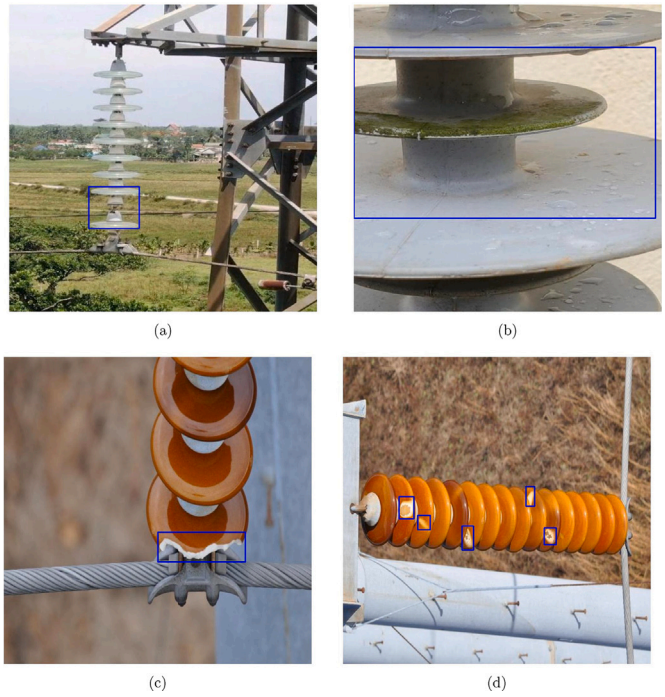

Fig. 4. Samples from the dataset used in this study: (a) glass missing; (b) contaminated polymer surface; (c) broken disc; and (d) flashover damage.

Fig. 4(a), a glass component is partially missing, revealing structural loss that may compromise insulation performance. Fig. 4(b) illustrates a polymer insulator exhibiting surface dirt or contamination, which typically leads to degraded dielectric properties. Fig. 4(c) presents a broken disc with clear fragmentation, posing a critical risk for mechanical failure. Finally, Fig. 4(d) captures the aftermath of a pollution-induced flashover event, identifiable by surface burn marks and carbonization patterns.

4.2. Experimental setup

The experiments were conducted using a workstation equipped with 8×RTX 5000 graphics processing units, each with 32 GB of memory.

Table 3
Insulator detection results based on structure variations (class 0).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.99892	**1.00000**	**0.99500**	0.97398	4.56
	s	0.99897	**1.00000**	**0.99500**	0.97848	5.92
	m	0.99898	**1.00000**	**0.99500**	0.97996	8.01
	l	0.99892	**1.00000**	**0.99500**	0.97046	12.23
	x	0.83109	**1.00000**	0.98182	0.89120	32.11
IAT-YOLO	n	0.99896	**1.00000**	**0.99500**	0.97381	3.69
	s	0.00000	0.00000	0.00000	0.00000	**1.90**
	m	0.99902	**1.00000**	**0.99500**	0.97756	7.40
	l	0.98797	0.92308	0.99275	0.93613	34.41
	x	0.66482	**1.00000**	0.95755	0.61176	44.06
SAE-YOLO	n	0.99888	**1.00000**	**0.99500**	0.98056	3.60
	s	0.99887	**1.00000**	**0.99500**	0.98098	3.89
	m	0.99895	**1.00000**	**0.99500**	0.97950	6.54
	l	0.99896	**1.00000**	**0.99500**	0.97646	8.29
	x	0.99892	**1.00000**	**0.99500**	0.97935	17.53
STN-YOLO	n	0.99892	**1.00000**	**0.99500**	0.97447	3.18
	s	0.99899	**1.00000**	**0.99500**	0.98128	4.65
	m	**1.00000**	**1.00000**	**0.99500**	**0.98513**	9.44
	l	0.99664	**1.00000**	**0.99500**	0.97685	8.70
	x	0.99896	**1.00000**	**0.99500**	0.97815	13.79

Best results in bold.

All models were implemented using Python and PyTorch, with training and evaluation performed in a Linux-based environment. The total processing time reported includes both the training and inference stages for each YOLO-based model across all dataset splits.

The evaluation focused on object detection performance for insulator fault identification and classification. Accordingly, standard metrics from the object detection domain were adopted: Precision, Recall, F1-score, and mAP at two IoU thresholds: @0.5 and @0.5:0.95, defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{29}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{30}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{31}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The mAP metric was computed as the average of interpolated precision values over recall levels for each class, with aggregation across classes [57].

Evaluations were conducted separately per class and per model scale (n, s, m, l, x) for standard YOLO and each structural variation (STN-YOLO, SAE-YOLO, IAT-YOLO). To assess generalization and robustness, each model was evaluated on a fixed test set with balanced representation of the seven classes: two insulator columns, dirty glass, glass missing, contaminated polymer, broken disc, standard insulator, and pollution flashover.

All samples were resized to 640×640 pixels, and the same augmentation techniques were applied during training across models to ensure comparability. Finally, the results from the individual STN-YOLO, SAE-YOLO, and IAT-YOLO models were compared against the latest YOLOv10, YOLOv11, and YOLOv12 architectures to assess improvements in detection accuracy and efficiency across multiple scales.

Table 3 presents the insulator fault detection performance for class 0, which corresponds to the two-insulator-column category. Results are reported for four architectural configurations: standard YOLO, in short STD-YOLO, and the structural variants based on Spatial Transformer Networks (STN-YOLO), Input Attention Transformers (IAT-YOLO), and Squeeze-and-Excitation blocks (SAE-YOLO), each evaluated across five model sizes: n, s, m, l, and x.

Across all architectures, high precision and recall values were observed, particularly for the STD-YOLO, STN-YOLO, and SAE-YOLO variants. Most models achieved near-perfect results, with precision and recall values close to 1.000 and mAP@[0.5] consistently fixed at 0.995, indicating extremely reliable detection performance for class 0. The mAP@[0.5:0.95] metric revealed subtle differences between models, highlighting the discriminative capability of structural enhancements at varying IoU thresholds.

The STN-YOLO variant demonstrated the highest mAP@[0.5:0.95] overall, with the m-sized model achieving 0.98513, slightly outperforming its standard YOLO counterpart (0.97996) and SAE-YOLO (0.97950). The SAE-YOLO variant also performed robustly across all scales, with the x-sized model reaching a mAP@[0.5:0.95] of 0.97935 and maintaining high consistency across scales. This indicates that channel-wise attention via squeeze-and-excitation blocks improves the detection head without overfitting.

The IAT-YOLO variant, while achieving strong performance in most sizes, revealed a clear case of non-convergence in the s-sized model, where all metrics were zero. This likely indicates a training failure or unstable gradient propagation due to the attention mechanism at this scale. Despite that, the IAT-YOLO-m and IAT-YOLO-n configurations still performed competitively, showing mAP@[0.5:0.95] values of 0.97756 and 0.97381, respectively. However, the performance of larger models such as IAT-YOLO-l and IAT-YOLO-x declined more noticeably, especially in the x configuration (mAP@[0.5:0.95] = 0.61176), suggesting sensitivity to overparameterization or overfitting in this variant.

Regarding training time, the x-sized models consistently required the most time, particularly in the IAT-YOLO and STN-YOLO configurations (over 13 and 44 hours, respectively), while n and s models completed training in under 6 hours. The STN-YOLO variant offered the best balance between training time and detection accuracy for class 0, particularly in the m configuration. These results indicate that class 0 is effectively detected by all evaluated models, with STN-YOLO-m and SAE-YOLO-x yielding the best performance-complexity trade-offs.

Table 4 presents the detection performance for class 1, corresponding to a dirty or polluted glass surface. Unlike class 0, the results for class 1 reveal greater variability and lower overall accuracy across all models and architectural variants. These results indicate that this class poses a more challenging detection task, due to subtle visual features or intra-class variability.

Table 4
Insulator detection results based on structure variations (class 1).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.63728	0.59246	0.60283	0.35573	9.40
	s	0.50004	0.53498	0.50793	0.26124	10.11
	m	0.83854	0.52326	0.61703	0.36959	12.40
	l	0.81550	0.66819	0.69171	0.45933	19.42
	x	0.63407	0.54402	0.47303	0.26224	41.82
IAT-YOLO	n	0.65528	0.53488	0.53349	0.29995	5.09
	s	0.00000	0.00000	0.00000	0.00000	1.74
	m	0.79003	0.61255	0.66067	0.43568	10.08
	l	0.00000	0.00000	0.00000	0.00000	16.65
	x	0.00000	0.00000	0.00000	0.00000	15.36
SAE-YOLO	n	0.63176	0.63839	0.62375	0.35993	4.56
	s	0.78188	0.58356	0.67900	0.42748	5.68
	m	0.60465	0.67522	0.42587	0.62762	9.06
	l	0.67442	**0.71311**	0.47395	**0.64052**	14.56
	x	0.83141	0.65116	0.69323	0.46950	19.94
STN-YOLO	n	0.00000	0.00000	0.00000	0.00000	**1.11**
	s	0.71424	0.58140	0.63456	0.39060	5.83
	m	0.78949	0.63953	0.68803	0.46026	9.77
	l	0.76554	0.62791	0.65055	0.46175	14.55
	x	**0.85502**	0.68574	**0.73552**	0.50617	22.58

Best results in bold.

Among the STD-YOLO models, the best-performing configuration was the l-sized model, achieving a precision of 0.81550, a recall of 0.66819, an mAP@[0.5] of 0.69171, and an mAP@[0.5:0.95] of 0.45933. Despite moderate gains in detection accuracy, the STD-YOLO-x variant underperformed significantly with mAP@[0.5:0.95] dropping to 0.26224, concluding that larger model scales do not necessarily translate to better performance for this class.

The STN-YOLO models consistently outperformed the STD-YOLO baseline in this class. The x-sized STN-YOLO model achieved the highest values across all metrics: precision equals 0.85502, recall equals 0.68574, mAP@[0.5] equals 0.73552, and mAP@[0.5:0.95] equals 0.50617, clearly outperforming all other configurations. For the IAT-YOLO variant, results were mixed. While the m-sized model achieved reasonable performance (precision equals 0.79003, mAP@[0.5:0.95] equals 0.43568), the s, l, and x configurations didn't converge, as indicated by zero values across all metrics.

The SAE-YOLO variant demonstrated stable and competitive results. The x-sized SAE-YOLO model had strong performance (precision equal to 0.83141, mAP@[0.5:0.95] equal to 0.46950), closely following the best results achieved by STN-YOLO-x. Additionally, the S-sized SAE-YOLO model showed good generalization with low training time, offering a favorable balance between computational cost and accuracy. Overall, the STN-YOLO-x model was the top performer for class 1, with SAE-YOLO-x and SAE-YOLO-l also showing high efficacy.

Table 5 reports the detection performance for class 2, which corresponds to the glass loss or missing glass condition. This class presents a moderate level of detection difficulty, based on intermediate performance metrics across the evaluated YOLO-based structures and scales. Among the STD-YOLO models, the m-sized model achieved the best overall performance within this group, reaching a precision of 0.84893, a recall of 0.54286, an mAP@[0.5] of 0.62707, and an mAP@[0.5:0.95] of 0.37820. However, despite larger configurations such as STD-YOLO-x having slightly higher precision (0.85863), they did not yield significant improvements in recall or mAP.

The STN-YOLO variant achieved the best detection results for class 2 overall. The STN-YOLO-l model achieved the top mAP@[0.5] and mAP@[0.5:0.95] values (0.71230 and 0.45427, respectively), combined with a balanced precision of 0.86354 and recall of 0.60000. The STN-YOLO-m model also performed well, with mAP@[0.5] of 0.64718. In contrast, the IAT-YOLO configuration showed significant training instabilities. Models of size s, m, and l completely failed to converge (metrics equal to 0.00000), while only the x-sized IAT-YOLO converged with a mAP@[0.5] of 0.57317 and mAP@[0.5:0.95] of 0.36344. The SAE-YOLO-l and SAE-YOLO-x models were the top-performing models for class 2, with the highest detection precision and mAP scores. SAE-YOLO-l reaching a mAP@[0.5] of 0.71984 (highest among all configurations) and mAP@[0.5:0.95] of 0.45423.

Table 6 summarizes the detection performance for class 3, which refers to the presence of contamination on polymer insulator surfaces. This class exhibits robust performance across all evaluated model configurations. The STD-YOLO models already achieved strong baselines, particularly the l-sized STD-YOLO model, which obtained a precision of 0.91686, a recall of 0.89120, mAP@[0.5] of 0.93267, and mAP@[0.5:0.95] of 0.68646. These results reflect both high detection confidence and consistency across IoU thresholds. However, the performance of the x-sized model slightly decreased in the recall and mAP metrics.

The STN-YOLO variants consistently improved the detection accuracy for this class. While STN-YOLO-l and STN-YOLO-x both had the highest mAP@[0.5] scores (0.88703 and 0.91110, respectively), the x-sized STN-YOLO model stood out with a strong balance between precision (0.88847) and recall (0.86869), and a high mAP@[0.5:0.95] of 0.67350.

The IAT-YOLO models achieved better results at the smaller scales (n, s, m), with the s-sized IAT-YOLO precision equal to 0.92265, recall equal to 0.77778, and mAP@[0.5] equal to 0.86361. However, a drop in performance occurred for the l and x configurations in recall and mAP@[0.5:0.95].

The SAE-YOLO models were the top performers for this class. The SAE-YOLO-l model reached the highest mAP@[0.5:0.95] of 0.69063, while the SAE-YOLO-n model achieved the best precision overall (0.94356). Moreover, the x-sized SAE-YOLO model achieved strong overall scores across all metrics, with mAP@[0.5] equal to 0.91927 and mAP@[0.5:0.95] equal to 0.66960.

Table 7 presents the detection results for class 4, which corresponds to the condition of a broken disc in insulators. Overall, the models demonstrate strong performance in identifying this fault type, with precision and recall values generally exceeding 0.85 across most configurations.

Table 5
Insulator detection results based on structure variations (class 2).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.65218	0.53593	0.56152	0.29810	9.32
	s	0.88055	0.48571	0.56651	0.31528	9.03
	m	0.84893	0.54286	0.62707	0.37820	12.18
	l	0.81273	0.57143	0.59283	0.36983	17.85
	x	0.85863	0.51429	0.61157	0.36827	51.28
IAT-YOLO	n	0.50056	0.37143	0.40201	0.23243	4.81
	s	0.00000	0.00000	0.00000	0.00000	1.50
	m	0.00000	0.00000	0.00000	0.00000	2.50
	l	0.00000	0.00000	0.00000	0.00000	12.09
	x	0.78958	0.54286	0.57317	0.36344	53.24
SAE-YOLO	n	0.00000	0.00000	0.00000	0.00000	**0.88**
	s	0.83093	0.51429	0.60259	0.36830	5.36
	m	0.80094	0.48571	0.59631	0.34008	9.38
	l	**0.89514**	0.60000	**0.71984**	0.45423	14.30
	x	0.85052	0.57143	0.71355	0.45276	22.55
STN-YOLO	n	0.58663	0.65714	0.59306	0.35317	4.74
	s	0.41503	0.20000	0.21319	0.11769	1.59
	m	0.82169	0.54286	0.64718	0.37551	9.48
	l	0.86354	0.60000	0.71230	**0.45427**	14.34
	x	0.65391	**0.64800**	0.61463	0.33816	22.35

Best results are in bold.

Table 6
Insulator detection results based on structure variations (class 3).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.86896	0.80379	0.86310	0.57291	10.03
	s	0.91632	0.74747	0.87103	0.59221	9.94
	m	0.82800	0.84848	0.80475	0.53537	9.48
	l	0.91686	**0.89120**	**0.93267**	0.68646	16.84
	x	0.89148	0.78788	0.84398	0.59879	49.75
IAT-YOLO	n	0.87308	0.73737	0.80277	0.54186	5.07
	s	0.92265	0.77778	0.86361	0.56831	6.08
	m	0.80170	0.81674	0.87166	0.59610	10.07
	l	0.62612	0.58586	0.64360	0.43467	37.70
	x	0.70882	0.56566	0.63591	0.39451	33.03
SAE-YOLO	n	**0.94356**	0.76768	0.87041	0.60233	**4.55**
	s	0.92290	0.81818	0.90095	0.64272	5.72
	m	0.92803	0.78156	0.89315	0.65168	9.62
	l	0.85091	0.88889	0.92407	**0.69063**	14.54
	x	0.92186	0.87879	0.91927	0.66960	23.22
STN-YOLO	n	0.91765	0.78787	0.85089	0.57570	4.78
	s	0.91216	0.81818	0.88920	0.62798	5.73
	m	0.88354	0.79798	0.87880	0.62382	9.67
	l	0.93532	0.81818	0.88703	0.68306	14.57
	x	0.88847	0.86869	0.91110	0.67350	23.04

Best results in bold.

Within STD-YOLO variants, the s-sized model achieved the highest precision at 0.96597 and a recall of 0.87368, with a mAP@[0.5] of 0.94417 and an mAP@[0.5:0.95] of 0.53449. The l-sized STD-YOLO model was the second-best model in STD-YOLO variants, with a lower precision (0.94670) but higher recall (0.93493), resulting in competitive mAP scores. The STN-YOLO variants showed consistent and competitive performance, with the l-sized STN-YOLO model closely matching the STD-YOLO-l in recall (0.93684) and mAP@[0.5] (0.93403). However, the highest precision in STN-YOLO models was obtained by the s-sized model (0.94359).

The IAT-YOLO models achieved comparable results, particularly the m-sized model, which reached a precision of 0.94177, a recall of 0.92632, and an mAP@[0.5] of 0.94025, matching closely with the top STD-YOLO and STN-YOLO models. Larger-scale IAT-YOLO models maintained strong precision and recall, though with increased training times.

The SAE-YOLO models also demonstrated robust performance, especially the x-sized model, which obtained the highest mAP@[0.5] at 0.93762 and the best mAP@[0.5:0.95] of 0.54963. The SAE-YOLO-m and SAE-YOLO-l models similarly produced high precision and recall metrics. Training times varied as expected, with larger models requiring up to approximately 47 hours for the x-sized variants, while smaller models completed training significantly faster (less than 11 hours), demonstrating practical trade-offs between computational cost and detection accuracy.

Table 8 presents the detection performance for class 5, representing the baseline or standard insulator class. This class shows generally excellent detection results across all model structures and scales. Among the STD-YOLO models, the l-sized STD-YOLO achieved the highest precision of 0.99942, a recall of 0.99387, and maximum mAP@[0.5] and mAP@[0.5:0.95] scores of 0.995 and 0.97108, respectively. The m-sized

Table 7
Insulator detection results based on structure variations (class 4).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.90395	0.77895	0.86528	0.49260	10.68
	s	**0.96597**	0.87368	**0.94417**	0.53449	10.52
	m	0.91346	0.92632	0.93108	0.52965	8.61
	l	0.94670	0.93493	0.94024	0.54007	16.30
	x	0.89563	0.90526	0.91630	0.53528	46.85
IAT-YOLO	n	0.82317	0.83158	0.88503	0.48436	5.06
	s	0.93790	0.90526	0.93723	0.54902	6.11
	m	0.94177	0.92632	0.94025	0.53509	10.10
	l	0.91089	0.85263	0.92817	0.52801	51.75
	x	0.90564	0.90526	0.92675	0.53633	49.35
SAE-YOLO	n	0.89788	0.86316	0.88833	0.51099	**4.60**
	s	0.90566	0.90952	0.90407	0.52014	4.90
	m	0.94699	0.89474	0.92232	0.54024	8.04
	l	0.92323	0.92632	0.92257	0.53823	14.73
	x	0.93471	**0.93684**	0.93762	**0.54963**	18.92
STN-YOLO	n	0.91827	0.85263	0.90775	0.50175	4.74
	s	0.94359	0.88046	0.91762	0.52924	4.68
	m	0.91367	0.91579	0.92330	0.52333	9.71
	l	0.94434	**0.93684**	0.93403	0.54502	14.75
	x	0.90440	0.90526	0.91058	0.53221	18.79

Best results in bold.

Table 8
Insulator detection results based on structure variations (class 5).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.98717	**1.00000**	0.99488	0.96226	9.46
	s	0.99386	0.99322	0.99077	0.95719	9.30
	m	0.99319	**1.00000**	**0.99500**	0.96639	12.06
	l	0.99942	0.99387	**0.99500**	**0.97108**	15.51
	x	0.91776	0.91411	0.96964	0.79654	30.26
IAT-YOLO	n	0.99386	0.99342	0.99494	0.95671	**4.13**
	s	0.99352	0.99387	0.99494	0.97013	6.12
	m	0.99957	**1.00000**	**0.99500**	0.96847	8.29
	l	0.98778	0.99184	0.99451	0.93971	48.61
	x	0.85590	0.98390	0.96509	0.84524	42.71
SAE-YOLO	n	**1.00000**	0.99094	**0.99500**	0.96505	5.04
	s	0.99913	**1.00000**	**0.99500**	0.96723	5.85
	m	0.99390	0.99927	**0.99500**	0.96619	8.33
	l	0.99929	0.99387	**0.99500**	0.96742	15.50
	x	0.99822	0.99387	**0.99500**	0.97022	18.95
STN-YOLO	n	0.99386	0.99294	0.99242	0.95538	4.52
	s	0.98753	0.99387	0.99475	0.96643	5.82
	m	**1.00000**	0.99277	**0.99500**	0.96920	8.00
	l	0.99982	**1.00000**	**0.99500**	0.96989	13.35
	x	0.99879	0.99387	**0.99500**	0.96693	21.72

Best results in bold.

STD-YOLO and s-sized STD-YOLO models also performed near this top level.

The STN-YOLO variants maintained similarly high accuracy. The l-sized STN-YOLO model achieved a perfect recall of 1.000 and mAP@[0.5] of 0.995, while the m-sized and x-sized STN-YOLO models also demonstrated precision and recall values exceeding 0.99, confirming that spatial transformation helps maintain detection robustness. The IAT-YOLO models showed very competitive performance in smaller sizes (n, s, m), with precision and recall close to or equal to 1.00 and mAP@[0.5] around 0.995. However, larger-scale IAT-YOLO models (l and x) showed some reduction in precision and mAP, potentially due to increased model complexity affecting training stability for this class.

The SAE-YOLO models consistently matched or exceeded other configurations, with the n-sized SAE-YOLO reaching perfect precision (1.000) and the s-sized SAE-YOLO attaining perfect recall (1.000) and mAP@[0.5] of 0.995. The SAE-YOLO models generally achieved the highest precision–recall balance with mAP@[0.5:0.95] scores close to 0.97 across scales. The time taken for training correlated with the size of the model, ranging from approximately four to 30 hours. Larger models required more computational resources. Overall, class 5 detection benefits from all architectural variations, with very high accuracy observed across all model sizes. Both STD-YOLO-l, STN-YOLO-l, and SAE-YOLO-n/s/l models exhibit top-tier detection performance.

Table 9 reports the detection performance metrics for class 6, corresponding to the flashover damage in insulators. This class presents more challenging detection scenarios compared to others, as reflected by generally lower mAP values across all model variants and scales. Among the STD-YOLO models, the l-sized STD-YOLO variant achieved the highest precision at 0.87637 and a recall of 0.83219, with mAP@[0.5] of 0.86282 and mAP@[0.5:0.95] of 0.38770. The m-sized STD-YOLO model performed comparably, with a lower precision (0.84420) but a

Table 9
Insulator detection results based on structure variations (class 6).

Structure	Model Size	Precision	Recall	mAP [0.5]	mAP [0.5:0.95]	Training time (h)
STD-YOLO	n	0.79893	0.76011	0.78325	0.34403	9.25
	s	0.85458	0.76957	0.83179	0.37760	9.23
	m	0.84420	**0.86522**	**0.88439**	**0.39872**	11.26
	l	0.87637	0.83219	0.86282	0.38770	16.99
	x	0.79872	0.83478	0.84562	0.39745	36.45
IAT-YOLO	n	0.80215	0.71304	0.77893	0.35702	5.02
	s	0.79951	0.81490	0.84010	0.38229	6.12
	m	0.86111	0.80435	0.86367	0.39812	7.29
	l	0.84471	0.81739	0.85732	0.40017	45.35
	x	**0.87782**	0.84348	0.85883	0.39693	38.92
SAE-YOLO	n	0.81516	0.74777	0.81143	0.36142	4.97
	s	0.84556	0.78557	0.82956	0.37877	5.69
	m	0.84513	0.84348	0.87654	0.36531	8.56
	l	0.86229	0.81739	0.86037	0.39635	14.31
	x	0.85089	0.81874	0.85555	0.38513	16.69
STN-YOLO	n	0.80341	0.73043	0.80646	0.35475	**4.94**
	s	0.82257	0.80870	0.82731	0.37580	5.67
	m	0.84095	0.80460	0.83689	0.37674	6.48
	l	0.83260	0.84336	0.85018	0.39848	15.18
	x	0.82963	0.81304	0.83481	0.39019	17.25

Best results in bold.

higher recall (0.86522), with the best mAP@[0.5] among STD-YOLO models at 0.88439.

The l-sized STN-YOLO model achieves the highest recall among STN-YOLO's variants (0.84336) and a mAP@[0.5] of 0.85018, slightly below the STD-YOLO-l model. Precision values for STN-YOLO models were generally lower than those of STD-YOLO but still competitive. However, the x-sized STN-YOLO model achieved a better precision value compared to STD-YOLO models (0.82963). The IAT-YOLO models achieved the highest precision in this class overall (0.87782), along with a recall of 0.84348 and mAP@[0.5] of 0.85883. Medium-sized IAT-YOLO variants also presented balanced metrics, though smaller sizes showed slightly reduced recall.

In SAE-YOLO variants, the l-sized SAE-YOLO variant showed a precision of 0.86229, recall of 0.81739, and mAP@[0.5] of 0.86037. Training times scaled with model size as expected, ranging from about 5 to 37 hours, with the largest models incurring significantly higher computational costs. In summary, the STD-YOLO-l, STN-YOLO-l, IAT-YOLO-x, and SAE-YOLO-l models provide the best overall balance of precision and recall, indicating that combining spatial and channel attention mechanisms alongside architectural variations can improve detection robustness for pollution-flashover faults.

4.3. Overall analysis

In summary, the analysis showed that all three proposed structural variations improved insulator fault detection in different ways. SAE-YOLO and STN-YOLO delivered consistent gains in accuracy and robustness across model sizes, with STN-YOLO particularly excelling in geometric alignment and SAE-YOLO in channel feature learning. IAT-YOLO demonstrated competitive performance in some configurations but also suffered from instability, especially in smaller and over-parameterized models where training failed to converge. The results confirmed that the proposed enhancements can outperform standard YOLO models, achieving mAP scores up to 0.995 in certain classes, validating their effectiveness for automated power grid inspection tasks.

For a comprehensive comparative analysis between classes, we present Table 10. Class 0 (insulating chains) obtained considerably higher results, reaching an mAP@[0.5] of 0.995. This result was possibly because images of insulator chains represent most of the total image, and there is a greater difference between an insulator chain and specifically defective insulators. Another class that had high results was class 5 (standard insulator reference), with an mAP@[0.5] of up to 0.99494. This result was possibly due to the larger number of samples in this class, which is common for this type of analysis. The class with the lowest mAP results was class 2 (glass missing), with the best performance in just mAP@[0.5] of 0.65391. The other classes had intermediate results, ranging from mAP@[0.5] of 0.67778 for class 1 (dirty glass surface) to mAP@[0.5] of 0.9264 for class 4 (broken disc).

Regarding the model inference time (tested with never-seen images), the models that presented the best results were STD-YOLO and SAE-YOLO. This was also observed in relation to training time, showing that the STD-YOLO and SAE-YOLO models are lighter models, suitable for use in embedded systems.

The proposed research can be extended beyond insulator fault detection. Since the architectural enhancements (input attention, squeeze-and-excitation, and spatial transformer modules) improve spatial awareness, feature selection, and geometric invariance, they apply to a wide range of object detection problems. Potential extensions include infrastructure monitoring (such as bridges, roads, and railways), medical imaging for anomaly detection, industrial quality inspection, agricultural monitoring of crops and livestock, and even autonomous driving, where robust detection under varying conditions is essential. By adapting the models to different datasets and retraining for specific domains, the method can contribute to broader societal applications where accuracy, robustness, and safety are critical.

4.4. Advantages and limitations

The proposed methods introduce clear advantages and some limitations that should be acknowledged. On the positive side, the architectural enhancements, namely the integration of input attention, squeeze-and-excitation, and spatial transformer modules, demonstrate significant improvements in feature representation, robustness to geometric distortions, and overall detection accuracy compared to standard YOLO variants. These modifications also show strong potential for handling complex fault patterns in insulator inspection, with several configurations achieving near-perfect mAP scores.

Certain disadvantages were also observed, particularly the instability of IAT-YOLO in specific configurations, where training failed to converge or performance degraded in larger models. In addition, the added architectural complexity increased computational requirements and training time, which may pose challenges for deployment in resource-constrained

Table 10
Insulator detection results based on structure variations.

Class	Model	Precision	Recall	mAP		Training time (h)	Inference time (ms)
				[0.5]	[0.5:0.95]		
0	STD-YOLO	0.99893	**1.00000**	**0.99500**	0.98609	2.23	2.24
	IAT-YOLO	0.99889	**1.00000**	**0.99500**	0.98470	5.47	3.88
	SAE-YOLO	**0.99875**	**1.00000**	**0.99500**	**0.98647**	**2.07**	**2.09**
	STN-YOLO	0.99803	**1.00000**	**0.99500**	0.98508	8.14	3.48
1	STD-YOLO	0.60486	0.58742	0.58696	0.37528	**2.77**	**2.15**
	IAT-YOLO	0.62484	0.48420	0.53535	0.28498	8.49	3.35
	SAE-YOLO	**0.77638**	0.56527	**0.67778**	**0.47392**	**2.77**	2.23
	STN-YOLO	0.72782	**0.62791**	0.66754	0.45715	5.87	3.01
2	STD-YOLO	**0.76974**	0.48571	0.60225	0.36928	2.71	2.36
	IAT-YOLO	0.66121	0.48571	0.51963	0.30492	6.92	3.47
	SAE-YOLO	0.69839	**0.62857**	**0.65391**	**0.37419**	**2.65**	**2.32**
	STN-YOLO	0.72077	0.51429	0.56733	0.32393	5.88	3.35
3	STD-YOLO	**0.96224**	0.77222	0.88359	0.61897	**2.76**	**2.08**
	IAT-YOLO	0.87665	0.78967	0.84414	0.55547	6.32	3.44
	SAE-YOLO	0.89232	**0.83703**	**0.92448**	**0.65031**	2.77	2.28
	STN-YOLO	0.89814	0.80159	0.87194	0.61036	5.30	3.54
4	STD-YOLO	0.90939	0.84524	0.89809	0.52740	2.76	2.34
	SAE-YOLO	**0.91099**	0.86194	0.90158	**0.52971**	2.77	2.26
	IAT-YOLO	0.90359	**0.90526**	**0.92640**	0.51618	4.78	3.83
	STN-YOLO	0.86852	0.80000	0.88359	0.49783	5.05	3.17
5	STD-YOLO	**0.99334**	0.99386	0.99493	0.96471	**2.40**	**2.36**
	IAT-YOLO	0.98642	**1.00000**	0.99488	0.96031	4.93	3.30
	SAE-YOLO	0.99134	**1.00000**	0.99476	**0.96752**	2.61	2.38
	STN-YOLO	0.98683	**1.00000**	**0.99494**	0.95632	5.04	3.49
6	STD-YOLO	0.83888	0.69130	0.77861	0.35242	**2.53**	**2.14**
	IAT-YOLO	0.81181	**0.80870**	0.84085	0.38971	4.77	3.86
	SAE-YOLO	0.84118	0.74348	0.80754	0.37132	2.78	2.47
	STN-YOLO	**0.86617**	0.80000	**0.86658**	**0.39565**	5.05	3.78

Best results by class in bold.

or real-time environments. While the proposed method advances detection accuracy and robustness, future work should focus on mitigating instability and reducing computational overhead to maximize practical applicability.

While the experimental assumptions enabled controlled evaluation, they introduce some limitations. The dataset, although diverse, may not fully capture the complexity of environmental conditions (e.g., weather effects, occlusions, or hardware-induced noise) encountered during real inspections. The fixed image size and augmentation strategies may overlook other distortions, such as motion blur from UAV-based inspections. The reliance on high-end computational hardware raises questions about the feasibility of deployment on edge devices or real-time platforms. Addressing these assumptions in future work, through the use of broader datasets, resource-constrained evaluations, and operational metrics, would strengthen the generalizability and practical value of the research.

5. Conclusion

This paper presents a study of three architectural innovations for insulator fault detection in power grids: IAT-YOLO, SAE-YOLO blocks, and STN-YOLO. Extensive experiments on a real-world, multi-class dataset of insulator faults demonstrate that each architectural variation contributes unique strengths. SAE-YOLO modules enhance channel-wise feature learning, STN-YOLO improves geometric invariance and spatial alignment, and IAT-YOLO modules enable better contextual understanding via self-attention.

The structural variations may outperform baseline YOLO models, achieving up to 0.995 mAP on some classes and showing superior generalization across model scales. The results confirm the viability and effectiveness of combining attention and spatial transformation mechanisms in object detection for complex infrastructure inspection tasks.

Although the IAT-YOLO variant demonstrated promising results, the experiments revealed instability in certain configurations, particularly in the s, l, and x scales, where convergence issues and performance drops were observed. This instability may be related to sensitivity in gradient propagation within the self-attention mechanism when applied to smaller or over-parameterized models.

Future research should investigate systematic hyperparameter optimization (hypertuning), normalization strategies, and multi-head attention designs to improve stability across scales. Moreover, exploring lightweight variants of the attention mechanism, as well as training regularization and curriculum learning approaches, could mitigate non-convergence issues. By addressing these aspects, IAT-YOLO can be further strengthened, ensuring more robust generalization and consistent performance across different configurations.

Incorporating heterogeneity would strengthen model generalization and reduce the risk of performance drops when deployed in operational scenarios. The use of stronger validation strategies, including k-fold cross-validation and cross-dataset evaluation, is recommended to mitigate overfitting and provide more reliable evidence of robustness across different operating contexts. This is particularly relevant given that some variants, such as IAT-YOLO, exhibited instability at certain scales, which highlights the need for rigorous evaluation to ensure consistent performance. By addressing these aspects, future studies can bridge the gap between controlled experimental setups and real-world deployment, thereby increasing the trustworthiness and applicability of these models in safety-critical infrastructure monitoring.

CRediT authorship contribution statement

João Pedro Matos Carvalho: Writing – original draft, Software. **Stefano Frizzo Stefenon:** Writing – review & editing, Software, Methodology. **Valderi Reis Quietinho Leithardt:** Writing – review & editing. **Laio Oriel Seman:** Methodology. **Kin-Choong Yow:** Writing –

review & editing, Supervision. **Juan Francisco De Paz Santana:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Colaboración Consejería de Educación de la Junta de Castilla y León grupo de investigación ESAL-EXPERT SYSTEM AND APPLICATIONS LAB (ESALAB).

Data availability

Data will be made available upon request.

References

[1] Corso MP, Stefenon SF, Singh G, Matsuo MV, Perez FL, Leithardt VRQ. Evaluation of visible contamination on power grid insulators using convolutional neural networks. Electr Eng 2023;105:3881–94. https://doi.org/10.1007/s00202-023-01915-2
[2] Wang H, Yang Q, Zhang B, Gao D. Deep learning based insulator fault detection algorithm for power transmission lines. J Real Time Image Process 2024;21(4):115. https://doi.org/10.1007/s11554-024-01495-9z
[3] Cong Z, Liu Y, Yan Y, Wang K, Jiang X. Study on the mechanism and electrical characterization of the distribution porcelain insulator incipient fault in the non-effectively grounded system. IEEE Trans Power Delivery 2024;39(3):1840–51. https://doi.org/10.1109/TPWRD.2024.3380814
[4] Dai J, Gao Y, Cai C, Xiong W, Liu M. UAV-enabled inspection system with no-fly zones: drl-based joint mobile nest scheduling and UAV trajectory design. IEEE Access 2025;13:10844–56. https://doi.org/10.1109/ACCESS.2025.3529085
[5] Zachariades C, Peesapati V, Gardner R, Cwikowski O. Electric field and thermal analysis of 132 kv ceramic oil-filled cable sealing ends. IEEE Trans Power Delivery 2021;36(1):311–9. https://doi.org/10.1109/TPWRD.2020.2977728
[6] Chen X, Chen L, Chen L, Chen P, Sheng G, Yu X, Zou Y, et al. Modeling thermal infrared image degradation and real-world super-resolution under background thermal noise and streak interference. IEEE Trans Circuits Syst Video Technol 2024;34(7):6194–206. https://doi.org/10.1109/TCSVT.2023.3349182
[7] Zhang T, Tan J, Li S, Yang R. APE: anomaly-guided progressively balanced ensemble learning for insulator extraction from imbalanced LiDAR data. IEEE Trans Geosci Remote Sens 2025;63:1–19. https://doi.org/10.1109/TGRS.2025.3574343
[8] Haj YE, El-Hag AH, Ghunem RA. Application of deep-learning via transfer learning to evaluate silicone rubber material surface erosion. IEEE Trans Dielectr Electr Insul 2021;28(4):1465–7. https://doi.org/10.1109/TDEI.2021.009617
[9] Hu Z, Zhai B, Zhao Z, Zhai Y, Wang Q, Yang K. State-space-model-guided deep feature perception network for insulator defect detection in high-resolution aerial images. IEEE Trans Geosci Remote Sens 2025;63:1–14. https://doi.org/10.1109/TGRS.2025.3584663
[10] Singh L, Alam A, Kumar KV, Kumar D, Kumar P, Jaffery ZA. Design of thermal imaging-based health condition monitoring and early fault detection technique for porcelain insulators using machine learning. Environmental Technology & Innovation 2021;24:102000. https://doi.org/10.1016/j.eti.2021.102000
[11] Heiden PZ, Priefer J, Beverungen D. Predictive maintenance on the energy distribution grid–design and evaluation of a digital industrial platform in the context of a smart service system. IEEE Trans Eng Manag 2024;71:3641–55. https://doi.org/10.1109/TEM.2024.3352819
[12] Wang X, Xie X, Zhao S. Lstm-mm: efficient lstm-based mobility management for power inspection vehicles in smart grids. IEEE Access 2025;13:58992–9006. https://doi.org/10.1109/ACCESS.2025.3556712
[13] Stefenon SF, Oliveira JR, Coelho AS, Meyer LH. Diagnostic of insulators of conventional grid through LabVIEW analysis of FFT signal generated from ultrasound detector. IEEE Latin America Trans 2017;15(5):884–9. https://doi.org/10.1109/TLA.2017.7910202
[14] Zhu M, Zhang B, Zhou C, Zou H, Wang X. Target recognition of multi source machine vision pan tilt integrated inspection robot for power inspection. IEEE Access 2024;12:45693–708. https://doi.org/10.1109/ACCESS.2024.3378580
[15] Stefenon SF, Seman LO, Sopelsa Neto NF, Meyer LH, Mariani VC, Coelho LDS. Group method of data handling using Christiano-Fitzgerald random walk filter for insulator fault prediction. Sensors 2023;23(13):6118. https://doi.org/10.3390/s23136118
[16] Stefenon SF, Seman LO, Singh G, Yow K-C. Enhanced insulator fault detection using optimized ensemble of deep learning models based on weighted boxes fusion. Int J Electr Power Energy Syst 2025;168:110682. https://doi.org/10.1016/j.ijepes.2025.110682
[17] Fahim F, Hasan MS. Enhancing the reliability of power grids: a YOLO based approach for insulator defect detection. e-Prime Adv Electr Eng Electron Energy 2024;9:100663. https://doi.org/10.1016/j.prime.2024.100663
[18] Shukla PK, Deepa K. AI-based synthetic data generation techniques for improved fault classification in power systems. Ain Shams Eng J 2025;16(8):103485. https://doi.org/10.1016/j.asej.2025.103485
[19] Liu Y, Huang X, Liu D. Weather-domain transfer-based attention YOLO for multi-domain insulator defect detection and classification in UAV images. Entropy 2024;26(2):136. https://doi.org/10.3390/e26020136
[20] Stefenon SF, Cristoforetti M, Cimatti A. Automatic digitalization of railway interlocking systems engineering drawings based on hybrid machine learning methods. Expert Syst Appl 2025;281:127532. https://doi.org/10.1016/j.eswa.2025.127532
[21] Stefenon SF, Seman LO, Klaar ACR, Ovejero RG, Leithardt VRQ. Hypertuned-YOLO for interpretable distribution power grid fault location based on EigenCAM. Ain Shams Eng J 2024;15(6):102722. https://doi.org/10.1016/j.asej.2024.102722
[22] Sadykova D, Pernebayeva D, Bagheri M, James A. In-Yolo: real-time detection of outdoor high voltage insulators using UAV imaging. IEEE Trans Power Delivery 2020;35(3):1599–601. https://doi.org/10.1109/TPWRD.2019.2944741
[23] Josphineleela R, Kumar GVS, Ramesh T, Balamurugan KS. Optimized multiple object tracking with conformalized graph neural network and narwhal optimizer for embedded system IOT and mobile edge computing. Ain Shams Eng J 2025;16(10):103581. https://doi.org/10.1016/j.asej.2025.103581
[24] Corso MP, Stefenon SF, Couto VF, Cabral SHL, Nied A. Evaluation of methods for electric field calculation in transmission lines. IEEE Lat Am Trans 2018;16(12):2970–6. https://doi.org/10.1109/TLA.2018.8804264
[25] Zheng Z, Wang T, Bashir AK, Alazab M, Mumtaz S, Wang X. A decentralized mechanism based on differential privacy for privacy-preserving computation in smart grid. IEEE Trans Comput 2022;71(11):2915–26. https://doi.org/10.1109/TC.2021.3130402
[26] Zhou Z, Jia Z, Liao H, Lu W, Mumtaz S, Guizani M, Tariq M, et al. Secure and latency-aware digital twin assisted resource scheduling for 5G edge computing-empowered distribution grids. IEEE Trans Ind Informatics 2022;18(7):4933–43. https://doi.org/10.1109/TII.2021.3137349
[27] Stefenon SF, Seman LO, Pavan BA, Ovejero RG, Leithardt VRQ. Optimal design of electrical power distribution grid spacers using finite element method. IET Generation Transmission & Distribution 2022;16(9):1865–76. https://doi.org/10.1049/gtd2.12425
[28] Stefenon SF, Americo JP, Meyer LH, Grebogi RB, Nied A. Analysis of the electric field in porcelain pin-type insulators via finite elements software. IEEE Lat Am Trans 2018;16(10):2505–12. https://doi.org/10.1109/TLA.2018.8795129
[29] Glowacz A. Ventilation diagnosis of minigrinders using thermal images. Expert Syst Appl 2024;237:121435. https://doi.org/10.1016/j.eswa.2023.121435
[30] Glowacz A, Sulowicz M, Kozik J, Piech K, Glowacz W, Li Z, Brumercik F, Gutten M, Korenciak D, Kumar A, et al. Fault diagnosis of electrical faults of three-phase induction motors using acoustic analysis. Bull Pol Acad Sci Tech Sci 2024;72(1):e148440–e148440. https://doi.org/10.24425/bpasts.2024.148440
[31] Panigrahy S, Karmakar S. Real-time condition monitoring of transmission line insulators using the YOLO object detection model with a UAV. IEEE Trans Instrum Meas 2024;73:1–9. https://doi.org/10.1109/TIM.2024.3381693
[32] Li Z, Jiang C, Li Z. An insulator location and defect detection method based on improved yolov8. IEEE Access 2024;12:106781–92. https://doi.org/10.1109/ACCESS.2024.3436919
[33] Lu G, Li B, Chen Y, Qu S, Cheng T, Zhou J. Precision in aerial surveillance: integrating yolov8 with PConv and CoT for accurate insulator defect detection. IEEE Access 2025;13:49062–75. https://doi.org/10.1109/ACCESS.2025.3551289
[34] Zhang Q, Zhang J, Li Y, Zhu C, Wang G. ID-YOLO: a multimodule optimized algorithm for insulator defect detection in power transmission lines. IEEE Trans Instrum Meas 2025;74:1–11. https://doi.org/10.1109/TIM.2025.3527530
[35] He M, Qin L, Deng X, Liu K. MFI-YOLO: multi-fault insulator detection based on an improved yolov8. IEEE Trans Power Deliv 2024;39(1):168–79. https://doi.org/10.1109/TPWRD.2023.3328178
[36] Li Y, Zhu C, Zhang Q, Zhang J, Wang G. IF-YOLO: an efficient and accurate detection algorithm for insulator faults in transmission lines. IEEE Access 2024;12:167388–403. https://doi.org/10.1109/ACCESS.2024.3496514
[37] Liu B, Jiang W. Dfkd: dynamic focused knowledge distillation approach for insulator defect detection. IEEE Trans Instrum Meas 2024;73:1–16. https://doi.org/10.1109/TIM.2024.3485446
[38] Cheng Y, Liu D. Adin-detr: adapting detection transformer for end-to-end real-time power line insulator defect detection. IEEE Trans Instrum Meas 2024;73:1–11. https://doi.org/10.1109/TIM.2024.3420265
[39] Shi W, Lyu X, Han L. An object detection model for power lines with occlusions combining CNN and transformer. IEEE Trans Instrum Meas 2025;74:1–12. https://doi.org/10.1109/TIM.2025.3529073
[40] Hu M, Liu J, Liu J. DRR-YOLO: a study of small target multi-modal defect detection for multiple types of insulators based on large convolution kernel. IEEE Access 2025;13:26331–44. https://doi.org/10.1109/ACCESS.2025.3539831
[41] Zhang Q, Zhang J, Li Y, Zhu C, Wang G. Il-Yolo: an efficient detection algorithm for insulator defects in complex backgrounds of transmission lines. IEEE Access 2024;12:14532–46. https://doi.org/10.1109/ACCESS.2024.3358205
[42] Ding L, Rao ZQ, Ding B, Li SJ. Research on defect detection method of railway transmission line insulators based on gc-yolo. IEEE Access 2023;11:102635–42. https://doi.org/10.1109/ACCESS.2023.3316266
[43] Li Y, Feng D, Zhang Q, Li S. Hrd-yolox based insulator identification and defect detection method for transmission lines. IEEE Access 2024;12:22649–61. https://doi.org/10.1109/ACCESS.2024.3363430
[44] Shen P, Mei K, Cao H, Zhao Y, Zhang G. Lddfsf-yolo11: a lightweight insulator defect detection method focusing on small-sized features. IEEE Access 2025;13:90273–92. https://doi.org/10.1109/ACCESS.2025.3569970

[45] Qiu Z, Zhu X, Liao C, Shi D, Qu W. Detection of transmission line insulator defects based on an improved lightweight yolov4 model. Appl Sci 2022;12(3):1207. https://doi.org/10.3390/app12031207
[46] Sumagayan MU, Aleluya ER, Mangorsi RB, Alagon FJ, Aleluya ER, Mangorsi YAB, Premachandra HWH, Premachandra C, Kawanaka H, et al. Foreign object obstruction evaluation for distribution network inspection. IEEE Access 2024;12:172325–42. https://doi.org/10.1109/ACCESS.2024.3484156
[47] Li J, Zhou H, Lv G, Chen J. A2mada-YOLO: attention alignment multiscale adversarial domain adaptation YOLO for insulator defect detection in generalized foggy scenario. IEEE Trans Instrum Meas 2025;74:1–19. https://doi.org/10.1109/TIM.2025.3541814
[48] Li T, Zhu C, Li J, Cao H, Bai H. A real-time insulator condition detection model for UAV inspection based on fg-yolo. Meas Sci Technol 2025;36(5):056208. https://doi.org/10.1088/1361-6501/adcc4a
[49] Mahapatra U, Rahman MA, Islam MR, Hossain MA, Sheikh MRI, Hossain MJ. Adversarial training-based robust model for transmission line's insulator defect classification against cyber-attacks. Electr Power Syst Res 2025;245:111585. https://doi.org/10.1016/j.epsr.2025.111585
[50] Ren K, Yan T, Hu Z, Han H, Zhang Y. Image attention transformer network for indoor 3D object detection. Sci China Technol Sci 2024;67(7):2176–90. https://doi.org/10.1007/s11431-023-2552-x
[51] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on Computer vision and pattern recognition; 2018. p. 7132–41.
[52] Jin X, Xie Y, Wei X-S, Zhao B-R, Chen Z-M, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognition 2022;121:108159. https://doi.org/10.1016/j.patcog.2021.108159
[53] Wang L, Peng J, Sun W. Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification. Remote Sens 2019;11(7):884. https://doi.org/10.3390/rs11070884
[54] Yu H, Xu Z, Zheng K, Hong D, Yang H, Song M. Mstnet: a multilevel spectral–spatial transformer network for hyperspectral image classification. IEEE Trans Geosci Remote Sens 2022;60:1–13. https://doi.org/10.1109/TGRS.2022.3186400
[55] Zhong Z, Li Y, Ma L, Li J, Zheng W-S. Spectral–spatial transformer network for hyperspectral image classification: a factorized architecture search framework. IEEE Trans Geosci Remote Sens 2022;60:1–15. https://doi.org/10.1109/TGRS.2021.3115699
[56] Roboflow. Insulator faults detection computer vision dataset. 2023. https://universe.roboflow.com/project-vmgqx/insulator-faults-detection (visited on 21 Feb 2026).
[57] Stefenon SF, Singh G, Souza BJ, Freire RZ, Yow K-C. Optimized hybrid YOLOu-Quasi-ProtoPNet for insulators classification. IET Gener Transm Distrib 2023;17(15):3501–11. https://doi.org/10.1049/gtd2.12886

Author biography

João Pedro Matos Carvalho received the M.Sc. (Hons.) and Ph.D. degrees in electrical and computer engineering from FCT NOVA, Portugal, in 2017 and 2021, respectively. He is currently an Assistant Professor with the Department of Informatics, Faculty of Sciences, University of Lisbon. Since 2025, he has been an Integrated Member of LASIGE and a Collaborator of the Center of Technology and Systems (CTS), UNINOVA, and COPELABS. He won the highly competitive Scientific Employment Stimulus (CEEC institutional) FCT Grant in 2021. He has published more than 50 papers in international journals and international conferences in the fields of remote sensing, pattern recognition, machine learning, sensor networks, and signal processing, with significant recognition and impact in the research community.