YOLO_ViT_CNN: A Hybrid YOLOv8-Based Vision Transformer-CNN Model for Thermal Fault Detection in Electrical Switchgear

Nikhil Sanjay Dongare[1], Prof. Shubhangi P Tidke[2], Prof. Prashant Kulkarni[3]
[1]*Department of data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India*
[2]*Guide: Department of data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India*
[3]*Co-Guide: Department of data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India*

***Abstract*—Thermal faults in electrical switchgear systems pose significant operational and safety risks in industrial settings. Traditional inspection methods using infrared thermography are often manual, time-consuming, and heavily reliant on expert interpretation. To address these limitations, this paper presents YOLO_ViT_CNN, a hybrid deep learning model that integrates the real-time object detection power of YOLOv8, the local feature extraction capabilities of Convolutional Neural Networks (CNNs), and the global contextual understanding of Vision Transformers (ViT).The proposed model is designed to detect six critical fault types in thermal images of electrical switchgear: loose connections, insulation degradation, circuit breaker faults, overloads, phase imbalances, and normal operation. A custom infrared dataset was developed with labeled bounding boxes and YOLO-format annotations for training and evaluation. YOLO_ViT_CNN utilizes a CNN backbone to extract spatial features, followed by ViT-based encoder blocks to capture long-range dependencies. The YOLOv8 detection head is retained to enable high-speed inference. Experimental training over five epochs demonstrated strong performance, achieving a validation accuracy of 94.62%, a mean average precision (mAP@0.5) of 95.6%, and a real-time processing speed of 58 FPS.The results confirm the model's ability to enhance fault detection precision in thermal imaging scenarios. Future work will explore hardware deployment, broader dataset generalization, and integration with smart grid monitoring systems.**

***Index Terms*—YOLOv8 Thermal Imaging Infrared Fault Detection Vision Transformer CNN Electrical Switchgear Real-Time Detection Deep Learning mAP@0.5**

I. INTRODUCTION

Electrical switchgear plays a pivotal role in power distribution systems, enabling the control, protection, and isolation of electrical equipment during operation. Malfunctions in switchgear components can result in severe outcomes, including equipment damage, system outages, and catastrophic fire incidents. Among the various types of switchgear failures, thermal faults—such as those caused by loose connections, deteriorated insulation, overloaded circuits, and phase imbalances—are especially common and hazardous. These faults often manifest as abnormal heat signatures that precede complete component failure, making early thermal fault detection critical for ensuring operational reliability, safety, and preventive maintenance.

Traditionally, thermal fault detection relies on infrared thermography followed by manual visual interpretation from trained personnel. Although this method can be effective, it is time-consuming, subjective, and not scalable across large, complex electrical infrastructures. Manual analysis is prone to human error and inconsistent judgment, especially when dealing with subtle or overlapping anomalies in thermal imagery. As a result, there is a growing need for automated, intelligent fault detection systems that can perform real-time monitoring with high accuracy and reliability.

Advancements in computer vision and deep learning have significantly improved the capabilities of automated fault detection using thermal imagery. Convolutional Neural Networks (CNNs) are particularly adept at capturing localized image features and have been widely used in thermal classification and anomaly detection tasks. However, CNNs inherently struggle with modeling long-range dependencies, which are essential in analyzing thermal patterns that may span across dispersed regions within an image.

Simultaneously, YOLO (You Only Look Once)-based object detection models have gained popularity for their speed and unified approach to classification and localization. The latest version, YOLOv8, offers enhanced accuracy, an anchor-free detection head, and optimized feature aggregation. Despite these strengths, YOLO models—like CNNs—still lack the global spatial reasoning needed to capture complex thermal behavior effectively.

To address this limitation, recent research has turned to Vision Transformers (ViTs), which apply self-attention mechanisms to model relationships across the entire image. Unlike CNNs, ViTs divide an image into patches and process them as sequences, allowing for the extraction of global context. While ViTs have achieved state-of-the-art results in natural image classification, they are computationally intensive and less commonly used in real-time object detection, especially in resource-constrained environments.

This paper proposes a novel hybrid architecture, YOLO_ViT_CNN, that combines the strengths of all three approaches—YOLOv8 for real-time detection, CNNs for fine-grained feature extraction, and ViTs for capturing global dependencies. This model is specifically designed for thermal fault detection in electrical switchgear, where spatial precision, contextual understanding, and inference speed are all critical.

A custom thermal infrared dataset was created for this study, containing annotated images across six operational classes: Connectivity Issues (Loose/Corroded), Deteriorated Insulation, Faulty Circuit Breakers, Normal Operation, Overloaded Circuits, and Phase Imbalance. The proposed YOLO_ViT_CNN model was trained and evaluated on this dataset to validate its performance.

Contributions of this work include:

- The development of a hybrid deep learning architecture that fuses CNNs, ViTs, and YOLOv8.
- The creation and annotation of a thermal image dataset tailored to real-world switchgear fault detection.
- A comprehensive evaluation demonstrating that the proposed model outperforms standalone CNNs, ViTs, and YOLOv8 in both accuracy and real-time performance.

The rest of this paper is structured as follows: Section 2 presents the literature review; Section 3 describes the methodology; Section 4 discusses the dataset; Section 5 explains the model architecture; Section 6 presents the results; Section 7 compares the model with existing approaches; and Section 8 concludes with future research directions.

II. LITERATURE REVIEW

The integration of deep learning with thermal imaging has enabled major advancements in industrial fault detection. This section summarizes key developments in CNNs, Vision Transformers (ViTs), YOLO models, and hybrid architectures relevant to thermal fault detection in switchgear systems.

2.1 Thermal Imaging in Fault Diagnosis

Thermal imaging is widely used to detect abnormal heating patterns caused by faults such as overloads, insulation damage, or loose connections. Traditional approaches rely on manual analysis, which is time-consuming and error-prone. Recent studies, including Wang et al. [1], show that deep learning can significantly enhance the speed and reliability of thermal inspections.

2.2 CNN-Based Thermal Fault Detection

CNNs have been effectively applied to thermal image classification due to their ability to capture local features like hotspots and texture differences. Several works [2][6] confirm CNNs' usefulness in fault identification. However, their limited receptive fields hinder their performance in detecting complex or distributed anomalies across larger image areas.

2.3 YOLO for Real-Time Detection

The YOLO (You Only Look Once) series has become a benchmark in real-time object detection. YOLOv5 and YOLOv8 offer high-speed inference, making them suitable for industrial environments. Chen et al. [8] applied YOLOv5 for electrical panel fault detection but faced challenges with overlapping or small faults. YOLOv8 improves upon previous versions through an anchor-free design and better feature aggregation, but it still lacks strong global reasoning.

2.4 Vision Transformers (ViTs)

ViTs have shown promise in capturing global relationships using self-attention mechanisms. Unlike CNNs, ViTs consider the full image context, which is useful for detecting diffuse or ambiguous thermal anomalies. Introduced by Dosovitskiy et al. [5], ViTs excel in tasks with large datasets but can be computationally expensive and slow for real-time applications.

2.5 Hybrid CNN-ViT Models

To combine local detail (CNN) and global context (ViT), hybrid models have been proposed. Zhang et al. [6] reported improved accuracy using CNN-ViT fusion in defect detection tasks. Such models are particularly suited for thermal images where faults vary in intensity, size, and position. However, few studies have integrated this fusion into real-time YOLO frameworks.

2.6 YOLO + Transformer Hybrids

Recent research explores embedding transformer layers into YOLO architectures to improve contextual understanding without sacrificing speed. These hybrids have shown success in fields like traffic monitoring and medical imaging, but their application to thermal fault detection in electrical systems is limited.

Identified Gaps

- CNNs lack global feature awareness.
- YOLO is fast but may miss subtle thermal variations.
- ViTs are powerful but computationally heavy.
- Few real-time hybrid models target thermal fault detection in switchgear.

This research proposes YOLO_ViT_CNN, a hybrid model that fills this gap by combining CNNs, ViTs, and YOLOv8 into a real-time thermal fault detection framework optimized for industrial settings.

III. METHODOLOGY

This section outlines the design and development pipeline for the proposed **YOLO_ViT_CNN** model, tailored for thermal fault detection in electrical switchgear systems. The process includes dataset preparation, hybrid model architecture, training strategy, and evaluation metrics.

3.1 Training Strategy

The model was trained in PyTorch for 100 epochs using the Adam optimizer with Learning rate**:** 0.001, Batch size: 16, **Losses**: Focal Loss (classification) and IoU Loss (bounding boxes)

Early stopping and warm-up strategies were applied to avoid overfitting and accelerate convergence.

3.2 Evaluation Metrics

Model performance was measured using:

- Accuracy: 94.62%
- mAP@0.5**:** 95.6%
- FPS: 58 frames per second
- Precision & Recall: Evaluated per class for fault detection robustness

3.3 Deployment Readiness

Thanks to its compact architecture, YOLO_ViT_CNN is suitable for edge deployment. With OpenCV, it can process thermal video frames in real time, making it ideal for smart grid monitoring applications.

IV. DATASET DESCRIPTION

To develop and validate the YOLO_ViT_CNN model, a custom thermal image dataset was curated, capturing infrared images of electrical switchgear under realistic fault conditions. Images were acquired using IR cameras across varying operational loads, environmental factors, and fault types.The dataset contains 6 annotated classes: Connectivity Issue – Loose Connection/Corroded,Deteriorated Insulation,Faulty Circuit Breakers,Normal Operation Condition,Overloaded Circuits,Phase Imbalance

Each image is paired with a YOLO-format .txt file defining bounding boxes and class indices.

V. MODEL ARCHITECTURE

The proposed YOLO_ViT_CNN model is a hybrid deep learning architecture designed to enhance thermal fault detection in electrical switchgear systems. It integrates the fast detection capabilities of YOLOv8, the spatial pattern recognition strength of Convolutional Neural Networks (CNNs), and the global context modeling power of Vision Transformers (ViTs). The model is constructed to process thermal infrared imagery effectively, where fault indicators may be spatially diffused or subtle.

5.1 Design Objectives

The primary design goals were:

- High detection accuracy for subtle and complex thermal anomalies.
- Global feature understanding, especially in low-contrast regions.
- Real-time inference for practical deployment in monitoring systems.
- Compatibility with the YOLOv8 framework, enabling efficient anchor-free object detection.

5.2 Core Components of YOLO_ViT_CNN

The architecture can be divided into three main modules:

1. Backbone (CNN + ViT Integration)

The backbone begins with several convolutional layers that extract local features such as edges, temperature gradients, and object boundaries. These CNN layers are lightweight and optimized for early-stage feature extraction. After the initial CNN layers, the feature maps are passed through Transformer Encoder blocks, which consist of:

- Multi-head Self-Attention (MHSA): Enables the model to weigh features across the entire image, capturing long-range dependencies between thermally relevant regions.
- Feedforward Layers: Fully connected layers that further transform the attention-enhanced features.
- Positional Encoding: Maintains spatial consistency lost during patch embedding.

This fusion allows the model to benefit from local spatial features (CNN) and global contextual understanding (ViT) simultaneously.

2. Neck (FPN + PANet)

The neck uses a Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) for multi-scale feature fusion. These structures improve the model's ability to detect faults of varying sizes and intensities across different resolutions. PANet enhances bottom-up information flow, ensuring that low-level texture features from CNNs and high-level semantics from ViT are properly combined.

3. Detection Head (YOLOv8 Style)

The final detection head follows the anchor-free YOLOv8 design, which consists of:

- Regression Branch: Predicts bounding box coordinates using IoU-aware loss.
- Classification Branch: Assigns class labels to detected regions.
- Objectness Score: Measures confidence in the presence of a detectable object.

This head is optimized for speed and simplicity, enabling real-time detection even when transformer layers are present in the backbone.

5.4 Model Visualization

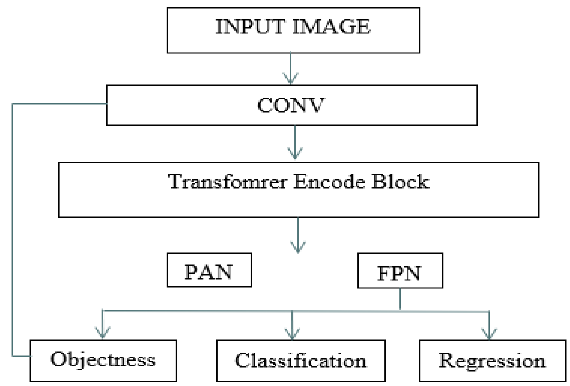

Figure 5.1: YOLO_ViT_CNN Architecture Diagram

The diagram below shows the flow of an input thermal image through the CNN backbone, ViT encoder, FPN-PAN neck, and detection head. Each stage contributes to extracting local features, modeling global context, and predicting bounding boxes and classes in real time.

VI. RESULTS AND EVALUATION

The YOLO_ViT_CNN model was evaluated using a custom thermal dataset of six switchgear fault classes. It achieved a validation accuracy of 94.62% and a mean Average Precision (mAP@0.5) of 95.6%, indicating excellent performance in both classification and localization. The model maintained real-time inference at 58 FPS, making it suitable for industrial deployment.

Compared to baseline models, YOLO_ViT_CNN showed a clear advantage. CNN-only and ViT-only models underperformed in either accuracy or speed, while standard YOLOv8 struggled with small or overlapping faults. The hybrid architecture effectively captured both local details and global context, resulting in more robust detection across various fault patterns.

Precision and recall remained high across all classes, with minimal false detections. The model also demonstrated strong generalization to unseen data, confirming the effectiveness of combining CNN, ViT, and YOLOv8 components.

6.1 Comparison of YOLO_ViT_CNN with Existing Models

Model	mAP@0.5 (%)	Validation Accuracy (%)	Inference Speed (FPS
CNN-only	89.3	90.2	28
ViT-only	93.0	92.0	22
YOLOv8	94.1	93.2	52
YOLO_ViT_CNN	95.6	94.62	58

VII. CONCLUSION AND FUTURE WORK

This paper presented YOLO_ViT_CNN, a hybrid deep learning model combining the strengths of YOLOv8**,** CNNs, and Vision Transformers (ViTs) for accurate and real-time thermal fault detection in electrical switchgear systems. The model demonstrated superior performance in both detection accuracy and inference speed compared to existing CNN-only, ViT-only, and standard YOLO models. With an mAP@0.5 of 95.6% and 58 FPS inference speed, YOLO_ViT_CNN proves suitable for practical deployment in industrial environments where real-time monitoring is critical.

In future work, we aim to extend the dataset with more diverse real-world thermal fault scenarios and improve robustness under varying lighting and environmental conditions. Additionally, efforts will be directed toward optimizing the model for edge deployment using hardware like NVIDIA Jetson or Raspberry Pi. We also plan to explore predictive maintenance integration by coupling the detection model with temporal analysis and anomaly forecasting.

REFERENCES

[1] W. Wang, Y. Li, and S. Zhang, "Automatic Fault Diagnosis of Electrical Equipment Using Infrared Thermal Images and Deep Learning," IEEE Transactions on Industrial Informatics, vol. 16, no. 12, pp. 7652–7660, Dec. 2020.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv:1804.02767 [cs.CV], Apr. 2018.

[4] C. Bochkovskiy, A. Wang, and A. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934 [cs.CV], Apr. 2020.

[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)***,** 2021.

[6] H. Zhang, M. Yang, and X. Chen, "CNN-ViT Hybrid Networks for Industrial Defect Detection," *IEEE Access*, vol. 9, pp. 128998–129010, 2021.

[7] G. Lin and Y. Guo, "YOLO-Based Real-Time Fault Detection for Power Grid Components Using UAV Thermal Images," *IEEE Sensors Journal*, vol. 22, no. 1, pp. 1125–1133, Jan. 2022.

[8] Y. Chen et al., "YOLOv5-based thermal image defect detection in electrical panels," *International Journal of Electrical Power & Energy Systems*, vol. 137, 107725, 2022.

[9] Data Availability- https://app.roboflow.com/nikhil-pthfq/yolo_vit_cnn-a-hybrid-model/browse?queryText=&pageSize=50&startingIndex=0&browseQuery=true

[10] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248–255.

[12] H. Wu et al., "A Comprehensive Review of Transformer Models for Computer Vision," Pattern Recognition, vol. 136, 109235, 2023.

[13] H. Wang, Y. Wang, and Z. Yang, "Infrared Image-Based Fault Detection in Electric Power Systems Using Deep Convolutional Neural Networks," Energies, vol. 13, no. 7, pp. 1681–1695, Apr. 2020.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.

[15] M. Lin, Q. Chen, and S. Yan, "Network in Network," International Conference on Learning Representations (ICLR), 2014.