

Advancing Thermal Image Processing: Modified YOLOv8 Algorithm for Surveillance System

MANDAR KHATAVKAR^{a,*}, BHAGAVAT JADHAV
JSPM's Rajarshi Shahu College of Engineering Tathwade,
Savitribai Phule Pune University, Pune,
INDIA

^aORCID: <https://orcid.org/0000-0002-9243-8469>

**Corresponding Author*

Abstract: - Thermal imaging plays a vital role in surveillance systems, specifically low-light or nighttime conditions. However, traditional methods have challenges in accomplishing accurate object recognition as well as skeleton generation in thermal images due to varying environmental conditions. To address these challenges and improve object recognition and skeleton generation accuracy in thermal images, the proposed You Only Look Once version 8 using Efficient net as backbone for Thermal Image (YOLO8-EFTI) algorithm was evaluated and compared with the traditional YOLO v8 to assess its effectiveness in recognizing objects and key points in thermal images. This study introduces a YOLO8-EFTI applied to thermal images, aimed at enhancing accuracy through performance parameters. It was hypothesized that the YOLO8-EFTI algorithm would show greater accuracy compared to state-of-the-art methods. The results of the experiments indicate that the proposed YOLO8-EFTI method enhances performance, achieving a 9.5% higher precision, 37.84% higher recall, 8.53% higher mAP@50, and a 7.95% higher mAP@50-95 than the traditional YOLOv8 architecture in object detection.

Key-Words: - Deep learning, Object detection, Surveillance system, Thermal imaging, YOLO algorithm.

Received: March 18, 2026. Revised: May 13, 2026. Accepted: June 12, 2026. Published: October 6, 2026.

1 Introduction

The surveillance system has become important part of today's era in the concern with safety, crime prevention and threats detection. One of the advancement in this field is use of Thermal Imaging in which object can be recognized based on heat rather than visible light. This technology is very useful in night as well as smoky condition where visibility of human eye is lower. Surveillance in such condition it is essential to use object detection and skeleton generation using thermal image.

There is advancement in traditional machine learning and deep learning methods for object detection and skeleton generation in thermal imaging, the conventional methods like Convolution Neural Network and Deep Learning models are significantly used in situation where data or computational resources are limited. The Models are able to learn from complex and multilayer features which effectively interpret and analyze the scenario.

In recent year, various models like You Only Look Once (YOLO), Faster R-CNN, and Single Shot MultiBox Detector (SSD), have been

developed for thermal object detection. There are development in architectural design, working mechanism in special layers which causes effectively detect small and low resolution objects.

Key strategies in deep learning-based thermal object detection include the following:

Model Architecture Optimization: Popular detection models (e.g., YOLOv8) are modified with fewer parameters and attention enhancements to suit the thermal imaging requirements.

Attention Modules: Channel and spatial attention mechanisms help models focus on the most informative regions within thermal images, [1].

Transfer learning: the models which designed for visible light images were converted for thermal imaging by fine tuning model parameters.

Small Object Detection: The small objects that are blurred in thermal images are detected by adding a specific layer or modifying respective field, [2].

Data Augmentation: The model generalization are improved by transformations such as rotation, flipping, and noise addition in order to differentiate

the training data.

Multi-Modal Fusion: The overall robustness and accuracy of the system are improved by integration of thermal images with visible spectrum data or inputs from other sensors, [3].

Preprocessing Techniques: To make object boundaries more distinguishable, methods like histogram equalization are commonly used in thermal images.

Efficient Inference: The lightweight architectures and model pruning techniques are widely used to achieve real-time performance on embedded systems, [4].

There are several factors such as environmental variability, domain differences, and the lack of large annotated thermal datasets makes several challenges to limit large-scale deployment of these systems.

To overcome these challenges, this work proposes the YOLO8-EFTI algorithm, for object detection and skeleton generation in thermal imagery. It is expected that this approach will achieve better results compared to existing models in terms of both speed and accuracy.

The rest of the paper is structured as follows. The Related Work section reviews existing research and recent advancements in this area. The Methodology section explains the proposed methodology, with model design and its key components. The detailed analysis of experimental findings is presented in Results and Discussion section, and finally, the contributions and outlines for future research is summarized in Conclusion.

2 Related Work

In surveillance application Thermal imaging has increased significant attention, particularly for detecting objects in low-light or nighttime environments. Researchers have proposed a lightweight version of YOLOv8 that incorporates a Global Attention Mechanism (GAM) and depth-wise convolutions, resulting in a 2% increase in precision and a 12.8% improvement in processing speed, while reducing the parameter count by approximately 75%.

Another study introduced a multi-scale, multi-target thermal infrared traffic detection method based on an improved YOLOv8. This approach integrates a novel Parm_C2f module and a DNC_Detect Head that utilizes deformable convolution, thereby enhancing model performance in complex traffic environments, [4].

For low-resolution infrared imagery, an enhanced YOLOv8 model incorporating a Spatial Pyramid Downsampling (SPD) module was

developed. This modification notably improves the small-object detection accuracy and increases the inference speed, demonstrating superior performance on datasets such as LLVIP and VEDAI, [5].

In the context of nighttime wildlife monitoring, the YOLOv8-night model introduces a channel attention mechanism, achieving higher mean Average Precision (mAP) scores than the standard YOLOv8. This makes it a reliable tool for ecological studies and wildlife conservation, [6].

The vehicle detection and classification system elaborates and evaluates an ensemble-based deep learning architecture, a collaborative framework merging “Efficient-Det and YOLOv8 for enhanced vehicle detection. The ensemble framework was significantly enhanced. Overall predictive performance, contributing to more robust and reliable results”, [7].

Modern surveillance systems operate across multiple environments, including ground-based, aerial, surface water, and underwater domains, [8]. While advancements in technology have significantly improved surveillance capabilities, they do not fully replace the role of human operators. In practice, CCTV operators still face cognitive challenges such as visual overload and lapses in attention, which must be considered when designing effective surveillance support systems, [9]. Furthermore, in the recent Internet of Things (IoT) technology large-scale data management, potential security risks included within video surveillance networks which introduce additional complexity, [10].

Accuracy and efficiency are comparatively improved in Deep learning approach for object detection and skeleton generation in thermal imagery, than traditional techniques. In particular, deep learning-based methods have advanced thermal image reconstruction and detection by addressing inherent challenges such as the lack of distinct visual features and the presence of thermal reflections.

For instance, deep learning approaches have demonstrated superior performance in reconstructing thermal images. A convolutional autoencoder-based image patch matching technique was developed for infrared imagery, which outperformed conventional matching methods and enhanced the accuracy of 3D object reconstruction, [11]. Similarly, a deep learning-based method for compressed sensing in MR image reconstruction was proposed, which operates without the need for extensive pre-training or large datasets, making it suitable for practical clinical applications, [12].

In the context of object detection, deep learning methods have also shown strong potential in handling challenges unique to thermal imagery, such as differentiating between actual objects and heat-induced reflections. A lightweight Mask R-CNN model has been introduced for this purpose, incorporating a three-channel input conversion strategy to improve detection performance, [13]. This method effectively reduces errors caused by reflections on surrounding surfaces due to thermal radiation.

In the summary, both the thermal image reconstruction and object detection achieved by the deep learning approaches with better performance. To make highly effective for analyzing thermal imagery these methods provides better image clarity, higher reconstruction accuracy, and reliable detection performance. The progress in development of deep learning algorithms and their applications in thermal imaging are expected to address the remaining challenges and expand the use of thermal cameras in various fields. The YOLO algorithm is used for object detection, [14]. These techniques can be used for surveillance and abnormality detection [15], [16], [17], [18], [19], [20], [21], [22].

The modern object detection and pose estimation frameworks, including RCNN-based methods, Detectron architectures, and YOLO-based models are used for object detection as well as skeleton generation. The use of these framework are depending on the application requirements. RCNN-based Approaches:

Region-based Convolutional Neural Networks (RCNN) and their developed versions like Fast RCNN and Faster RCNN, gives significant advances in object detection. These models are based on a two-stage detection pipeline, I first stage where candidate regions are generated and object classification and bounding box refinement process in nest stage. [23] the high localization accuracy and reliable detection performance are achieved due to this hierarchical approach; but, computational complexity are increased due to the multi-stage processing and reduced inference speed, particularly in real-time scenarios, [24].

Detectron Frameworks:

The human pose estimation tasks and applications such as robotics, motion analysis, and clinical rehabilitation the Detectron based frameworks are used. The detailed skeletal representations by accurately estimating joint positions provided by these frameworks, [25]. Detectron performs especially well in environments involving complex human-object interactions, where contextual and

spatial dependencies between detected regions can enhance pose accuracy. Nevertheless, in certain specialized rehabilitation tasks, alternative pose estimation models may achieve superior performance for specific joint movements, [26].

YOLO-based Methods:

The YOLO (You Only Look Once) family of models which can be used in object localization and classification. This design of YOLO helps to significantly reduces computational overhead make suitable for real-time applications. Although YOLO has slightly lower precision than two-stage detectors, but it consistently delivers fast and reliable detection, making it effective in time critical systems, [27], [28].

In summary, each detection method presented above has it unique strength: with the help of detailed region analysis the RCNN-based models gives better accuracy, YOLO methods are well known for their speed and efficiency for real-time applications, and In complex pose estimation tasks the Detectron based methods are widely used. Consequently, the selection of an method depends on the specific demands of the application, including accuracy requirements, computational constraints, and the complexity.

3 Proposed Methodology

In the YOLOv8 architecture, the backbone+ neck and head structure, the backbone+ neck is used for feature extraction. In the state-of-art method, CSPDarknet53 along with a Spatial Pyramid Pooling Fast (SPPF) network is used. The CSPDarknet53 network comprises a C2F network, which is an advanced version of the C2 network.

The splitting and merging is a principle operation of C2F network. The SPPF block consist of convolution block and a sequence of maxpool layers, and every resulting feature map is concatenated and ends with a convolution block. The SPPF block enables a neural network to work with images of different resolutions through pooling operations.

As shown in Figure 1, the head consists of a detection block that is responsible for detecting the object. The detection block contains two tracks: the first is for bounding box prediction, and the second is for class prediction. Both tracks have two convolution layers, followed by Conv2D layers, resulting in the bbox and class losses.

Thermal cameras were captured the thermal reflections of the objects and cerates Thermal Images. Figure 2 shows a block diagram of the

proposed method.

During the preprocessing stage, the blurring effects were mitigated, and the thermal images were enhanced to a higher resolution. In the subsequent stage, to enhance the accuracy of the object detection technique, modifications were required in the existing YOLOv8 architecture. YOLOv8 operates in two components: the backbone, neck, and head. In the proposed methodology YOLO8-EFTI, the backbone and neck structures are replaced with an efficient net network to improve the performance parameters.

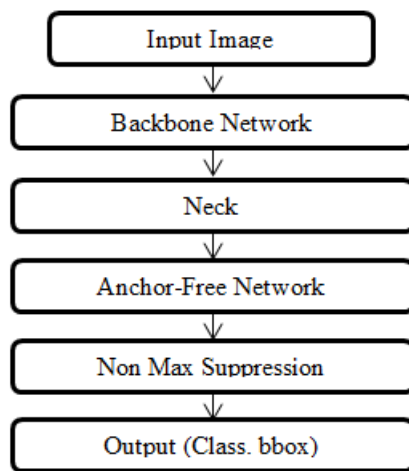


Fig. 1: YOLO v8 architecture
Source: created by the authors

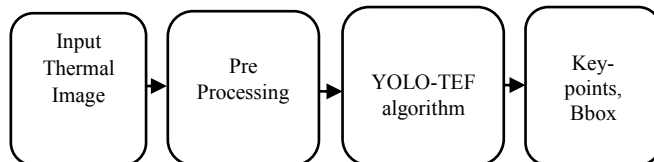


Fig. 2: Proposed YOLO-EFTI methodology
Source: created by the authors.

EfficientNet is highly optimized for extracting features from an image, and it offers a family of models (B0–B7), which provides a choice for selecting the backbone depending on the resources and accuracy. EfficientNet operates in three stages: convolution, feature extraction, and multiscaling. The initial convolution is performed using the Conv2D function with kernel, stride, and padding, the output of which is given to the mobile inverted residual Bottleneck Convolution (MB Conv.) block.

Initial Convolution

$$F_1 = \text{Conv2D}(I, K_1, S_1, P_1) \quad (1)$$

Where: K Kernel, S Stride, P Padding

MB Conv. This was used to improve feature quality and computational efficiency. The components used in the MB Conv. Block are:

Point-wise convolution (1x1) which was used to combine the information across channels, that is, expansion.

Point-wise convolution (expansion)

$$F_{\text{exp}} = \text{conv2D}(f_{\text{in}}, K_1, S_1, P_1)W_e + b_e \quad (2)$$

Where W_e, b_e are learnable weights

Depth-wise convolution is a specialized type of convolution that operates separately on each channel.

Depth-wise convolution

$$F_{\text{dw}} = \text{DepthwiseConv}(f_{\text{exp}}, K_d, S_d, P_d) \quad (3)$$

Point-wise convolution is used to reduce the number of channels back to their original size that is, the projection stage.

Point-wise convolution (expansion)

$$F_{\text{proj}} = \text{conv2D}(f_{\text{dw}}, K_1, S_1, P_1)W_p + b_p \quad (4)$$

The squeeze and excitation (SE) module is the main component of the MB Conv. This module recalibrates the channel-wise features of the input data. It selectively enhances and suppresses useful and less useful features, respectively.

Squeeze and excitation block

$$F_{\text{se}} = F_{\text{proj}} \cdot \sigma \left(F_c \left(\text{GlobalAvgPool}(F_{\text{proj}}) \right) \right) \quad (5)$$

Where σ is Activation Function

Stage-wise scaling is carried out, and then global pooling and features are extracted.

$$F_{\text{global}} = \text{GlobalAvgPool}(F_{\text{final}}) \quad (6)$$

The SE (Squeeze-and-Excitation) module operates in two main stages: squeeze and excitation. In the squeeze stage, average pixel value is calculated by global average pooling layer which is applied to each channel. This process effectively summarizes the entire feature map by capturing the most important information from each channel and compressing it into a single representative value.

In the excitation stage, these compressed values are passed through a small neural network. The appropriate weights are signed by this network by learning the relative importance of each channel, which improves both the efficiency and accuracy of the model.

The YOLOv8, directly predicts the center of an object which simplifies object localization, rather than calculating offsets from predefined anchor boxes. This removes the dependency on anchor boxes improves the detection process. Additionally,

the use of non-maximum suppression (NMS) in the post-processing eliminates the redundant and overlapping bounding boxes, which retained relevant bounding boxes.

In skeleton generation, the YOLO EFTI pose architecture improves feature representation, particularly for thermal imagery. These Feature maps are extracted from multiple stages of the EfficientNet backbone, which enables the model to capture both fine details and high-level semantic information.

Since these extracted features differ in channel dimensions from those in the original YOLOv8 design, depthwise-separable convolutional adaptation layers are introduced to align the feature dimensions while maintaining important spatial information needed for accurate keypoint detection. To further improve performance in low-contrast thermal environments, channel-wise attention modules are applied after feature adaptation. These modules help the network focus on human-related thermal patterns while reducing background interference.

A lightweight Path Aggregation Feature Pyramid Network (PAN-FPN) is then used to combine multi-scale features in a structured manner. This fusion enhances contextual understanding across different resolutions and improves the detection of small or partially occluded human figures. The refined feature maps are then fed into the YOLOv8-Pose head, which performs joint prediction of bounding boxes, confidence scores, and human skeletal keypoints in an end-to-end fashion.

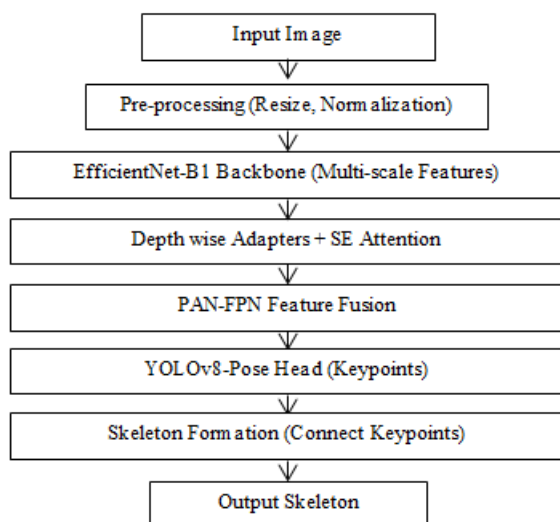


Fig. 3: Flowchart of YOLO-EFTI
Source: created by the authors.

Overall, the proposed architecture achieves a

balance between computational efficiency and improved detection performance, making it suitable for real-time human pose estimation in thermal imaging. The flowchart of the proposed skeleton generation method is illustrated in Figure 3.

Performance parameter

To assess the effectiveness of the YOLO8-EFTI algorithm for object detection in thermal images, several standard performance metrics were used:

Precision indicates the accuracy of positive predictions, Recall This reflects how well the model identifies all relevant objects in an image, The mAP@50 mean Average Precision at an IoU threshold of 0.5 evaluates detection accuracy and The mAP@50–95 provides a more comprehensive evaluation by averaging precision over multiple IoU thresholds ranging from 0.5 to 0.95, The overall performance measurement of the YOLO8-EFTI model is calculated by computed above metrics for each class and then averaged to obtained. To determine whether the proposed approach achieves improved accuracy in detecting objects within thermal images compare these values with those of existing methods.

4 Result and Discussion

The YOLO8-EFTI algorithm proposed in this work is designed to enhance object detection in thermal images, particularly for surveillance applications.

This study specifically compares the Object detection and skeleton generation state-of-the-art methods with proposed method and Furthermore, evaluating the model's performance in real-time scenarios is essential to confirm its practical applicability in surveillance systems.

To validate its effectiveness, the proposed method is compared with existing object detection models, including Faster R-CNN (FRCNN), EfficientDet, SSD, and YOLOv8. The comparison is carried out using standard performance metrics such as Precision, Recall, and F1-score. The same thermal image database included four classes (Bus, Car, Cycle, and Person) for training purposes. The Google Colab platform was used, and a T4 GPU was configured for training.

The Table 1 shows comparison between existing deep learning architecture and proposed YOLO-EFTI method, which indicated 7.04% improvement in Precision value, 3.04% improvement in Recall and 2.40% improvement in F1 Score value, A summary of the scores mentioned in Table 1 is graphically shown in Figure 4, [29].

Table 1. Comparison between Deep Learning Architecture and Proposed method

Methods	Precision	Recall	F1 Score
FRCNN	0.55	0.46	0.50
Efficient-Det	0.67	0.41	0.51
SSD	0.42	0.43	0.42
YOLO-8	0.71	0.44	0.54
YOLO-EFTI	0.76	0.60	0.67

Source: created by the authors.

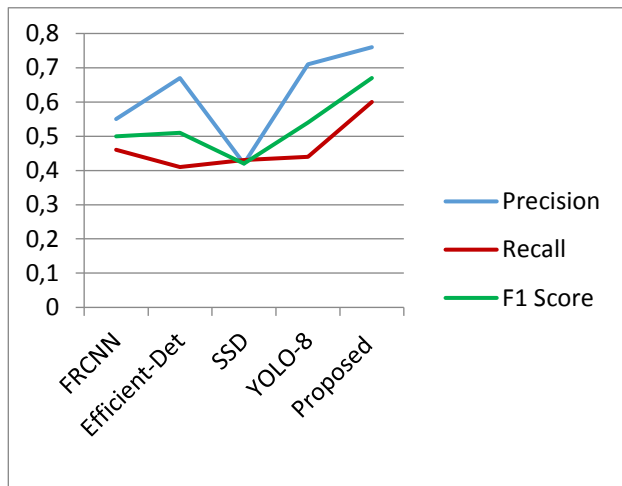


Fig. 4: Comparison of YOLO8-EFTI and CNN architecture

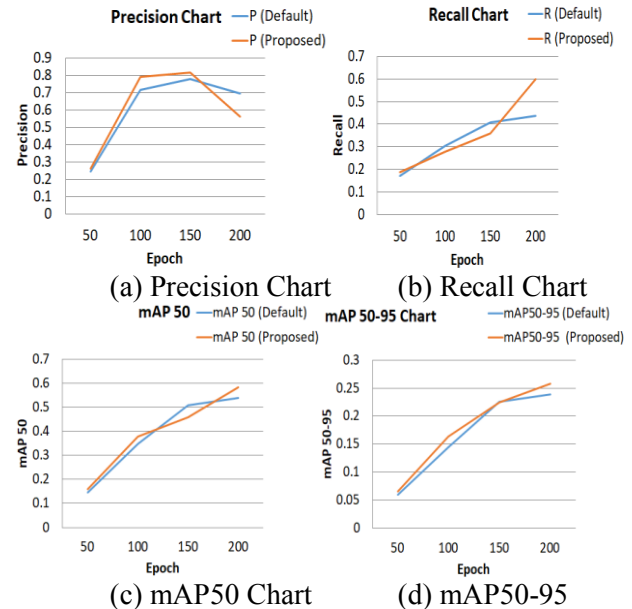
Source: created by the authors.

Performance was evaluated on the FLIR Thermal Dataset using epochs from 0–200 with a learning rate of 0.001. At the final epoch, YOLOv8-EFTI achieved higher performance with 76.1% precision and 60.1% recall compared to 69.5% and 43.6% for the default YOLOv8 as represented in Table 2 (Appendix). The proposed model also improved detection accuracy, achieving 58.5% mAP@50 and 25.8% mAP@50–95, outperforming the baseline YOLOv8 (53.9% and 23.9%).

Figure 5 illustrates performance variation across training epochs (1–200) for the default YOLOv8 and the proposed YOLO8-EFTI model, showing improvements in performance parameters during training. Table 3 (Appendix) compares keypoint detection results of the default YOLOv8 pose model and YOLO8-EFTI at different epochs (50–200) using OKS-based metrics. The proposed model achieves the highest precision and improves Pose mAP@50–95 while maintaining similar Pose mAP@50 compared to the baseline.

Although a slight decrease in recall is observed, the results show that the proposed YOLO8-EFTI model delivers more precise and reliable keypoint predictions, particularly at higher IoU thresholds. Overall, the findings indicate that the modified YOLO8-EFTI pose architecture consistently

enhances pose estimation accuracy and achieves comparable or better mAP performance across training epochs. This highlights its effectiveness in accurately localizing keypoints when compared to the standard YOLO pose model.



Chart

Fig. 5: Comparison of proposed YOLO8-EFTI method and YOLOv8 Default architecture

Source: created by the authors.

Table 4. Testo 686 Thermal Camera specification

Sr. No.	Parameter	Specification
1	Infrared resolution	160 X 120 pixel
2	Super Resolution (Pixel)	320 X 240 pixel
3	Thermal sensitivity	<0.08° C
4	Spectral range	7.5 to 14 μ m
5	Image refresh rate	9 Hz

Source: created by the authors

The proposed YOLO8-EFTI methodology was also applied to a real-time dataset captured on campus using a Testo686 thermal camera, the specifications of which are listed in Table 4. The some sample images are shown in Figure 7

The real-time image dataset consists of four classes: Bus, Car, Cycle, and Person. All images were pre-processed using the Testo software. These processed images were applied to the proposed YOLO8-EFTI method on the Google Colab platform using a T4 GPU. The results are shown in Figure 8.

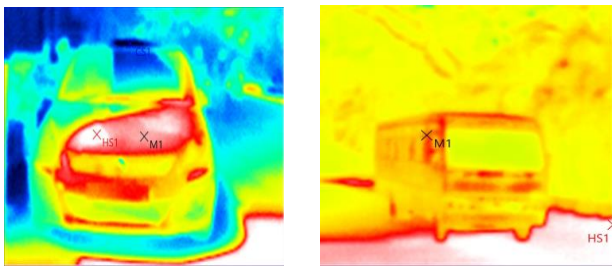


(a) Predicted Image I (b) Predicted Image II

Fig. 6: Predicted thermal images obtained using proposed YOLO8-EFTI method.

Source: created by the authors.

Figure 6 shows the predicted results of the thermal images from the test dataset of the FLIR thermal dataset.



(a) Car (b) Bus

Fig. 7: Real time thermal images captured using Testo 686

Source: created by the authors.

The thermal image dataset were pre-processed and applied to the proposed YOLO8-EFTI method on the Google Colab platform using a T4 GPU. The proposed YOLO8-EFTI method was applied to one class i.e. Person with 6 keypoints (head, body, left hand, right hand, left leg and right leg). The results are shown in Figure 9.

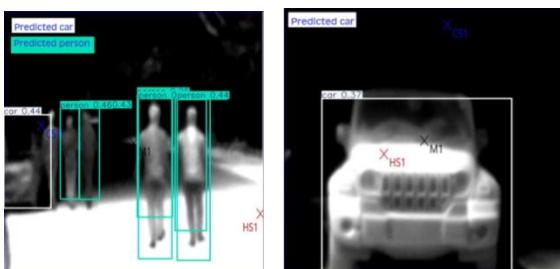


Fig. 8: Output of proposed YOLO8-EFTI method

Source: created by the authors.

Table 6 presents a comparative performance analysis of different EfficientNet architectures (B0–B7) when used as backbones in an object detection or image analysis model (likely YOLO or a similar detector). Table 5 (Appendix) evaluates the

performance using standard metrics: Precision, Recall, F1 Score, mAP@50, and mAP@50-95.

Stochastic Gradient Descent (SGD) is a widely used optimization method due to its simplicity and effectiveness in training large-scale models, [30]. The randomness can introduce gradient variance and slow convergence, [31]. To address this, improved variants such as mini-batch SGD and advanced sampling methods (e.g., typicality sampling) have been developed to reduce noise and enhance training efficiency, [32].



Fig. 9: Keypoint Prediction using proposed YOLO8-EFTI method

Source: created by the authors.

Table 6. Analysis of Efficient Net Backbone architecture

Architecture	P	R	F1 Score	mAP@ 50	mAP@ 50-95
B0	0.561	0.601	0.290	0.585	0.258
B1	0.785	0.614	0.345	0.639	0.295
B2	0.561	0.601	0.290	0.585	0.258
B3	0.703	0.584	0.319	0.628	0.306
B4-B7	0.561	0.601	0.290	0.585	0.258

Source: created by the authors.

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} J(\theta_t) \quad (7)$$

Where,

θ_t : model parameters at step t .

$\nabla_{\theta} J(\theta_t)$: Gradient of the loss function J with respect to parameters

η : Learning rate

The Adam optimizer is commonly used in deep learning due to its ability to adapt the learning rate for each parameter and handle sparse gradients efficiently. It combines the strengths of AdaGrad and RMSProp, resulting in faster and more stable convergence compared to standard Stochastic Gradient Descent, [33]. Adam works by maintaining moving averages of both the gradients and their squared values to compute adaptive learning rates.

However, in applications involving differential privacy, the introduction of noise can affect performance by introducing bias in the estimation of

the second moment. To address this issue, approaches such as DP-AdamBC have been proposed, which apply bias correction techniques to improve accuracy under noisy conditions, [34].

The first moment (mean of gradients) (8), second moment (uncentered variance) (9), and their corresponding bias-corrected estimates (10) are calculated as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \cdot g_t \quad (8)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \cdot g_t^2 \quad (9)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (10)$$

Where,

m_t : momentum, smooth the gradient

v_t : scales the learning rate adaptively

β_1 : Exponential decay rate for first moment

β_2 : Exponential decay rate for second moment

$g_t = \nabla_{\theta} J(\theta_t)$: Gradient of the loss function

An extension of the AdamW optimizer introduces weight prediction to further improve its performance. In this approach, future weights are estimated before each mini-batch update by utilizing the AdamW update rule. These predicted weights are then used during both the forward and backward passes, allowing the optimizer to achieve faster convergence and improve overall training efficiency in deep neural networks, [35].

AdamW itself is an improved variant of the Adam optimizer, where weight decay (11) is separated from the gradient update process. This decoupling helps in better regularization, leading to improved generalization and reduced overfitting. As a result, AdamW is particularly well-suited for training complex deep learning architectures.

$$g_t = \nabla_{\theta} J(\theta_t) + \lambda \cdot \theta_t \quad (11)$$

Where,

λ : weight decay coefficient

In general, for simple models Stochastic Gradient Descent (SGD) can be effective, Adam and AdamW are often preferred in modern deep learning applications due to their faster convergence and better performance.

Following Table 5 (Appendix) shows a comparative analysis using three optimizers: SGD, Adam, and AdamW to compare the YOLO default architecture and the proposed YOLO8-EFTI approach

Table 7. Analysis based on different Learning rate

Learning Rate	P	R	F1 Score	mAP@ 50	mAP@ 50-95
0.0001	0.629	0.216	0.322	0.243	0.0896
0.0005	0.821	0.516	0.634	0.563	0.247
0.001	0.792	0.519	0.627	0.582	0.28
0.005	0.813	0.651	0.723	0.691	0.34
0.01	0.771	0.524	0.624	0.649	0.283

Source: created by the authors.

In the further analysis hyper parameter learning rates are varied to compare proposed YOLO8-EFTI approach, with YOLOv8 architecture with the AdamW optimizer, as presented in Table 7. This tuning the learning rate highlights the optimal performance in thermal image detection tasks.

Similarly, Table 8 presents comparison between YOLOv8 and proposed YOLO8-EFTI approach which indicates 3.70% in Precision, 15.38% in Recall, 9.72% in F1 Score, 18.84% in mAP@50 and 8.10% in mAP@50-95 improvement.

Table 8. Analysis based on different Batch Size

Batch Size	P	R	F1 Score	mAP@ 50	mAP@ 50-95	Time (hr)
4	0.53	0.51	0.52	0.51	0.22	0.14
8	0.72	0.62	0.67	0.69	0.30	0.11
16	0.81	0.65	0.72	0.69	0.34	0.09
32	0.84	0.75	0.79	0.82	0.37	0.10

Source: created by the authors.

[36], served as the primary source of thermal images used in this research. The dataset contains grayscale thermal images captured under diverse environmental conditions, which is suitable for training and evaluating proposed YOLO8-EFTI model.

[37] was used for annotation, and preprocessing. The platform facilitated the manual labeling of class objects keypoints and bounding boxes, ensuring high-quality annotations compatible with the proposed YOLO8-EFTI model.

5 Conclusion

This study proposes the YOLO8-EFTI model, a modified version of YOLOv8 designed to improve object and pose detection in thermal surveillance environments. The backbone is enhanced using EfficientNet-B1, which achieved the best balance of detection performance with 78.5% precision, 68.9% F1-score, 63.9% mAP@50, and 29.5% mAP@50-95 on the FLIR Thermal Dataset. EfficientNet-B3 provided the highest mAP@50-95, indicating stronger fine-grained detection capability, while

other EfficientNet variants produced similar results under identical training conditions.

For keypoint detection, the YOLO8-EFTI model improved pose precision by 2.94% and increased pose mAP@50–95 by 3.10%, demonstrating better keypoint localization accuracy and reduced false positives compared with the default YOLOv8 pose model. Across all evaluation metrics—precision, recall, F1-score, mAP@50, and mAP@50–95—the proposed method consistently outperformed the baseline architecture.

Among optimizers, Adam optimizer achieved the highest precision and F1-score, while AdamW optimizer provided the best generalization with the highest mAP@50–95. Experiments with learning rates showed optimal performance at 0.005, while excessively high rates slightly reduced accuracy.

Batch size experiments shows improved performance, with the best results obtained at a batch size of 32, achieving the highest precision, recall, F1-score, and detection accuracy.

Overall, by experimentation it is proved that modified version of YOLOv8, YOLO8-EFTI framework significantly enhances detection accuracy, robustness, and training stability for thermal image surveillance applications.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process The authors wrote, reviewed and edited the content as needed and they have not utilised artificial intelligence (AI) tools. The authors take full responsibility for the content of the publication.

References:

- [1] Liu, Z.; El-Sousy, F.F.M.; Larik, N.A.; Quan, H.; Ji, T. Riemannian Geodesic Discriminant Analysis–Minimum Riemannian Mean Distance: A Robust and Effective Method Leveraging a Symmetric Positive Definite Manifold and Discriminant Algorithm for Image Set Classification. *Mathematics* 2024, 12, 2164. <https://doi.org/10.3390/math12142164>.
- [2] Kim, Y. S., Choi, H. J., & Park, M. (2025). A Lightweight Network Based on YOLOv8 for Improving Detection Performance and the Speed of Thermal Image Processing. *Electronics*, 14(3), 783. <https://doi.org/10.3390/electronics14030783>.
- [3] Liu, Y., Xue, W., Zhu, H., Wang, S., & Chen, L. (2024). Research on Infrared Dim Target Detection Based on Improved YOLOv8. *Remote Sensing*, 16(16), 2878. <https://doi.org/10.3390/rs16162878>.
- [4] Ren, Z., Zhang, T., & Li, Y. (2024). Adaptive Multi-Scale Fusion for Infrared and Visible Object Detection in YOLOv8. *Journal of Theory and Practice of Engineering Science*, 12(4), 224–233.
- [5] Chen, Y., Li, X., & Zhou, W. (2024). Multiscale and Multitarget Thermal Infrared Traffic Detection Based on Improved YOLOv8. In *Proceedings of the International Conference on Artificial Intelligence and Signal Image Processing (AASIP 2024)*.
- [6] Ren, Z. (2024). Enhanced YOLOv8 Infrared Image Object Detection Method with SPD Module. *Journal of Theory and Practice in Engineering and Technology*, 1(2), 1–7, [Online]. <https://woodyinternational.com/index.php/jtpe/article/view/42> (Accessed Date: January 15, 2026).
- [7] Zhang, Y., Li, X., & Wang, J. (2025). Nighttime Wildlife Object Detection Based on YOLOv8-Night. *Electronics Letters*, 61(3), 123–125. <https://doi.org/10.1049/el12.13305>.
- [8] Lv C, Mittal U, Madaan V, Agrawal P. Vehicle detection and classification using an ensemble of EfficientDet and YOLOv8. *PeerJ Comput Sci.* 2024 Aug 13;10:e2233. doi: 10.7717/peerj-cs.2233. PMID: 39314728; PMCID: PMC11419654.
- [9] Chun, W. H., & Papanikolopoulos, N. (2016). *Robot Surveillance and Security* (pp. 1605–1626). Springer. https://doi.org/10.1007/978-3-319-32552-1_61.
- [10] Hodgetts, H. M., Vachon, F., Chamberland, C., & Tremblay, S. (2017). See No Evil: Cognitive Challenges of Security Surveillance and Monitoring. *Journal of Applied Research in Memory and Cognition*, 6(3), 230–243. <https://doi.org/10.1016/j.jarmac.2017.05.001>.
- [11] Wang, R., Li, Q., Liu, C., Tsai, W.-T., He, J., & Deng, E. (2019). A Video Surveillance System Based on Permissioned Blockchains and Edge Computing. 1–6. <https://doi.org/10.1109/bigcomp.2019.8679354>.
- [12] Knyaz, V. A., Gorbatshevich, V., Luhmann, T., Conen, N., Kniaz, V. V., Vizilter, Y., & Vygolov, O. (2017). Deep Learning of Convolutional Auto-Encoder for Image Matching and 3D Object Reconstruction in the Infrared Range. 9907, 2155–2164. <https://doi.org/10.1109/icevw.2017.252>.
- [13] Zhao, D., Zhao, F., & Gan, Y. (2020). Reference-Driven Compressed Sensing MR Image Reconstruction Using Deep

- Convolutional Neural Networks without Pre-Training. *Sensors*, 20(1), 308. <https://doi.org/10.3390/s20010308>.
- [14] Batchuluun, G., Nguyen, D. T., Arsalan, M., Pham, T. D., Park, K. R., & Kang, J. K. (2021). Deep Learning-Based Thermal Image Reconstruction and Object Detection. *IEEE Access*, 9, 5951–5971. <https://doi.org/10.1109/access.2020.3048437>.
- [15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *IEEE*, 2016
- [16] D. Tan, K. Huang, S.Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, Aug. 2006, pp. 1000–1003.
- [17] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, Jun. 2014, pp. 201–208.
- [18] D. Gangodkar, P. Kumar, and A. Mittal, "Segmentation of moving objects in visible and thermal videos," in *Proc. Int. Conf. Comput. Commun. Informat.*, Coimbatore, India, Jan. 2012, pp. 1–5.
- [19] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Proc. Int. Conf. Comput. Vis., Graph. Image Process.*, Madurai, India, Dec. 2006, pp. 528–539.
- [20] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," in *Proc. 2nd Int. Conf. Comput. Res. Develop.*, Kuala Lumpur, Malaysia, May 2010, pp. 747–751.
- [21] J.W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understand.*, vol. 106, nos. 2–3, pp. 162–182, May 2007.
- [22] Heena Patel and Kishor P. Upla, "Night Vision Surveillance: Object Detection using Thermal and Visible Images", 2020 International Conference for Emerging Technology (INCET)Belgaum, India. Jun 5-7, 2020
- [23] Lokanath, M., Kumar, K. S., & Keerthi, E. S. (2017). Accurate object classification and detection by faster-RCNN. *IOP Conference Series: Materials Science and Engineering*, 263(5), 052028. <https://doi.org/10.1088/1757-899x/263/5/052028>.
- [24] Chandan, G., Jain, H., Jain, A., & Mohana, M. (2018, July 1). Real Time Object Detection and Tracking Using Deep Learning and OpenCV. <https://doi.org/10.1109/icirca.2018.8597266>.
- [25] Hernández, Ó. G., Ramon, J. L., Morell, V., & Jara, C. A. (2021). Human Pose Detection for Robotic-Assisted and Rehabilitation Environments. *Applied Sciences*, 11(9), 4183. <https://doi.org/10.3390/app11094183>.
- [26] Chao, Y.-W., Deng, J., Zeng, H., Liu, X., & Liu, Y. (2018, March 1). Learning to Detect Human-Object Interactions. <https://doi.org/10.1109/wacv.2018.00048>.
- [27] Krishna, N. M., Reddy, M. S. C., Sudham, G., Madhav, K. P., & Reddy, R. Y. (2021). Object Detection and Tracking Using Yolo. 1–7. <https://doi.org/10.1109/icirca51532.2021.9544598>.
- [28] Lavanya, G., & Pande, S. D. (2023). Enhancing Real-time Object Detection with YOLO Algorithm. *EAI Endorsed Transactions on Internet of Things*, 10. <https://doi.org/10.4108/eetiot.4541>.
- [29] M. D. Khatavkar, B. D. Jadhav (2025), *A Comparative Analysis Of CNN Models For Object Detection In Thermal Imaging*, *Int. J. Environ. Sci.*, pp. 2709–2718, Jul. 2025, doi: 10.64252/cz9zbp65.
- [30] Tian, Y., Zhang, H., & Zhang, Y. (2023). Recent Advances in Stochastic Gradient Descent in Deep Learning. *Mathematics*, 11(3), 682. <https://doi.org/10.3390/math11030682>.
- [31] De, S., & Goldstein, T. (2016). *Efficient Distributed SGD with Variance Reduction*. 111–120. <https://doi.org/10.1109/icdm.2016.0022>.
- [32] Peng, X., Wang, F.-Y., & Li, L. (2020). Accelerating Minibatch Stochastic Gradient Descent Using Typicality Sampling. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11), 4649–4659. <https://doi.org/10.1109/tnnls.2019.2957003>.
- [33] Mehmood, F., Whangbo, T. K., & Ahmad, S. (2023). An Efficient Optimization Technique for Training Deep Neural Networks. *Mathematics*, 11(6), 1360. <https://doi.org/10.3390/math11061360>.
- [34] Tang, Q., Shpilevskiy, F., & Lécuyer, M. (2024). DP-AdamBC: Your DP-Adam Is Actually DP-SGD (Unless You Apply Bias Correction). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14),

15276–15283.

<https://doi.org/10.1609/aaai.v38i14.29451>.

- [35] Guan, L. (2023). *Weight Prediction Boosts the Convergence of AdamW* (pp. 329–340). springer nature switzerland. https://doi.org/10.1007/978-3-031-33374-3_26.
- [36] FLIR Systems, “FLIR ADAS Thermal Dataset,” FLIR Systems, Inc., 2018, [Online]. <https://www.flir.in/oem/adas/adas-dataset-form> (Accessed Date: January 23, 2026).
- [37] Roboflow, “Roboflow: Computer Vision Dataset Management and Annotation Tool,” Roboflow Inc., [Online]. <https://roboflow.com> (Accessed Date: January 23, 2026).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflicts of Interest

The authors declare that they have no conflict of interest. The author is pursuing Ph. D. so there is no involvement in commercial and financial relationships.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

Table 2. Comparison between YOLOv8 Default Architecture and YOLO8-EFTI for object detection.

Epoch	YOLOv8 Default Architecture				Proposed Method			
	P	R	mAP@50	mAP@50-95	P	R	mAP@50	mAP@50-95
50	0.246	0.172	0.145	0.060	0.262	0.187	0.159	0.066
100	0.716	0.305	0.347	0.144	0.792	0.277	0.379	0.164
150	0.78	0.409	0.508	0.226	0.816	0.358	0.459	0.224
200	0.695	0.436	0.539	0.239	0.761	0.601	0.585	0.258

Source: created by the authors.

Table 3. Keypoint Comparison between Yolo Default Pose and YOLO8-EFTI pose detection

Epoch	Pose Precision (OKS)		Pose Recall (OKS)		F1 Score		Pose mAP@50 (OKS)		Pose mAP@50-95 (OKS)	
	Default	Proposed	Default	Proposed	Default	Proposed	Default	Proposed	Default	Proposed
50	0.909	0.921	0.858	0.829	0.883	0.873	0.884	0.877	0.454	0.438
100	0.951	0.977	0.936	0.957	0.943	0.967	0.975	0.983	0.637	0.628
150	0.917	0.949	0.978	0.936	0.947	0.942	0.983	0.970	0.690	0.651
200	0.951	0.979	0.978	0.893	0.964	0.934	0.982	0.982	0.645	0.665

Source: created by the authors.

Table 5. Analysis based on Optimizers

Optimiser	YOLO Default Architecture					Proposed YOLO8-EFTI Method				
	P	R	F1	mAP50	mAP50-95	P	R	F1	mAP50	mAP50-95
SGD	0.629	0.496	0.277	0.511	0.242	0.65	0.531	0.292	0.601	0.276
Adam	0.752	0.643	0.347	0.74	0.374	0.892	0.71	0.395	0.79	0.379
AdamW	0.77	0.753	0.381	0.778	0.42	0.828	0.669	0.370	0.775	0.422

Source: created by the authors.