ESKİŞEHİR TECHNICAL UNIVERSITY
FACULTY OF ENGINEERING

Deep Learning-Based Object Detection in High-Altitude Thermal Imagery from UAVs

Duygu HALİSYAMA
Furkan DÖNMEZ

A Bachelor of Science Project
Department of Computer Engineering

May 2025

Deep Learning-Based Object Detection in High-Altitude Thermal Imagery from UAVs

by

Duygu HALİSYAMA
36086147230

Furkan DÖNMEZ
48346699694

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Engineering is approved by the following scientific committee members.

Date of Approval **:** Month xx, 20xx

Member (Advisor) **:** Asst. Prof. Dr. Cahit Perkgöz

Member **:** Assoc. Prof. Dr. Mehmet Koç

Member **:** Assist. Prof. Dr. Mehmet Kılıçarslan

Member **:** Dr. Ahmet Aydın

ABSTRACT

Unmanned Aerial Vehicles (UAVs) equipped with thermal cameras enable night-time and low-visibility object detection over wide areas, but high-altitude thermal imagery suffers from low contrast and small target size. In this study, we propose a real-time detection pipeline based on the medium variant of YOLOv8 (YOLOv8m) enhanced by Contrast-Limited Adaptive Histogram Equalization (CLAHE). We compare three configurations: (1) YOLOv8m baseline with CLAHE preprocessing, (2) CLAHE + YOLOv8m with Convolutional Block Attention Module (CBAM), and (3) CLAHE + YOLOv8m with Coordinate Attention (CoordAtt). All models are trained and evaluated on the HIT-UAV dataset (2 898 images, 5 classes) using standard metrics (Precision, Recall, F1-score, mAP@0.5, mAP@0.5:0.95). Results show that CLAHE preprocessing alone boosts recall by 12.3 % and mAP@0.5 by 9.1 % over the baseline, achieving the highest overall detection performance. While attention modules improve precision (up to 88.4 % with CBAM), they incur notable recall drops and lower mAP gains compared to CLAHE only. Our findings indicate that simple contrast enhancement can outperform more complex attention mechanisms in low-detail thermal imagery, providing practical guidance for lightweight, real-time UAV-based object detectors in search-and-rescue, surveillance, and defense applications.

Keywords: UAV, thermal imagery, object detection, YOLOv8, CLAHE, attention mechanisms

ÖZET

Yüksek irtifadan elde edilmiş termal görüntüler, geniş alan taraması ve gece-görüş imkânı sağlarken düşük kontrast ve küçük hedef boyutu nedeniyle nesne tespitinde zorluklar yaratır. Bu çalışmada, YOLOv8 orta boyutlu modeli (YOLOv8m) temelli, gerçek zamanlı bir tespit hattı önerilmektedir. Model, kontrast arttırma için CLAHE (Contrast-Limited Adaptive Histogram Equalization) ön işleme ile üç farklı mimari olarak incelenmiştir: (1) yalnızca CLAHE + YOLOv8m, (2) CLAHE + YOLOv8m + CBAM (Convolutional Block Attention Module) ve (3) CLAHE + YOLOv8m + CoordAtt (Coordinate Attention). Tüm modeller, HIT-UAV veri seti (2 898 görüntü, 5 sınıf) üzerinde Precision, Recall, F1-skor, mAP@0.5 ve mAP@0.5:0.95 metrikleriyle değerlendirildi. Elde edilen sonuçlar, yalnızca CLAHE ön işlemenin, Recall'u %12,3 ve mAP@0.5'i %9,1 oranında artırarak en yüksek genel performansı sunduğunu göstermiştir. Dikkat modülleri (özellikle CBAM) doğruluğu %88,4'e kadar yükseltse de Recall'da belirgin düşüşe ve daha düşük mAP artışına yol açmıştır. Bulgularımız, düşük detaylı termal görüntülerde basit kontrast iyileştirmesinin karmaşık dikkat mekanizmalarından daha etkili olabileceğini ortaya koyarak, arama-kurtarma, gözetleme ve savunma uygulamaları için hafif, gerçek zamanlı UAV tabanlı nesne tespit sistemlerinin tasarımına yol göstermektedir.

Anahtar Kelimeler: UAV, termal görüntü, nesne tespiti, YOLOv8, CLAHE, dikkat mekanizmaları

ACKNOWLEDGEMENT

This work was supported by the TÜBİTAK 2209A Research Project Support Programme for Undergraduate Students grant TBTK-0159-7014.

We would like to express our deepest gratitude to our families for their unwavering support and encouragement throughout this journey, and to our advisor, Asst. Prof. Dr. Cahit PERKGÖZ, for his invaluable guidance and insightful advice which have been instrumental in the completion of this thesis.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABBREVIATIONS

AP	Average Precision
BCE	Binary Cross-Entropy
CBAM	Convolutional Block Attention Module
CIoU	Complete Intersection over Union
CNN	Convolutional Neural Network
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CoordAtt	Coordinate Attention
COCO	Common Objects in Context
DCT	Discrete Cosine Transform
ES	Exhaustive Search
F1	F1-score
FPN	Feature Pyramid Network
FLIR	Forward Looking InfraRed
HIT-UAV	High Altitude Infrared Thermal Dataset for Unmanned Aerial Vehicles
IoU	Intersection over Union
LWIR	Long-wave Infrared
mAP	Mean Average Precision
NMS	Non-Maximum Suppression
PAN	Path Aggregation Network
RGB	Red, Green, Blue
SAR	Search and Rescue
SPPF	Spatial Pyramid Pooling-Fast
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once

1. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are rapidly becoming ubiquitous platforms for sensing and data collection across civilian and military domains. Their flexibility, low cost, and ease of deployment allow drones to greatly improve operations such as disaster relief, search-and-rescue (SAR), and wide-area surveillance. For example, Lyu et al. note that UAVs have led to "a significant improvement in the efficiency of search and rescue (SAR) operations," enabling tasks once "difficult or impossible" for humans. Drones' high mobility, ability to hover in place, and low maintenance cost make them ideal for monitoring hazardous or remote areas. Modern UAVs often carry advanced sensors (optical, LiDAR, thermal), extending their utility into challenging conditions. In particular, integrating high-resolution infrared thermal cameras allows UAVs to operate effectively at night or in low-visibility environments. These capabilities have widened the use of UAVs to include border surveillance, infrastructure inspection, traffic monitoring, and environmental reconnaissance (Lyu et al., 2023).

(a) Visual light image (b) Infrared thermal image

Figure 1. The difference between visual and thermal imaging at night (Suo et al., 2023).

In *Figure 1*, the left panel shows a nighttime aerial scene as captured by a visible-light camera; the right panel shows the same scene in infrared thermal. In the thermal image, parked vehicles and other objects appear as bright hot spots, whereas the visible-light image is nearly dark and uninformative. As Suo et al. (2023) demonstrate, thermal infrared cameras "readily identify car and bicycle objects" in night scenes that are nearly invisible to RGB cameras. This inherent advantage makes thermal UAV imaging

especially valuable for night-time search, surveillance, and reconnaissance. Moreover, high-altitude thermal imagery can cover very large areas; for example, datasets collected at 60–130 m altitude include urban and outdoor scenes that span hundreds of meters. In these high-altitude images, a single thermal frame may contain dozens of object instances, enabling UAVs to scan wide regions more efficiently than low-flying aircraft. Together, these factors – thermal contrast and wide coverage – allow UAVs to detect humans, vehicles, and other targets even in darkness or adverse conditions (Suo et al., 2023).

Despite these advantages, object detection in high-altitude UAV thermal imagery remains challenging. Targets such as people and cars often occupy very few pixels and have low contrast, so their thermal signatures can be faint against complex backgrounds. Dash et al. (2025) point out that "thermal signatures can be subtle and easily confused with background noise" in high-altitude imagery. Traditional image-processing methods (e.g. thresholding, edge or blob detectors) and classical machine learning approaches rely on hand-crafted features or simple models, which typically fail under these conditions (Kumar and Singh, 2023). Occlusion, scene clutter, and changes in ambient temperature further confound rule-based detectors. In practice, such conventional techniques yield many missed detections and false alarms in low-light aerial images (Dash et al., 2025). A recent survey of infrared small-target detection confirms that modern deep learning methods substantially outperform traditional algorithms on this problem (Kumar and Singh, 2023). In short, the combination of small object size, sensor noise, and dynamic environments renders thermal UAV data a difficult domain for classic vision pipelines.

Deep learning offers a powerful solution to these challenges. Modern convolutional neural networks (CNNs) learn rich hierarchical features directly from image data, eliminating the need for manual tuning of low-level filters. In the past decade, deep learning has revolutionized object detection: for example, Redmon et al. (2016) introduced the You Only Look Once (YOLO) framework, which frames detection as a single regression problem over bounding boxes and class probabilities. A YOLO network processes an entire image in one pass and predicts all object locations and labels simultaneously, achieving both high accuracy and real-time speed. Such end-to-end detectors can be trained on diverse datasets and generalized to new conditions.

One particularly promising approach is the family of YOLO architectures. YOLO has evolved through many versions (v1–v8), each improving speed and accuracy. Notably, YOLOv8 (the latest Ultralytics model) incorporates key innovations for small and distant object detection. It uses an anchor-free design together with a decoupled head architecture, where separate network branches independently predict objectness, class probabilities, and bounding-box offsets (Terven et al., 2023). This design allows each task-specific head to focus on its subproblem, yielding more precise predictions. As Terven et al. (2023) explain, YOLOv8's anchor-free, decoupled head "improves the model's overall accuracy" by letting the objectness, classification, and regression branches specialize. In addition, YOLOv8 employs advanced loss functions (CIoU, Distributive Focal Loss) and data augmentations that particularly boost small-object detection performance. Combined with a streamlined backbone, these features allow YOLOv8 to achieve state-of-the-art accuracy on benchmarks while running at hundreds of frames per second. Indeed, recent studies note YOLOv8's "accuracy and speed advantages" make it suitable for real-time UAV applications (Terven et al., 2023).

In this thesis, we propose a contrast-enhanced YOLOv8m framework for real-time detection of humans and vehicles in high-altitude UAV thermal imagery. First, we apply CLAHE preprocessing to amplify faint thermal contours. Then, we integrate and compare two attention modules—CBAM and CoordAttention—against the CLAHE-only baseline. Through an ablation study, we demonstrate that simple contrast enhancement alone yields the largest mAP gains, whereas attention mechanisms introduce precision–recall trade-offs under low-detail conditions. Our results provide new methodological guidance for designing lightweight, real-time detectors in low-contrast infrared domains, with direct relevance to search-and-rescue, surveillance, and defense applications.

2. METHODOLOGY

2.1. Data Acquisition and Dataset Description

The HIT-UAV dataset is an open-source high-altitude unmanned aerial vehicle (UAV) infrared thermal image dataset specifically collected for object detection (Suo et al., 2023). It contains 2,898 thermal images (each 640×512 pixels) extracted from 43,470 video frames captured over varied urban and semi-urban scenes (e.g., schools, parking lots, roads, playgrounds). This dataset is publicly available under a CC BY 4.0 license and was designed for UAV-based object detection of pedestrians and vehicles (Suo et al., 2023). All annotated object categories are used in our study, including *Person*, *Car*, *Bicycle*, *OtherVehicle*, and a "*DontCare*" label for ambiguous regions. In total, there are 24,899 annotated object instances in HIT-UAV. The class distribution is unbalanced: The Person category has the most instances, while Car and Bicycle are also substantial; OtherVehicle is less frequent (the "DontCare" instances are typically ignored during evaluation) (Suo et al., 2023).

All images were collected using a DJI Matrice M210 V2 UAV equipped with a dual-sensor DJI Zenmuse XT2 camera (Suo et al., 2023). The XT2 integrates a FLIR long-wave infrared (LWIR) thermal imager with a 640×512–pixel uncooled microbolometer sensor (25 mm lens) and a visible camera (4K video, 12 MP stills). The thermal sensor captures frames at up to 30 Hz (model-dependent) in the 7.5–13.5 μm band, yielding high-sensitivity thermal imagery. The high-altitude platform and LWIR sensor ensure that images are privacy-preserving (humans appear as heat blobs) and effective in varied lighting (Suo et al., 2023).

Data were collected over a wide range of flight conditions. Videos were recorded at altitudes from 60 m to 130 m and at camera pitch angles from 30° to 90°. Recordings were made both during the day and at night to capture different thermal backgrounds (Suo et al., 2023). For each flight and frame, metadata including altitude, camera angle, date, and illumination (day/night) were recorded. These variations result in a dataset that spans diverse scenarios; for example, in *Figure 2*, thermal contrast is typically higher at night, which aids object discrimination (Suo et al., 2023).

(a) Night image sample (b) Day image sample

Figure 2. The samples of the night and day images (Suo et al., 2023).

The entire dataset (2,898 images) was used in our experiments. We partitioned the images randomly into training (70%), validation (10%), and test (20%) subsets. All object classes were retained in each split. This yielded 2018 training images, 299 validation images, and 581 test images with their corresponding annotations (bounding boxes provided in both standard and rotated formats) (Suo et al., 2023). Using this split, the dataset serves as a comprehensive benchmark for high-altitude UAV-based infrared object detection.

2.2. Data Preprocessing

2.2.1. Image Resizing

All thermal images from the HIT-UAV dataset were preprocessed to ensure consistent input and correct annotation format for YOLOv8 training. First, each image was uniformly resized to 512×512 pixels. This fixed size ensures that all input images have the same dimensions, which speeds up batch processing and provides a consistent scale for the neural network. Uniform resizing also aligns with the YOLOv8 configuration, which resizes images to the specified dimension before feeding them into the model (Ultralytics, 2025).

2.2.2. Annotation Formatting and Normalization

The HIT-UAV dataset annotations use a standard bounding-box format given as (x_c, y_c, w, h) – namely the center (x, y) of the box and its width and height (Suo et al., 2023). We parsed the annotation files to extract each object's class label and bounding-box coordinates from this format. To train YOLOv8, these values were converted into the YOLO label format: each object is represented by `<class> <x_center> <y_center> <width> <height>` with all coordinates normalized to the [0,1] range (Torres, 2025). In practice, this means dividing the box center coordinates and dimensions by the image width and height so that they are scale-invariant. Using normalized coordinates makes the training more stable, since the network predicts relative locations rather than pixel units (Torres, 2025). In summary, we read the annotation files, extracted (x_c, y_c, w, h) for each object, and then normalized these values by the image size to produce properly formatted YOLO labels.

2.2.3. Visualization

After resizing and label normalization, we visually verified the correctness of the annotations. We overlaid the bounding boxes on the corresponding thermal images to confirm that each box aligned with the target object. For example, the HIT-UAV codebase includes a visualization script for this purpose (Suo et al., 2023). By plotting the boxes on images before training, we could detect any parsing or labeling errors (such as misplaced boxes or incorrect classes) early. This visual check acts as a sanity check on the data and helps ensure that the model does not train on faulty annotations.

2.3. Contrast-Limited Adaptive Histogram Equalization (CLAHE)

Thermal images often exhibit low and uneven contrast, which can make object features subtle. As a possible enhancement, we considered applying Contrast-Limited Adaptive Histogram Equalization (CLAHE) to the input images. Unlike simple global adjustments, CLAHE works on localized regions of the image. Specifically, it divides the image into small tiles and applies histogram equalization independently to each tile (MathWorks, n.d.). By doing so, CLAHE increases local contrast in each region rather

than adjusting the whole image at once. Importantly, CLAHE imposes a contrast limit (clip limit) during equalization, which prevents any region from becoming over-enhanced and avoids amplifying noise (Li et al., 2024). The result is a more even enhancement of contrast across the image without creating overly bright or dark patches.

By contrast, simple image thresholding would binarize the image based on a single intensity cutoff. In thresholding, each pixel is set to black or white depending on whether its value is below or above the threshold (Buhl, 2023). While thresholding can separate foreground from background, it discards most of the grayscale information in the image. CLAHE, on the other hand, preserves the full range of intensity values while stretching them to improve local visibility. In other words, instead of converting the image to just two levels, CLAHE adaptively "spreads out" the local intensity distribution. This can reveal subtle thermal signatures (for example, the faint heat outline of a person or vehicle) that would be lost under hard thresholding.

(a) Original image (b) CLAHE-enhanced image

Figure 3. Original and CLAHE-enhanced thermal images from the HIT-UAV dataset.

Because CLAHE enhances contrast adaptively, it can make objects in low-contrast thermal scenes stand out more clearly against the background. It does so without saturating the entire image; the contrast limiting ensures that details and shadows are retained. Consequently, applying CLAHE may help YOLOv8 better distinguish objects in the thermal images by making their shapes and edges more pronounced. As illustrated in *Figure 3*, the original thermal image and its CLAHE-enhanced version clearly demonstrate this effect. The CLAHE-applied image reveals object boundaries and fine details that are less discernible in the original. In summary, unlike a fixed threshold that yields a binary segmentation, CLAHE offers a nuanced contrast enhancement that can boost object visibility in thermal imagery without introducing noise or artifacts.

2.4. Object Detection Models

2.4.1 YOLOv8 Baseline Model

The YOLOv8 model builds on the YOLO series with a modern CSP backbone and a hybrid FPN+PAN neck for efficient multi-scale feature fusion (Yaseen, 2024). Refer to *Figure 4* for a visual representation of the architecture, which illustrates how the CSP backbone enhances feature extraction and the FPN+PAN neck optimizes multi-scale detection. Importantly, its detection head is *anchor-free*: it no longer relies on predefined anchor boxes but instead directly predicts object centers and dimensions. As Yaseen (2024) notes, this anchor-free design “simplifies the prediction process, reduces the number of hyperparameters, and improves the model’s adaptability to objects with varying aspect ratios and scales”. In practice, removing anchors reduces computational complexity by eliminating the overhead of generating and refining anchors. The net result is a streamlined head that can more flexibly learn to localize objects of diverse shapes and sizes.

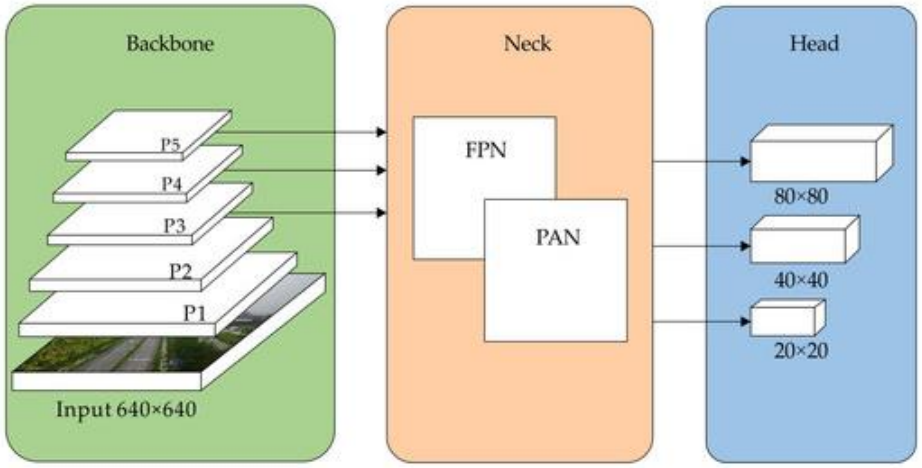


Figure 4. YOLOv8m architecture diagram illustrating the CSP backbone, FPN/PAN neck, and detection head (Li et al., 2025).

YOLOv8 also employs a *decoupled detection head*. In this design, the head splits into two parallel branches: one branch predicts object classes and confidences, and the other predicts bounding box coordinates. This contrasts with earlier "coupled" heads that share features for both tasks. As described in Ultralytics (2025) documentation, separate (decoupled) branches allow each task to specialize, which can improve overall accuracy. In practice, the classification branch focuses on distinguishing object categories, while the regression branch refines precise box localization. This architectural choice (inherited from other modern detectors) helps YOLOv8 balance accuracy and speed on detection tasks.

Among the YOLOv8 variants, we selected the *YOLOv8m (medium)* model as our baseline. This variant offers a compromise between model capacity and efficiency. Compared to smaller versions (e.g. YOLOv8s or n), YOLOv8m has more parameters and thus higher representational power. Compared to larger versions (YOLOv8l or x), YOLOv8m is significantly smaller. Given the need to process UAV imagery efficiently, YOLOv8m was chosen as a balanced trade-off: it is large enough to capture complex object patterns in aerial views, yet small enough for faster inference and training on available hardware.

For training, we followed the standard YOLOv8 setup. The model was trained for 50 epochs with a batch size of 16 and input images resized to 512×512 pixels. Mixed-precision (half-precision, 16-bit) arithmetic was used throughout to accelerate training

and reduce GPU memory usage. The choice of 50 epochs and moderate image size reflects common defaults for medium-scale YOLOv8 experiments. Other hyperparameters (learning rate, augmentation, etc.) were kept as in the reference implementation. The key hyperparameters are summarized in *Table 1* below.

Table 1. Key YOLOv8m training hyperparameters

Hyperparameter	Value	Default Value	Description
Epochs	50	100	Number of times the model sees the entire dataset during training.
Batch Size	16	32	Number of samples processed before the model's parameters are updated.
Image Size	512×512	640×640	Resolution to which all images are resized.
Mixed Precision	16-bit	32-bit	Use of half-precision to speed up training and reduce memory usage.
Learning Rate	0.01	0.01	Step size at each iteration while moving towards a minimum.
Augmentations	Default	Default	Standard YOLOv8 augmentations for data variability.
Loss Function	CIoU + BCE	CIoU + BCE	Combination of CIoU for bounding box regression and BCE for classification.

2.4.1.1 YOLOv8 Loss Function

The loss function in YOLOv8 is critical to its training process, as it guides the model in learning accurate object localization and classification. It is composed of two main components: *Bounding Box Regression Loss* and *Classification Loss*, each optimized to address the unique challenges of object detection.

YOLOv8 uses a sophisticated bounding box regression approach, relying primarily on the *Complete Intersection over Union (CIoU) Loss*. Unlike simpler overlap-based metrics, CIoU accounts not only for the overlap area but also for the center distance and aspect ratio of predicted and ground truth boxes. This makes it more sensitive to both spatial alignment and shape consistency, encouraging the model to produce tighter, more precise bounding boxes. The CIoU formulation improves localization performance, especially for objects with varying aspect ratios, and helps

reduce false positives and missed detections by incorporating a more comprehensive measure of box similarity (Zheng et al., 2020).

YOLOv8 employs a *Binary Cross-Entropy (BCE) Loss* for classification, which is applied independently to each class prediction. Unlike earlier YOLO versions, YOLOv8's classification head is fully decoupled from its localization head, meaning it handles class probabilities separately from box coordinates. This design choice eliminates the need for an explicit "objectness" score, simplifying the model architecture while maintaining high accuracy. By separating these tasks, YOLOv8 reduces the risk of feature interference and improves training stability (Ge et al., 2021).

Together, these two loss components are calculated in parallel, with separate branches for classification and regression. This decoupled design allows each component to specialize in its respective task, enhancing overall model accuracy and reducing computational overhead. During training, the total loss is computed as a weighted sum of these components, with each term scaled according to hyperparameters defined in the training configuration.

2.4.2 YOLOv8 with CBAM Attention Module

To enhance feature representation, we integrate the Convolutional Block Attention Module (CBAM) into the YOLOv8m model. CBAM is a lightweight, plug-in attention mechanism that sequentially applies channel attention and spatial attention to intermediate feature maps (Woo et al., 2018). In CBAM, the input feature tensor is first passed through a *channel attention* submodule (which learns to weight each feature channel) and then a *spatial attention* submodule (which learns to weight each spatial location). This sequential inference of channel- and spatial-attention maps adaptively refines the convolutional features with negligible computational overhead (Woo et al., 2018). Because CBAM is both lightweight and end-to-end trainable, it can be inserted into CNN architectures (like YOLOv8) without disrupting the base network's training process (Woo et al., 2018).

In our proposed architecture (denoted YOLOv8m+CBAM), the CBAM block is inserted immediately after the Spatial Pyramid Pooling – Fast (SPPF) layer at the end of the YOLOv8m backbone (i.e., just before the neck). The SPPF layer aggregates multi-

scale contextual features from the backbone, effectively pooling information at different scales (Torres, 2025). By placing CBAM after SPPF, the attention module operates on these rich, multi-scale feature maps. This allows CBAM to recalibrate salient feature channels and spatial regions before they are merged in the neck. In other words, CBAM helps the network focus on the most informative elements of the high-level feature map. A schematic of the modified YOLOv8m architecture with the CBAM block is shown in *Figure 5*.

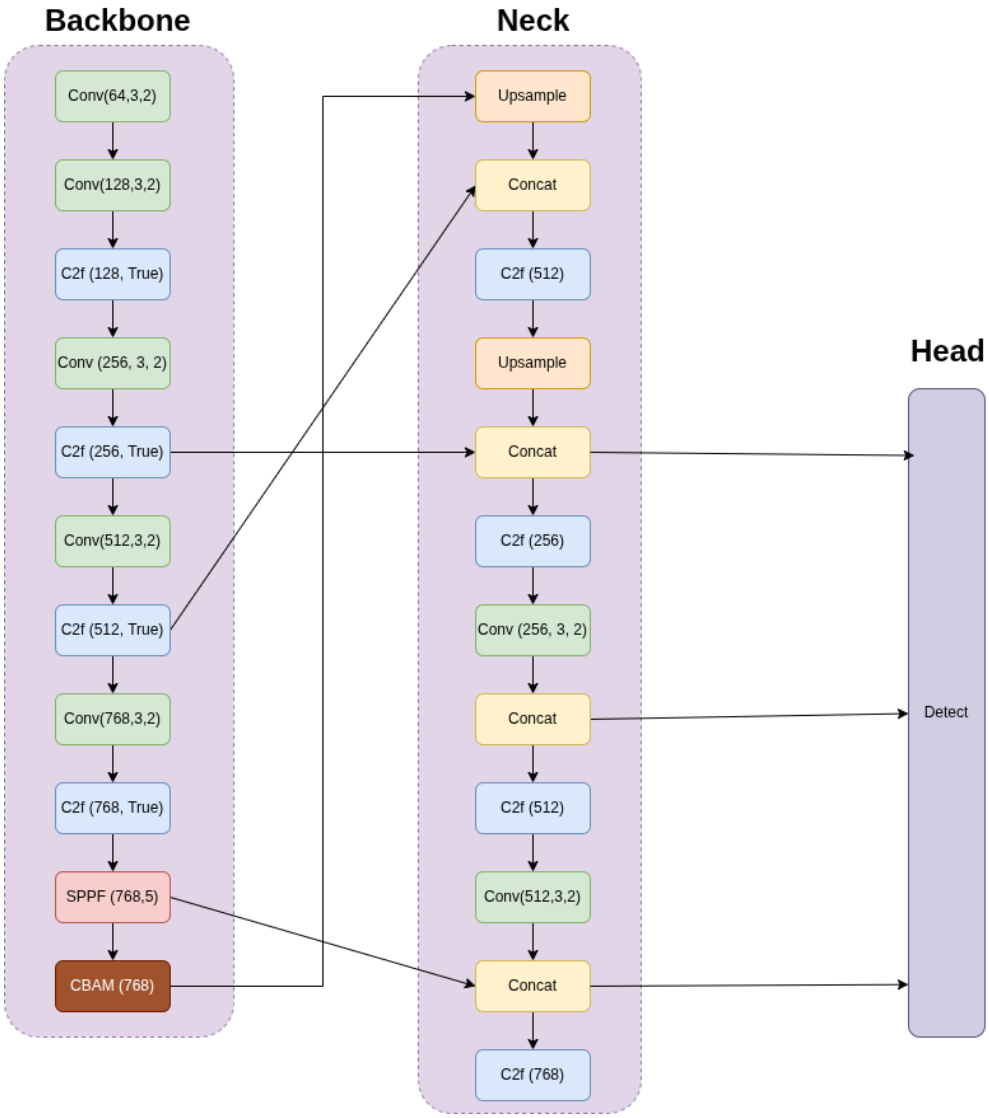


Figure 5. A visual illustration of the YOLOv8m+CBAM architecture.

The expected advantage of this modification is improved focus on small or partially occluded objects in UAV imagery. In aerial data, object sizes vary widely and targets may be obscured by clutter or overlap. The dual channel-and-spatial attention of

CBAM can help the detector emphasize relevant parts of the feature map that correspond to small or hidden objects. Previous work has shown that adding CBAM to YOLO-based detectors yields better attention to key features and can boost accuracy on challenging targets (Li et al., 2024). For example, Li et al. (2024) note that a YOLOv8 model with an added CBAM module was better able to detect small and occluded targets in UAV images. Similarly, Tahir et al. report that integrating CBAM into a YOLOv8-based model improved mAP by enabling the network to focus on previously misdetected objects. We therefore anticipate that the YOLOv8m+CBAM model will more effectively attend to informative regions across scales, leading to improved detection of small or partially occluded objects in the UAV domain.

All other aspects of the network and training regimen remain the same as in the YOLOv8m baseline. In particular, we retain the identical hyperparameters used for the baseline model (see Table 1 for details) to ensure a fair comparison. In summary, the proposed YOLOv8m+CBAM architecture augments the baseline by inserting a CBAM attention block after the backbone's SPPF layer. This design is intended to emphasize salient multi-scale features and thereby improve detection accuracy on small or obscured UAV targets.

2.4.3 YOLOv8 with CoordAtt Module

Coordinate Attention (CoordAtt) is a recent attention mechanism that augments channel attention with precise spatial context (Hou et al., 2021). Unlike conventional channel-only attention (e.g. SE) that pools spatially, CoordAtt factorizes the attention process into two parallel 1D streams along the height and width axes (Hou et al., 2021). Concretely, the module performs two separate global pooling operations – one aggregating features across each column (height) and one across each row (width) – yielding two "direction-aware" feature maps. Each pooled feature map is then processed to generate an attention map that captures long-range dependencies along one spatial direction while preserving precise positional information along the other (Hou et al., 2021). The two resulting attention maps are applied multiplicatively to the original feature tensor, creating a coordinate-aware representation that highlights important regions without losing object location cues. Hou et al. demonstrate that this design is

computationally simple and incurs nearly no extra overhead when plugged into common backbones, yet significantly improves downstream performance (e.g. object detection and segmentation) compared to standard attention blocks (Hou et al., 2021).

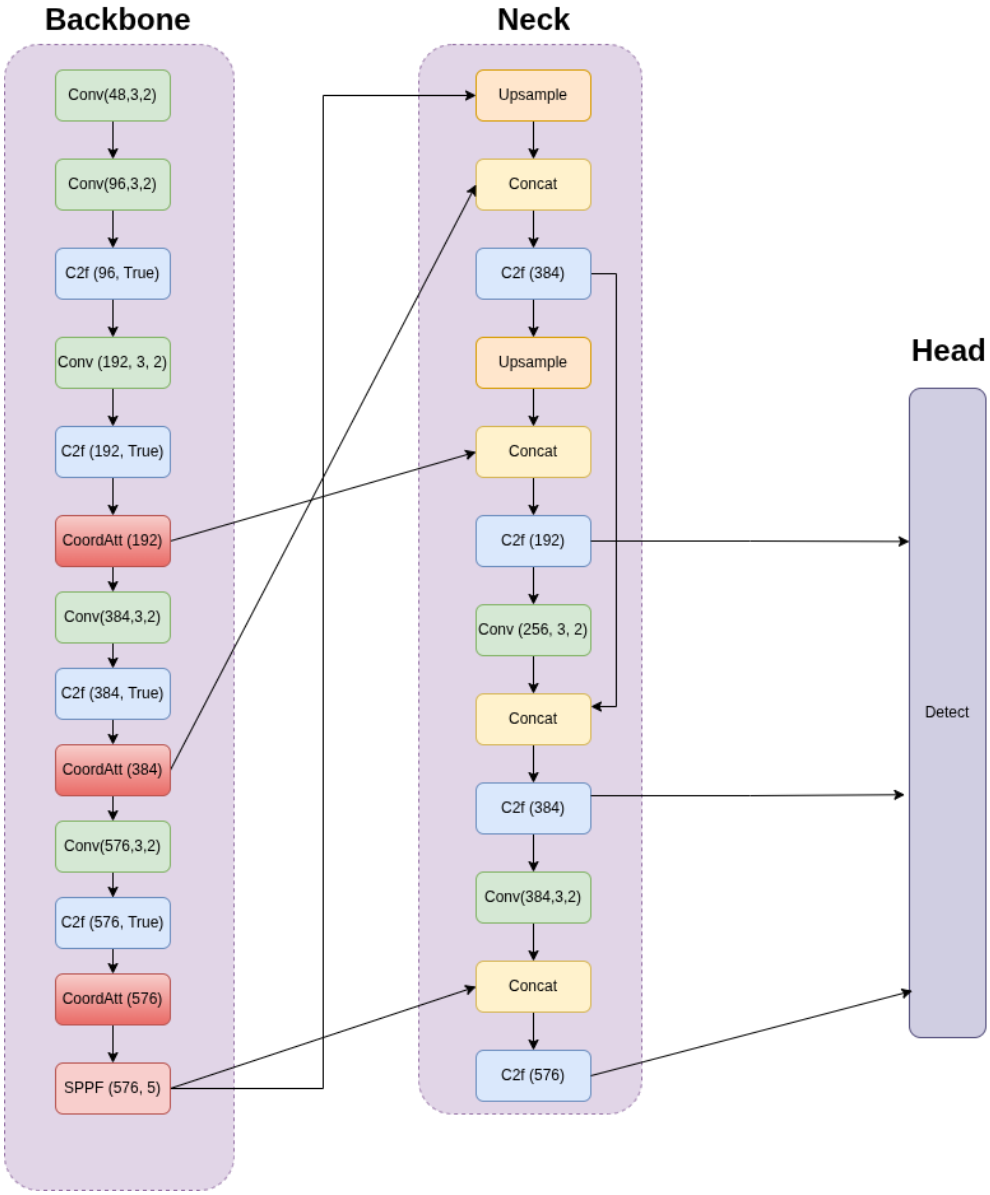


Figure 6. A visual illustration of the YOLOv8m+CoordAtt architecture.

This coordinate-aware approach is especially promising for UAV-based object detection, where targets tend to be small, elongated, or sparsely distributed in high-resolution aerial scenes. By preserving fine-grained spatial information along each axis, CoordAtt helps retain localization cues that might otherwise be lost in global pooling. In

practice, UAV imagery often features tiny ground objects (vehicles, people, etc.) against complex backgrounds, so an attention mechanism that explicitly encodes where features appear can improve detection accuracy. Indeed, recent studies in aerial vision report that adding CoordAtt enhances small-target localization: for example, Xu et al. (2024) note that encoding feature channels separately along horizontal and vertical axes "enhances the network's ability to locate targets" in a YOLOv8n UAV detector. Similarly, Wang et al. (2024) observe that CoordAtt "contemplates both inter-channel dependencies and the significance of spatial features for accurate object localization," providing broader spatial context than CBAM or SE and yielding more robust detection in cluttered UAV scenes. These findings suggest that inserting CoordAtt into YOLOv8m model could improve sensitivity to small or narrow objects by emphasizing their spatial signatures.

To evaluate this, we constructed a YOLOv8m+CoordAtt variant by integrating CoordAtt blocks into the backbone of YOLOv8m. Specifically, we inserted a CoordAtt module after each C2f stage in the backbone, i.e. immediately following the feature layers that output 192, 384, and 576 channels respectively. In each case the CoordAtt block takes the preceding feature map, applies the height- and width-axis attention operations described above, and outputs the same number of channels back into the network (see *Figure 6*). This ensures that multi-scale feature maps in the backbone are modulated by coordinate attention before proceeding to the neck and head. The resulting architecture (depicted in *Figure 6*) is otherwise identical to the baseline YOLOv8m. In particular, all training settings and hyperparameters were kept the same as in the original YOLOv8m experiments (see Table 1), so that any performance differences can be attributed to the architectural change.

2.5. Implementation Details

The implementation was carried out using the Python programming language in a Google Colaboratory (Colab) Pro environment. Colab provides a hosted Jupyter notebook interface with free GPU acceleration, making it well suited for deep learning tasks (Google Colab, n.d.). Key libraries included PyTorch (an optimized tensor library for deep learning) and the Ultralytics YOLOv8 library (for model definition and training), along with OpenCV, NumPy, and Matplotlib for image processing and

visualization. All codes were written in Python 3.11. Data handling made use of standard NumPy arrays and OpenCV image operations, and Matplotlib was used to generate charts and figures for results (Matplotlib is a comprehensive plotting library for Python).

Training was executed on a Google Colab Pro GPU instance NVIDIA A100. The dataset and model were loaded into GPU memory for accelerated training. After training, the best-performing model weights were saved for evaluation on the held-out validation and test sets.

2.6. Evaluation Metrics

The object detector is evaluated using standard metrics that capture different aspects of detection performance. *Precision* measures the proportion of predicted detections that are correct. In other words, it reflects how often the model's positive predictions are true positives (few false positives). *Recall* measures the proportion of all ground-truth objects that are successfully detected. Thus, high recall means the model finds most of the true objects (few false negatives), whereas high precision means its reported detections are mostly accurate. In summary, precision can be seen as a quality measure (few false alarms) and recall as a completeness measure (few missed objects).

To combine these two aspects into a single metric, the *F1-score* is used. The F1-score is defined as the harmonic mean of precision and recall. This means the F1-score is high only when both precision and recall are high, effectively balancing the trade-off between false positives and false negatives. In practice, the F1-score provides a single summary of detection accuracy by punishing the model if either precision or recall is low.

Another key metric for object detection is *mean Average Precision (mAP)*, which summarizes the precision–recall performance across classes. Average Precision (AP) is computed for each object category by integrating the precision–recall curve. The mAP is then the mean of these AP values over all classes. In our evaluations, we report two variants: *mAP@0.5* and *mAP@0.5:0.95*. The notation "@0.5" indicates that a detection is counted as correct only if its bounding box overlaps the ground truth by at least 50% (Intersection-over-Union of 0.5). Thus, mAP@0.5 is the average precision using this

fixed overlap threshold. In contrast, *mAP@0.5:0.95* is the average of mAP values computed over multiple thresholds from 0.5 up to 0.95 (in 0.05 increments). Using mAP@0.5:0.95 provides a more comprehensive assessment, since it requires the model to perform well across stricter overlap criteria. Overall, higher mAP values indicate better detection accuracy and robustness.

2.7. Comparative Analysis

In this study, we compare the medium-sized YOLOv8m baseline against two enhanced variants that incorporate attention modules (CBAM and CoordAtt) and were all trained identically on the HIT-UAV dataset with COCO-pretrained weights. he weight-transfer logs from model initialization reveal how many pretrained weights were reused. The training output reported “Transferred 234/478 items from pretrained weights” for YOLOv8m+CBAM and 135/505 for YOLOv8m+CoordAtt. These numbers denote that, out of 478 total learnable tensors in the CBAM model, only 234 matched the original COCO-pretrained architecture; similarly, only 135 of 505 tensors matched in the CoordAtt model. The remaining tensors (corresponding to the newly added attention layers) had no counterpart in the COCO weights and were therefore randomly initialized. This phenomenon is expected when modifying the network: any inserted modules break the correspondence with the pretrained model, so fewer weights can be loaded.

3. RESULTS AND DISCUSSION

3.1. Quantitative Results

Table 2 reports the key detection metrics for YOLOv8m variants, evaluated on 299 test images containing 2,453 ground-truth boxes.

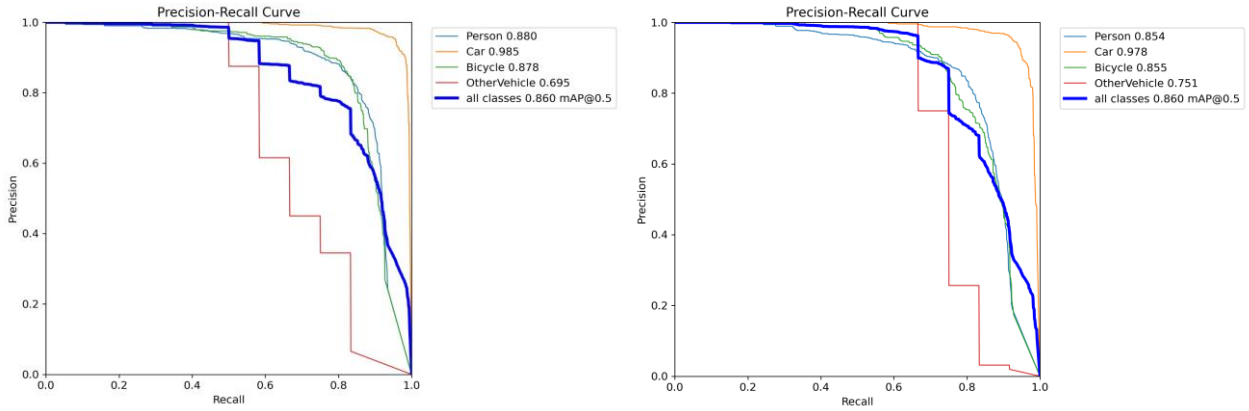
Table 2. Quantitative detection metrics for YOLOv8m variants on thermal images

Model	Precision (%)	Recall (%)	F1-score	mAP@0.50 (%)	mAP@0.5:0.95 (%)
YOLOv8m baseline	80.3	76.1	78.1	79.3	53.4
CLAHE+ YOLOv8m	81	88.4	84.5	88.4	61.1
CLAHE+ YOLOv8m+CBAM	88.4	78.1	82.9	86	57.8
CLAHE+ YOLOv8m+CoordAtt	85.6	79.1	82.2	86	57

Applying CLAHE alone yields a modest boost in precision (80.3 → 81.0 %) but a substantial gain in recall (76.1 → 88.4 %), pushing the F1-score from 78.1 to 84.5 %. This suggests that contrast enhancement makes faint targets more detectable, with fewer missed objects and a 9.1 % absolute improvement in mAP@0.50 (79.3 → 88.4 %) and a 7.7 % gain in the stricter mAP@0.5:0.95.

Inserting CBAM after CLAHE raises precision sharply to 88.4 %, indicating far fewer false positives. However, recall drops to 78.1 %, so the F1-score settles at 82.9 %. The mAP@0.50 (86.0 %) remains above the baseline but below the CLAHE-only model, illustrating that channel–spatial attention sharpens detections but at the cost of missing some objects.

CoordAtt yields a balance between the previous two: precision of 85.6 % and recall of 79.1 % (F1 = 82.2 %). Its mAP@0.50 (86.0 %) matches CBAM's, while mAP@0.5:0.95 (57.0 %) is slightly lower, suggesting that coordinate-aware weighting helps localize some small targets better than CBAM but does not recover all the recall lost by attention.



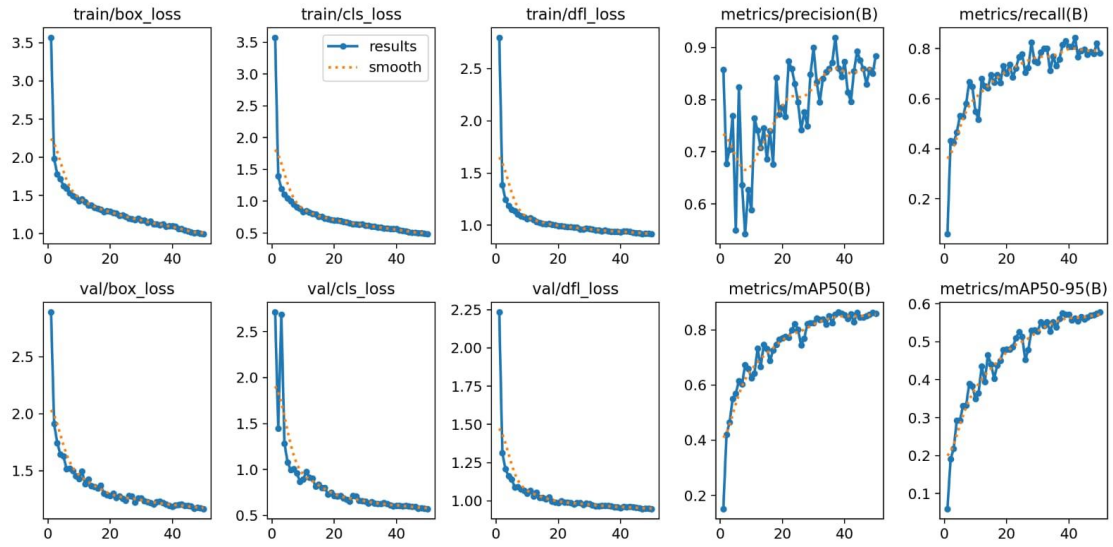
(a) (b)

Figure 7. Precision-Recall curves for the models: (a) CLAHE+ YOLOv8m+CBAM and (b) CLAHE+ YOLOv8m+CoordAtt.

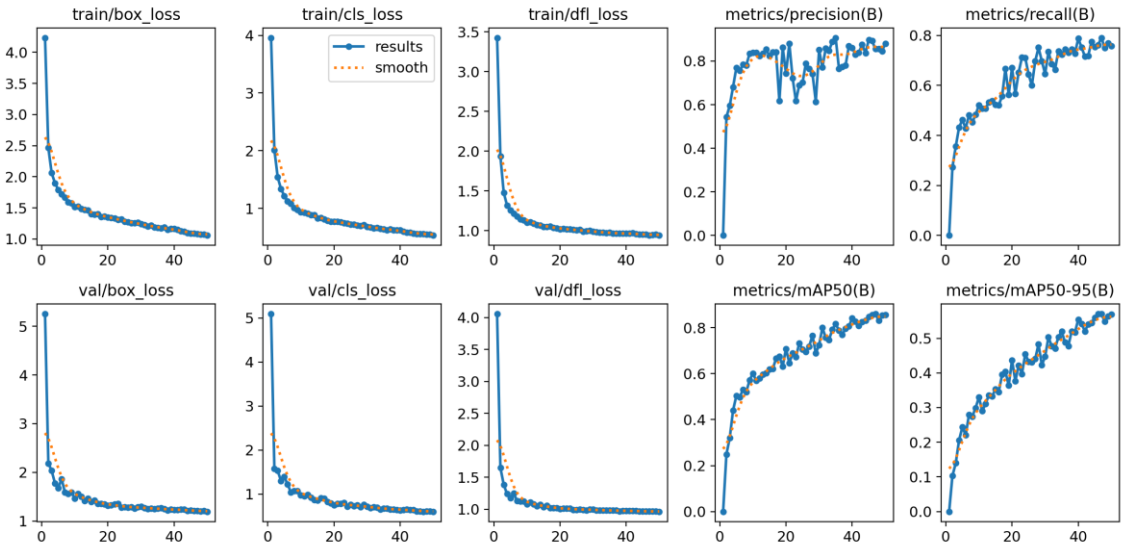
In *Figure 7a*, the curve starts at very high precision (> 95 %) for low recall (< 20 %), demonstrating that its highest-confidence predictions are extremely reliable. Precision then gradually declines as recall rises, crossing 80 % precision at around 60 % recall. The area under the curve (AUC) reflects its strong precision bias but limited maximum recall (~ 78 %).

In *Figure 7b*, this model attains its peak precision (~ 90 %) at lower recall levels (< 30 %) but maintains precision above 75 % even as recall approaches 80 %. Its AUC is slightly lower than CBAM's at high-precision thresholds but exceeds CBAM's AUC in the mid-recall range (50–80 %), indicating a more consistent trade-off. These curves confirm that CBAM emphasizes precision at the cost of recall, whereas CoordAtt offers a more balanced precision–recall profile.

CLAHE alone provides the largest boost in recall and overall mAP by enhancing faint thermal signatures. Adding CBAM shifts the balance toward precision, pruning false positives sharply but incurring more missed detections, while CoordAtt achieves a steadier precision–recall balance.



(a)



(b)

Figure 8. Training-process curves (a) CLAHE+ YOLOv8m+CBAM and (b) CLAHE+ YOLOv8m+CoordAtt.

Figure 8 shows that both CLAHE + YOLOv8m + CBAM (*Figure 8a*) and CLAHE + YOLOv8m + CoordAttention (*Figure 8b*) converge smoothly by around epoch 40, with bounding-box, classification, and distribution focal losses all decreasing steadily and validation curves closely tracking the training curves—an indication of

minimal overfitting. In the CBAM variant, precision fluctuates in the early epochs before climbing to approximately 0.90, while recall rapidly rises to about 0.80 and then plateaus; correspondingly, mAP@0.50 and mAP@0.5:0.95 reach roughly 0.88 and 0.58, respectively, reflecting CBAM's bias toward high-confidence detections. In contrast, the CoordAttention model exhibits a more uniform increase in both precision (settling near 0.87) and recall (around 0.79), leading to slightly lower peak mAP values (≈ 0.86 at 0.50 IoU and ≈ 0.57 at 0.5:0.95 IoU) but with fewer early-stage oscillations. Overall, both attention-augmented architectures train reliably, with CBAM offering marginally higher final precision and mAP at the expense of greater volatility, while CoordAttention provides a steadier, more balanced precision–recall trajectory.

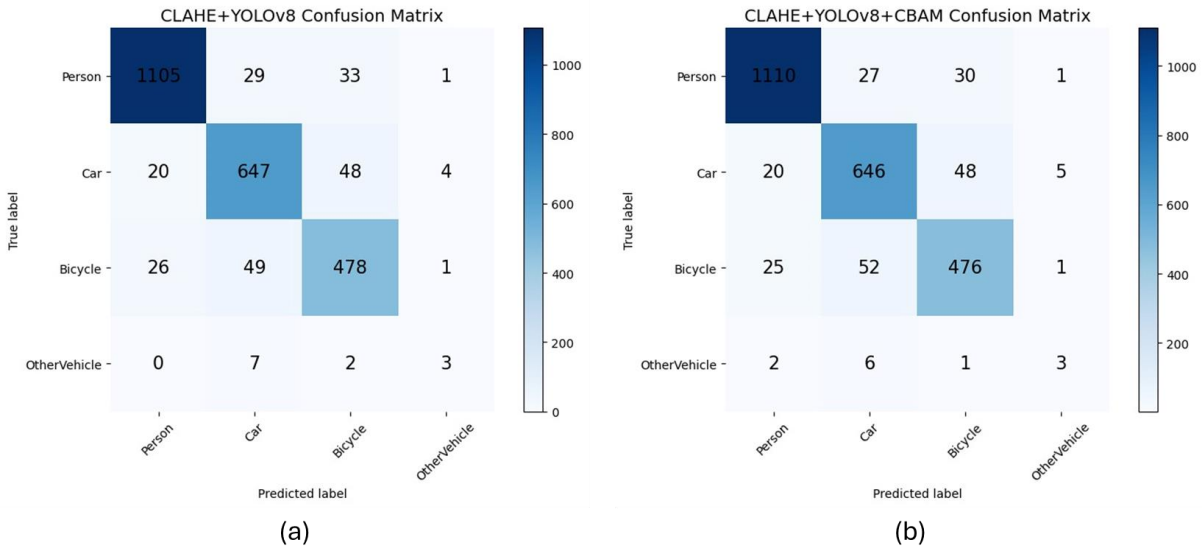
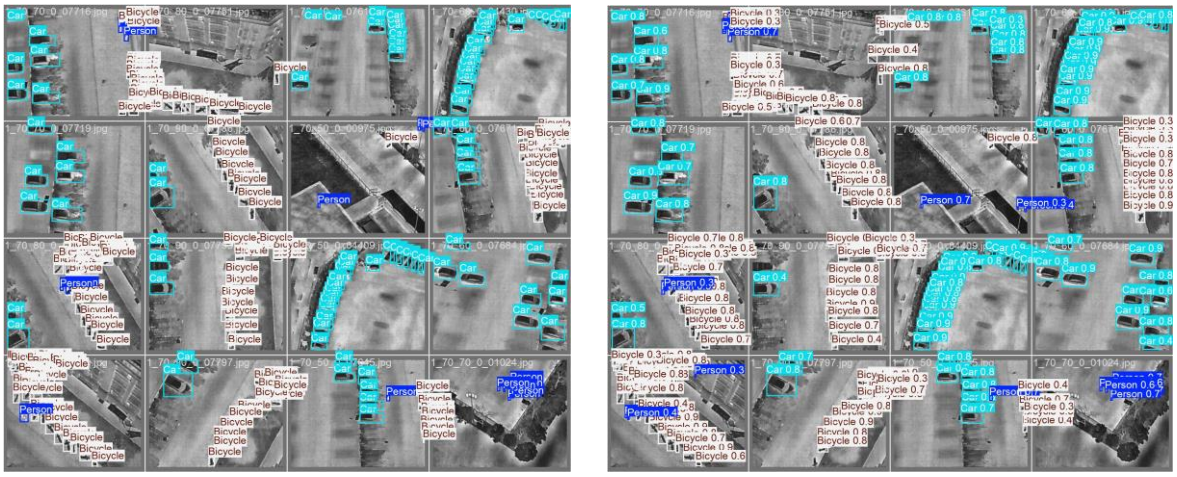


Figure 9. Confusion matrices on the test set for (a) CLAHE+YOLOv8m and (b) CLAHE+ YOLOv8m+CBAM.

Figure 9 evaluates class-level performance on the held-out test set. In the baseline CLAHE+YOLOv8m model (*Figure 9a*), the "Person" and "Car" categories dominate correct detections but still exhibit notable cross-class confusions—most frequently with each other and with "Bicycle"—while the "OtherVehicle" class remains underrepresented. Integrating CBAM (*Figure 9b*) further sharpens decision boundaries: true positives for the primary classes increase slightly, and misclassifications between "Person," "Car," and "Bicycle" decrease, yielding a cleaner separation of these

categories. The sparse “OtherVehicle” detections remain largely unchanged. Overall, the CBAM-augmented model delivers a more distinct, less ambiguous classification profile across all four object types.

3.2. Qualitative Results



(a) (b)

Figure 10. Detection results of the CLAHE+YOLOv8m+CBAM model: (a) ground-truth bounding boxes and (b) predicted bounding boxes.

This *Figure 10* presents a representative detection result from the CLAHE+YOLOv8m+CBAM model on a thermal image. As the other model variants (i.e., baseline YOLOv8m, CLAHE+YOLOv8m, and CLAHE+YOLOv8m+CoordAtt) yielded visually similar detection patterns in most scenarios, only one qualitative example is shown for illustration purposes.

3.3. Ablation Study

As shown in *Table 2*, applying CLAHE produces a modest improvement in precision but a substantial gain in recall and overall mAP. Specifically, the CLAHE+YOLOv8m model raises recall from 76.1% to 88.4% and increases mAP@0.50 from 79.3% to 88.4%. This indicates that contrast enhancement via CLAHE makes faint thermal targets more detectable, effectively reducing missed

detections. In effect, amplifying subtle thermal gradients allows the network to learn object features that were otherwise difficult to discern in the baseline imagery. These results underscore the value of simple contrast processing: by highlighting weak signals, CLAHE significantly boosts detectability of small or low-contrast objects in high-altitude thermal scenes.

Comparing CLAHE+YOLOv8m with the version including CBAM reveals a clear trade-off in performance. *Table 2* shows that adding CBAM raises precision sharply to 88.4% (up from 81.0% without CBAM) while recall drops to 78.1% (down from 88.4%). *Figure 7a* illustrates this behavior: the CBAM-enhanced model achieves over 95% precision at very low recall levels (<20%), but precision steadily declines as more objects are recalled. In other words, CBAM emphasizes only the strongest feature activations, pruning many false positives but suppressing weaker thermal signals. As a result, the overall mAP@0.50 (86.0%) remains above the baseline but below the CLAHE-only model. This suggests that CBAM's channel–spatial attention sharpens detections at the expense of missing some faint targets, yielding higher precision but lower recall compared to the simpler CLAHE-enhanced network.

A similar analysis applies to the coordinate attention (CoordAtt) variant. *Table 2* indicates that adding CoordAttention yields precision of 85.6% and recall of 79.1%, with mAP@0.50 of 86.0%. The Precision–Recall curve for CLAHE+YOLOv8m+CoordAtt (*Figure 7b*) shows a more balanced trade-off: precision peaks near 90% at low recall and remains above 75% even at 80% recall. CoordAttention is designed to preserve spatial locality by encoding coordinate information into the attention weights, which can improve localization of elongated or sparsely-represented objects. However, this spatially-aware attention still attenuates weaker signals relative to the CLAHE-only model, limiting recall. Thus, the overall mAP (86.0%) matches the CBAM-enhanced result but does not exceed the contrast-enhanced baseline. In practice, CoordAtt achieves a steadier precision–recall profile (better balance) than CBAM, but its enhancements are similarly offset by the increased missed detections.

In summary, although each attention-enhanced variant introduced targeted improvements (higher precision with CBAM, a steadier precision–recall balance with CoordAtt), the best overall performance was achieved by the CLAHE-enhanced

YOLOv8m model without any attention module. The CLAHE-only model delivered the largest recall and mAP gains by making weak thermal objects more detectable. These findings suggest that, for high-altitude UAV thermal imagery where object details are limited, simple contrast enhancement can outweigh the advantages of more complex attention mechanisms. In other words, amplifying subtle thermal signatures via CLAHE proved more effective than relying on attention modules to find those signals, emphasizing that complexity should be carefully justified in low-detail thermal object detection.

3.4. Comparison with Existing Studies and Contribution to the Literature

Recent work on thermal UAV object detection provides useful context for our results. For example, Dash et al. (2025) apply YOLOv8 with dynamic magnitude-based pruning and NMS optimizations to high-altitude infrared images. They report very high accuracy (up to 99.3% overall) and per-class mean average precision (mAP) values exceeding 94% on most classes. Importantly, Dash et al. note that thermal UAV images "lack distinct visual features" and suffer from reduced resolution at altitude. Similarly, Kumar and Singh's (2023) comprehensive review of small/dim IR target detection emphasizes that cluttered backgrounds and ambiguous thermal signatures make detection difficult. They conclude that deep-learning methods typically outperform classical IR-target detectors. These studies together suggest that advanced models and preprocessing (like pruning/NMS or deep networks) can improve performance, but also that IR images inherently have low-detail challenges (Dash et al., 2025; Kumar and Singh, 2023). Our approach aligns with this context: we also use a deep model (YOLOv8m) but focus on pre-processing and attention, and we report competitive detection accuracy on a similar high-altitude thermal dataset.

In our experiments, the YOLOv8m variant with only localized contrast enhancement (CLAHE) attained the highest mAP, outperforming the variants that included channel/spatial attention modules. Attention mechanisms like CBAM and CoordAttention are theoretically motivated to improve feature representation (e.g. Woo et al. (2018) showed CBAM improves detection performance, and Hou et al. (2021) demonstrated that coordinate attention enhances object detection). However, when

applied to our low-detail thermal images, adding CBAM or CoordAttention did not yield further gains. In fact, the CLAHE+YOLOv8m baseline (without any attention module) achieved the best results. This suggests that in the context of high-altitude infrared imagery, the extra capacity of attention modules may be less beneficial than simply boosting the local contrast of the input. In other words, CLAHE's local histogram equalization made subtle thermal object cues more detectable, whereas the inductive biases of CBAM/CoordAttention did not overcome the scarcity of strong features in these images.

Our architectural choices and research goal reflect these insights. Rather than solely chasing maximum accuracy, we explicitly examined how contrast enhancement and attention behave under constrained conditions. We deliberately incorporated CBAM and CoordAttention into some model variants as a test: if high-level attention could compensate for low-level detail loss. The study revealed that CLAHE's localized contrast boost alone provided the highest mAP on our thermal dataset, outperforming all attention-augmented versions. This finding is consistent with the observation that high-altitude thermal pixels often lack distinct features. In such a low-information environment, enhancing contrast at the input stage can be more effective than additional attention layers.

Our findings contribute to a better understanding of how contrast enhancement and attention mechanisms function in low-detail thermal UAV imagery. In particular, we show that while CBAM and CoordAttention are effective in other domains, they did not yield performance gains over CLAHE-enhanced YOLOv8m in our setting. This suggests that in high-altitude infrared scenes with sparse spatial detail, boosting local contrast (via CLAHE) offers more tangible benefits than adding channel or coordinate attention. By contextualizing our results with prior studies such as Dash et al. (2025) and Kumar and Singh (2023), we highlight that image-level enhancement can be a more impactful strategy than architectural complexity in low-information scenarios. Thus, our study provides empirical guidance on selecting enhancement strategies for infrared object detection under constrained visual conditions.

In summary, this study contributes valuable insights rather than asserting absolute superiority. It demonstrates that in high-altitude thermal imagery—characterized by limited pixel detail and contrast—prioritizing adaptive contrast enhancement (CLAHE)

can be more effective than adding attention modules. These findings refine our understanding of model design in the IR UAV domain: attention mechanisms are useful in general, but their benefit depends on the input quality, and contrast enhancement can be critical when thermal features are scarce.

3.5. Discussion and Limitations

An unexpected result of this study was that CLAHE-enhanced YOLOv8m outperformed the versions incorporating CBAM and CoordAttention. Although attention modules are known to improve feature discrimination by emphasizing important regions, they introduced additional parameters and computational complexity. Given the low-information nature of high-altitude thermal imagery-characterized by low contrast and weak texture-this complexity did not translate into better performance. In contrast, CLAHE improved object visibility at the input level by enhancing local contrast, leading to more effective detection without increasing model capacity.

Despite these promising findings, the present study has several notable limitations. First, the dataset size was relatively small by deep-learning standards, which restricts generality. A modest number of training images makes it difficult to fully train large neural networks and increases the chance that the model will overfit to idiosyncrasies of the data. Second, the data exhibited high variability due to changing UAV flight altitude and angle. In practice, thermal UAV datasets often span a wide altitude range (for example, 60–130 m in HIT-UAV (Suo et al., 2023)) and diverse camera perspectives. Such variation means that object scale and appearance can change dramatically from frame to frame, which complicates learning and can degrade accuracy. Third, we used only a single spectral band (thermal IR) for detection. Relying solely on thermal intensity omits complementary cues (such as visible color or multispectral information) that could help distinguish objects. In short, our evaluation setting – limited data, wide altitude range, and mono-spectral imagery – constrains how far the results can be extended. These factors should be borne in mind when comparing our method to others or when applying it in new contexts.

Future work may explore more advanced preprocessing techniques such as spectral filtering or adaptive denoising. Lightweight attention mechanisms could be

evaluated to reduce computational overhead while retaining spatial focus. Incorporating multispectral data, combining thermal with visible or near-infrared channels, could offer richer input representations. Finally, real-time UAV deployment and flight testing would be valuable to assess the model's performance under operational constraints such as motion, variable illumination, and onboard hardware limitations.

4. CONCLUSIONS

This thesis addressed the challenge of detecting objects (such as humans and vehicles) in high-altitude UAV thermal imagery. This problem is significant for applications like search-and-rescue and surveillance. However, thermal sensors on UAVs at high altitudes produce images with limited detail and low contrast, which makes reliable detection challenging. To address these issues, a YOLOv8m-based detection model was developed and evaluated. This approach incorporated image preprocessing using Contrast Limited Adaptive Histogram Equalization (CLAHE) to amplify faint thermal features, and it compared the effects of integrating two attention modules (CBAM and CoordAttention) into the network. The main focus was to determine which enhancements yield the most reliable detection performance under the constraints of thermal UAV imagery.

The experimental results showed that the YOLOv8m model using only CLAHE preprocessing achieved the best detection performance. CLAHE amplified subtle thermal contours, making objects more distinguishable in the otherwise low-contrast images. In contrast, adding the CBAM or CoordAttention modules did not improve performance and in fact sometimes introduced errors. On the high-altitude thermal dataset, the CLAHE-enhanced model outperformed all attention-augmented variants. This suggests that for such low-detail infrared images, boosting input contrast yields more benefit than adding additional attention layers. In other words, simple contrast enhancement made the faint object signatures more detectable, whereas the extra complexity of channel/spatial attention did not compensate for the scarcity of strong features in these images.

These findings represent the core contributions of this work. By systematically evaluating the effects of CLAHE, CBAM, and CoordAttention within the same YOLOv8m framework, the study clarifies how each component contributes to detection performance. The results indicate that a simple preprocessing step (CLAHE) can have a greater impact than complex attention mechanisms in this domain. This insight suggests that, when working with sparse thermal data, model designers should prioritize image-level enhancements (like contrast boosting) over additional network complexity. In particular, the superior performance of the CLAHE-only model highlights the limited

benefit of channel/spatial attention modules for imagery with weak signals. Overall, this work provides empirical guidance for the development of lightweight, real-time detectors in low-contrast infrared settings. The enhanced YOLOv8m pipeline and its documented evaluation enrich the literature on thermal UAV detection and inform practical design choices for search-and-rescue and surveillance systems.

Future work should build on these results in several ways. More advanced preprocessing techniques—such as adaptive noise filtering or spectral enhancement—could be investigated to further improve object visibility. Alternative attention mechanisms or architectural optimizations might be evaluated to achieve better efficiency or accuracy. Combining thermal imagery with other modalities (e.g., visible or near-infrared channels) could provide richer information and improve detection under challenging conditions. Finally, deploying and testing the model in actual UAV flight scenarios would be valuable to assess its robustness under realistic operational conditions (motion, varying illumination, hardware constraints). These directions would extend the understanding of deep-learning-based object detection in high-altitude thermal imagery and help advance practical UAV imaging systems.

In summary, this thesis demonstrates that adaptive contrast enhancement is a critical factor for reliable object detection in high-altitude UAV thermal imagery. The findings clarify that enhancing input contrast via CLAHE can be more effective than adding attention modules under low-detail conditions. These insights will guide future efforts to design efficient and accurate detection systems for aerial thermal imaging, ensuring that resources are focused on the most impactful enhancements.

REFERENCES

[1] Buhl, N. (2023). Image thresholding in image processing. Encord. https://encord.com/blog/image-thresholding-image-processing/

[2] Google Colab. (n.d.-a). Colab.google. https://colab.google/

[3] Contrast Limited Adaptive Histogram Equalization – MATLAB & Simulink. (n.d.-b). https://la.mathworks.com/help/visionhdl/ug/contrast-adaptive-histogram-equalization.html

[4] Dash, Y., Gupta, V., Abraham, A., & Chandna, S. (2025). Improving object detection in high-altitude infrared thermal images using magnitude-based pruning and non-maximum suppression. *Journal of Imaging*, *11*(3), 69. https://doi.org/10.3390/jimaging11030069

[5] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021, July 18). YOLOX: Exceeding YOLO series in 2021. *arXiv*. https://arxiv.org/abs/2107.08430

[6] Hou, Q., Zhou, D., & Feng, J. (2021, March 4). Coordinate attention for efficient mobile network design. *arXiv*. https://arxiv.org/abs/2103.02907

[7] Kumar, N., & Singh, P. (2023, November 27). Small and dim target detection in IR imagery: A review. *arXiv*. https://arxiv.org/abs/2311.16346

[8] Li, H., Wang, S., Li, S., Wang, H., Wen, S., & Li, F. (2024). Thermal infrared-image-enhancement algorithm based on multi-scale guided filtering. *Fire*, *7*(6), 192. https://doi.org/10.3390/fire7060192

[9] Li, H., Li, Y., Xiao, L., Zhang, Y., Cao, L., & Wu, D. (2025). RLRD-YOLO: An improved YOLOv8 algorithm for small object detection from an unmanned aerial vehicle (UAV) perspective. *Drones*, *9*(4), 293. https://doi.org/10.3390/drones9040293

[10] Li, Y., Li, Q., Pan, J., Zhou, Y., Zhu, H., Wei, H., & Liu, C. (2024). SOD-YOLO: Small-object-detection algorithm based on improved YOLOv8 for UAV images. *Remote Sensing, 16*(16), 3057. https://doi.org/10.3390/rs16163057

[11] Lyu, M., Zhao, Y., Huang, C., & Huang, H. (2023). Unmanned aerial vehicles for search and rescue: A survey. *Remote Sensing, 15*(13), 3266. https://doi.org/10.3390/rs15133266

[12] Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788).

[13] Suo, J., Wang, T., Zhang, X., et al. (2023). HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle–based object detection. *Scientific Data, 10*, 227. https://doi.org/10.1038/s41597-023-02066-6

[14] Tahir, N. U. A., Long, Z., Zhang, Z., Asim, M., & ELAffendi, M. (2024). PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8. *Drones, 8*(3), 84. https://doi.org/10.3390/drones8030084

[15] Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, *5*(4), 1680–1716. https://doi.org/10.3390/make5040083

[16] Torres, J. (2025a, January 2). YOLOv8 architecture; Deep dive into its architecture – YOLOv8. YOLOv8. https://yolov8.org/yolov8-architecture/

[17] Torres, J. (2025b, January 3). YOLOv8 label format: Step-by-step guide | YOLOv8. YOLOv8. https://yolov8.org/yolov8-label-format/

[18] Ultralytics. (2025, March 31). YOLOv8. https://docs.ultralytics.com/models/yolov8/

[19] Wang, Y., Zhang, J., & Zhou, J. (2024). Urban traffic tiny object detection via attention and multi-scale feature driven in UAV-vision. *Scientific Reports, 14*, 20614. https://doi.org/10.1038/s41598-024-71074-2

[20] Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018, July 17). CBAM: Convolutional block attention module. *arXiv*. https://arxiv.org/abs/1807.06521

[21] Xu, L., Zhao, Y., Zhai, Y., et al. (2024). Small object detection in UAV images based on YOLOv8n. *International Journal of Computational Intelligence Systems, 17*, 223. https://doi.org/10.1007/s44196-024-00632-3

[22] Yaseen, M. (2024, September 12). What is YOLOv9: An in-depth exploration of the internal features of the next-generation object detector. *arXiv*. https://arxiv.org/abs/2409.07813

[23] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020, April). Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 12993–13000).