Jurnal Nasional Teknik Elektro dan Teknologi Informasi
Volume 15 Number 2 May 2026

DOI: 10.22146/jnteti.v15i2.24931

A Comparison of SR and CBAM for Optimized Thermal Drone Object Detection

Helfy Susilawati[1], Akhmad Fauzi Ikhsan[1], Firman[1], Arief Suryadi Satyawan[2], Chandra Rahmana[1]

[1] Department of Electrical Engineering, Faculty of Engineering, Universitas Garut, Garut, Jawa Barat 44116, Indonesia
[2] Research Center for Telecommunication, National Research and Innovation Agency, Bandung, Jawa Barat 40135, Indonesia

[Received: 27 October 2025, Revised: 1 February 2026, Accepted: 16 March 2026]
Corresponding Author: Helfy Susilawati (email: helfy.susilawati@uniga.ac.id)

ABSTRACT — Human detection using thermal cameras is very useful in certain conditions, such as detecting people lost in mountainous areas that are difficult to explore. Rescue operations are usually conducted by deploying a search and rescue (SAR) team to the location, which is not always effective because this operation can only be carried out under certain conditions and may pose a risk to the SAR team itself. Therefore, one alternative approach is the use of drones equipped with human detection and recognition capabilities. In this context, thermal cameras are used because they can penetrate challenging environments, making them suitable for SAR operations. The object detection method used in this study was You Only Look Once (YOLO) version 8 or YOLOv8. This study aimed to compare the effectiveness of integrating enhanced super-resolution generative adversarial networks (ESRGAN) with YOLOv8 and incorporating a convolutional block attention module (CBAM) into the neck architecture of YOLOv8. The performance of ESRGAN with YOLOv8 and CBAM with YOLOv8 was evaluated using precision, mean average precision (mAP), and training loss. Based on the experimental results, the combination of ESRGAN with YOLOv8 outperformed the CBAM-based modification. This is indicated by higher precision and mAP values, as well as lower training loss in the ESRGAN-enhanced YOLOv8 detection framework. The experimental findings highlight that image enhancement using ESRGAN is more effective than CBAM-based modification in improving thermal image-based human detection performance for SAR applications.

KEYWORDS — SAR, YOLOv8, Super Resolution, CBAM, Object Detection.

I. INTRODUCTION

Mountain climbing activities carry the risk of accidents or emergencies, such as someone getting lost or missing. One common example is the case of people going missing in the mountains. The latest reports are cases of climbers missing on Mount Binaya [1] and Mount Buthak [2]. During the search and rescue (SAR) process for missing hikers in mountainous areas, various challenges are encountered, such as difficult terrain, extreme weather, limited visibility due to fog or nighttime conditions, and a shortage of rescue personnel. This often leads to delays in the rescue process, as conducting the rescue could also pose a risk to the SAR team. To overcome this, an unmanned aerial vehicle (UAV) is needed to survey the affected area and help establish a vital wireless communication link between survivors and the nearest available cellular infrastructure [3]. AI-controlled robotic systems are emerging to address the challenges that exist for SAR operations designed to save lives in disaster zones where hazardous environments and challenging terrain make the rescue effort more difficult [4]. One technology that can be used to support SAR teams is drones equipped with thermal cameras as search tools. Thermal cameras are widely used in diverse applications, such as drones used for close-range geophysical surveys to detect surface and subsurface anomalies [5]; plant growth prediction and monitor by capturing plant surface temperature information correlated with water content [6]; detection, map, localization of peat fires using the main sensor [7]; and population estimation of the Formosan sika deer using drone-based monitoring [8].

Several studies have explored the use of drones and thermal cameras for SAR rescue processes. For instance, thermal cameras have been used as the main sensor to detect and track humans in drone-based SAR missions using YOLOv5, with the parameters measured being total track life (TTL), mean track life (MTL), and track purities [9]. In addition to the use of AI in drones, research has also been conducted on the addition of multi-rotor drones. Thermal cameras have proven very effective for finding missing persons under difficult conditions (darkness, fog/obstacles), with the track segment association (TSA) method significantly increasing tracking accuracy up to > 200% [10]. Other research has focused on improving human detection using thermal cameras and drones, with detection models based on YOLO8m and YOLO5s [11]. The parameters compared were the values of precision, recall, mean average precision (mAP)@50, and mAP@50–95. There is also research on human detection using thermal cameras mounted on drones using YOLOv3 [12].

Thermal cameras are utilized due to their advantages over ordinary visual cameras, particularly their ability to detect objects in the dark and measure temperature differences [13]. Many studies have employed thermal image datasets from drones collected under different conditions, such as open areas and urban environments. Although these datasets are not directly related to the mountainous context, they remain important as they provide aerial thermal images of humans that are relevant to SAR scenarios.

The use of drones for human detection in mountains can be very effective if the system can run to recognize objects well. However, the problem that arises from images captured by drones is that they have low resolution, which limits the model to extract important features. This research utilized the available drone thermal dataset, allowing researchers to develop, train, and test the dataset, which was then used to compare the use of preprocessing with super resolution with

modifications to the architecture using convolutional block attention module (CBAM).

Super resolution was chosen because it can improve precision on very small objects [14]. Using super resolution without modifying the architecture is a simple and computationally efficient method, but it can increase the detection rate [15]. Using super resolution with object detection to detect small objects, especially vehicles in aerial and traffic images, increases significantly [16] and improves the visual quality of objects from low-resolution videos [17]. Super resolution is a technique that can be used to enhance pixels not only by enlarging their size but also by adding new details learned from training data (e.g., texture, object edges, fine patterns). The super resolution technique used in this study was enhanced super-resolution generative adversarial networks (ESRGAN), which was then integrated into YOLOv8.

CBAM is an attention mechanism added to the convolutional neural network (CNN) to help the model "focus" on the more important feature parts of an image. CBAM works by providing attention to two aspects: channel attention and spatial attention. CBAM was chosen because it integrates channel and spatial information to improve the model's focus on feature context and object location [18]. Additionally, CBAM is used to improve feature discrimination through channel and spatial attention [19]. CBAM is widely used in thermal cameras, including for detecting leaks on dam surfaces [20], and in the backbone to improve smoke feature representation [21]. This research applied CBAM into the backbone using YOLOv8 so that features entering the neck and head became cleaner and more informative, allowing the head to focus on detecting important features instead of cleaning up noises. After the CBAM process, spatial pyramid pooling–fast (SPPF) was also performed to enrich multi-scale information before it was forwarded to the head (neck + detection).

Novel strategic comparative perspectives and controlled empirical evidence that provide new insights into the relative effectiveness of using two approaches in UAV scenarios is presented in this paper. The first approach was a combination of ESRGAN with YOLOv8, which was expected to produce higher object detection parameters using the super resolution method. The second approach was the application of CBAM on the YOLOv8 backbone to improve the quality of attention to object detection without changing the input data. Unlike previous studies that generally evaluated the two approaches above separately, this study directly explored how different levels of intervention affected the performance of small object detection in low-resolution drone imagery. From the perspective of scientific contribution, this study distinguished two different performance improvement paradigms in YOLOv8-based UAV object detection, namely data-level image enhancement through super-resolution and model-level feature refinement through attention mechanism, where the dataset used the public MONET dataset. Meanwhile, the practical impact of this study lies in the potential application of ESRGAN-based detection models using YOLOv8 for SAR operations in mountainous areas. With enhanced human detection capabilities using thermal cameras, this system can accelerate the victim identification process, improve the safety of search and rescue teams, and reduce search time.

II. RELATED WORKS

A. THERMAL IMAGING FOR SEARCH AND RESCUE

The ability to detect infrared emissions from the human body has led to the widespread use of thermal cameras in SAR operations, where thermal cameras can be used even in low light, fog, or dense forest areas. This is in contrast to red, green, blue (RGB) cameras, which are highly dependent on lighting. Thermal cameras can provide high contrast between living objects and their surroundings. Several studies using thermal cameras for SAR operations have compared the performance of YOLOv7-Tiny with YOLOv8x to determine the most suitable model for SAR. The results showed an average precision (AP) of 95% for YOLOv8x, while the smallest variant, YOLOv8n, still maintained accuracy with a value of 91%. Meanwhile, YOLOv7-Tiny provided performance equivalent to YOLOv8n, but was lighter (48% smaller) [22]. Another study involved the use of a thermal image dataset taken directly with a drone that functions as a victim detection system in post-disaster search and rescue operations, which was trained using YOLOv3 and tested on NVIDIA Jetson TX2 with real-time results of 26.6 FPS [23], and the use of RTMDet, PP-YOLOE+s, YOLOv5s, YOLOv8s, and DINO for SAR, where a new UAV-based dataset with an infrared thermal camera called partially occluded person (POP) was previously collected, resulting in stable average precision and high accuracy up to an occlusion level of 70% [24].

B. OBJECT DETECTION IN THERMAL IMAGES

Research has been conducted on object detection in thermal images, including research on the use of deblur-SRRGAN + mask region-based CNN (R-CNN), which has succeeded in improving thermal image quality and the accuracy of object/thermal reflection detection compared to other methods, although it still has limitations in processing speed for real-time applications [25]. In addition, other research has employed multi-sensors integrating thermal + light detection and ranging (LiDAR) for automatic perception systems in maritime search. Using YOLOv8, the proposed approach produced precision of 93.5% and recall of 94.2% [26].

C. ATTENTION MECHANISMS IN DEEP LEARNING

An attention mechanism can be defined as a process that maps a query vector and a collection of keys–value vector pairs into a single output vector [27]. Research on the use of the channel-wise attention mechanism combined with GAN and residual network (ResNet), emphasizing important features in the channel dimension to focus on critical facial characteristics [28] is one of the studies related to attention mechanisms, where the results of the study obtained precision of 0.79, recall of 0.88, F1 score of 0.83, accuracy of 0.83, and receiver operating characteristic-area under the curve (ROC-AUC) of 0.825. Another study investigated the use of an efficient-vision transformer (ViT) combined with an attention mechanism for early detection of knee osteoarthritis. By comparing precision, recall, F1-score, and mAP, the result showed that YOLOv8-ViT outperformed YOLOv5n, YOLOv8n, and YOLOv9n [29].

D. YOLOV8 FOR OBJECT DETECTION

Research on the use of YOLOv8 has been extensively conducted, including its application for detecting small objects [30], [31], traffic signs [32], and potential landslides [33]. The four studies mentioned above demonstrated that the use of YOLOv8 for object detection improved the accuracy.

E. SUPER-RESOLUTION FOR SMALL OBJECT DETECTION

Super-resolution research was used, among other things, to improve the resolution and texture of small pineapple objects,

where image capture using drone results in a 17% increase in precision [34]. Another study is the use of the CSRGAN model with classification-oriented super-resolution, which significantly improves the quality of small object detection in UAV images [35]. From a perceptual perspective, one study demonstrated that SRGAN reconstructions with large upscaling factors were four times more photorealistic [36].

III. METHODOLOGY

This study used two approaches to improve the detection of objects from thermal images taken using drones. These two approaches will then be compared to determine which approach has the highest parameter value. The first approach is to apply Super Resolution in the pre-processing stage. The second approach is to modify the model using the convolutional block attention module (CBAM) and store it in the backbone. The dataset used in this study is the MONET dataset [37] with 7,610 thermal images for the training process, 1,676 thermal images for the validation process, and 3,443 thermal images for testing.

The super-resolution approach was used because the images produced by drones are relatively small. In the object detection process, small images present their own level of difficulty. Therefore, a method is needed to enlarge the image size without losing the details of the image, one of which is by using super-resolution. The super-resolution technique used in this study was ESRGAN, which is an improvement over super-resolution generative adversarial networks (SRGAN) but still follows the same workflow as SRGAN. The workflow of ESRGAN can be seen in Figure 1.

Figure 1 shows that the SRGAN process uses a generator and a discriminator. Super-resolution refers to the process of generating high resolution (HR) images from low resolution (LR) images. In SRGAN, LR images are the input for the generator. The discriminator compares the generator's output with the HR image, and the loss is calculated to adjust the weights and biases. The generator network continuously attempts to make the LR image closer to the actual HR image. Conversely, the discriminator network generates a discrimination probability that can be used to assess accuracy by trying to distinguish between the generated photo and the actual HR image.

Although the SRGAN and ESRGAN flows are relatively similar because they are part of GAN and super resolution (SR), the difference between SRGAN and ESRGAN is that ESRGAN uses residual-in-residual dense blocks (RRDB) in the generator section and does not use batch normalization in RRDB. By doing so, it can improve perceptual quality, which can provide better visual quality consistently with more realistic and natural textures compared to SRGAN [38].

In addition to using ESRGAN, this study also used CBAM as a way to determine the performance of object detection in thermal camera images. CBAM is an attention block that applies two sequential stages, namely channel attention followed by spatial attention. Channel attention serves to calculate the weight of each feature channel that focuses on "what" important information is, and spatial attention calculates a map of important locations in the feature that focuses on "where" the crucial information is.

In this study, CBAM was placed in the backbone, at the deepest feature map level after the C2f block, using 1024 channels. CBAM was then inserted into the YOLOv8 architecture. The CBAM architecture in YOLOv8 can be seen in Figure 2. Based on this figure, the YOLOv8 algorithm consists of four parts: input, backbone, neck, and output. The C2f module in YOLOv8 has fewer parameters and superior feature extraction capabilities. This approach maintains lightweight characteristics while capturing more abundant gradient flow information, which significantly improves model performance.

This study aimed to compare the results of using ESRGAN and CBAM on the YOLOv8 architecture on thermal images. The research started from detecting objects using the original thermal image with YOLOv8. The next step was to perform SR using ESRGAN. The data used in the SR process were the train image data. The valid and test image data were not subjected to SR so that the system could recognize the original images. After SR was performed, object detection was then carried out using YOLOv8. The objects detected were humans, cars, and animals. After detection by YOLOv8, the next step was to calculate the training loss, validation loss, precision, recall, and mean average precision (mAP) parameters.

After the SR process was complete, the next step was to modify the model in the backbone and neck sections using CBAM. The modified model was then applied to YOLOv8 for object detection. As with SR, the next step was to calculate the training loss, validation loss, precision, recall, and mAP parameter values.

IV. RESULTS AND DISCUSSION

The loss data analyzed in this study were train loss and validation loss. Train loss consisted of box loss, class loss, and distribution focal loss (DFL). Validation loss consisted of val box loss, val class loss, and val DFL. Box loss aimed to measure how well the bounding box prediction (x, y, w, h) compared to the ground truth. Class loss aimed to measure whether the model correctly guessed the object class. DFL is a special part of YOLOv7/YOLOv8 to improve the prediction of bounding box coordinates. Box loss focused more on the location and size of the bounding box, or what can be called the quality of the bounding box. Class loss focused more on the prediction of object classes, or what can also be called label quality. DFL focused more on the precision of the bounding box coordinates (distribution), or what can be called box accuracy. In the object detection process, this study used the default YOLOv8, namely stochastic gradient descent (SGD), with Initial LR = 0.001, final LR ratio = 0.1, scheduler = cosine decay, warmup = 5, batch size = 4, and epoch 50 using original data, SR, and CBAM.

A. OBJECT DETECTION USING ORIGINAL DATA WITH YOLOv8

The first study conducted was using the MONET dataset to determine the training loss, validation loss, precision, recall, and mAP values using YOLOv8. Figure 3 depicts the results of the parameters obtained using YOLOv8. Based on the research results, the accuracy (precision vs recall) value obtained was quite high (close to 0.869), meaning that the model was able to reduce false positives. The recall value was lower (max 0.626), meaning that the model still missed some objects (relatively high false negatives). The mAP50 value was quite good (0.717), but the mAP50-95 was lower (0.319), meaning that the model had difficulty maintaining performance across various intersections over union (IoU) scales, which can also be interpreted as inconsistent bounding box accuracy. The box loss and DFL values remained relatively stable. In contrast, the validation class loss value was quite high, indicating that the

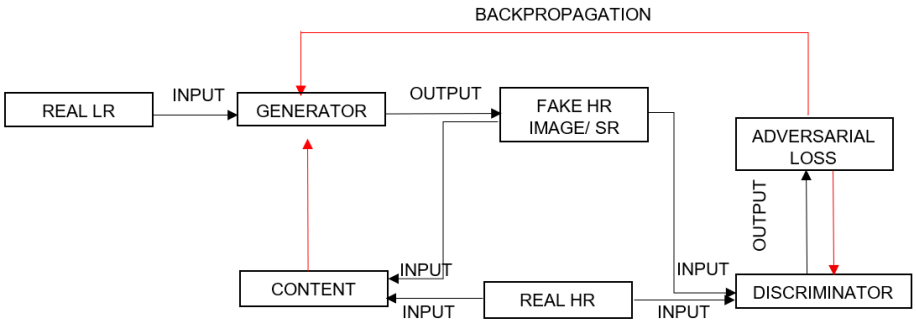

Figure 1. SRGAN flow.

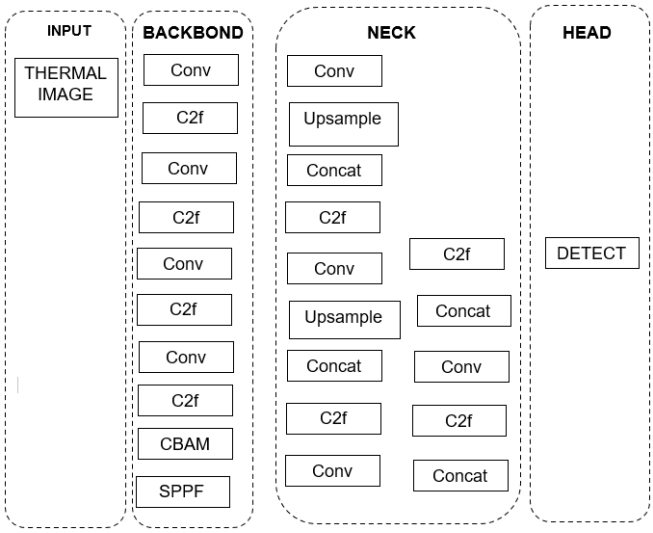

Figure 2. CBAM-YOLOv8 architecture.

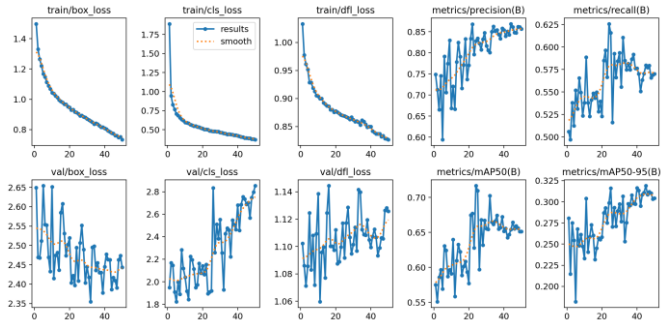

Figure 3. Results of parameter values using YOLOv8.

model still struggled to distinguish between classes. The learning rate reached a maximum value of 0.0012 and then gradually decreased to maintain stable training.

B. OBJECT DETECTION USING ESRGAN

The generator network architecture used in ESRGAN is residual-in-residual dense block network (RRDBNet). In this study, the input and output images consisted of three channels (red green blue, RGB), with an initial number of feature maps (channel width) of 64. The network architecture showed that there are 23 consecutive RRDB blocks in the network, and each dense block in the RRDB added 32 new feature channels. This study used ESRGAN for the super-resolution process with a scale of 4, making the ESRGAN image to have a pixel value 4 times that of the original image. The dataset used consisted of images with a resolution of 800 × 600, and the results from ESRGAN produced images with a size of 3200 × 2400. Figure 4 and Figure 5 show a comparison between the original image and the image produced by ESRGAN.

Figure 5 shows that the patterns in the ESRGAN image appear much sharper compared to the original image in Figure 4. For example, the lines of the rice field patterns appear smoother and more defined compared to the results of simple interpolation. In addition, object boundaries such as field edges and vehicles (white dots on the road) are more distinct. Based on Figure 5, it is also evident that in agricultural areas, ESRGAN clarifies the planting pattern, as seen from the more continuous vertical and horizontal lines. These fine details are close to photo-realistic, while in Figure 5 they appear blurrier.

Figure 6 depicts the results of the parameters obtained using SR training data with YOLOv8. Based on the research results, the accuracy (precision vs recall) and mAP values for images that had undergone SR were almost the same as those for the original images and the validation loss value was significantly higher, indicating potential overfitting. The learning rate was unstable, as seen from the train loss increasing when the learning rate got bigger, whereas it should decrease or at least remain the same when the learning rate is stable and appropriate.

C. OBJECT DETECTION USING CBAM

Training using YOLOv8 with CBAM modifications was performed on the original dataset. CBAM modifications were made to the backbone, which was then inserted into YOLOv8. This study used the channel attention and spatial attention modules. Channel attention focuses on “what” is important in each feature map channel, while spatial attention focuses on “where” the important features are located after the channel attention module.

The steps in channel attention include global pooling, multi-layer perceptron (MLP), and combination and sigmoid. In global pooling, the input feature x with dimensions $[B,C,H,W]$ was processed with global average pooling and global max pooling for each channel. This process was carried out through AdaptiveAvgPool2d and AdaptiveMaxPool2d. The result was two channel descriptor vectors with dimensions $[B,C,1,1]$. In MLP (bottleneck), each channel descriptor was passed to a multi-layer perceptron network implemented with two 1×1 convolutions: fc1 and fc2 with an intermediate ReLU layer. This process reduced the channel feature dimensions to a smaller size corresponding to the bottleneck ratio and then returned them to their original size. In combination and sigmoid, the outputs from the average and max pooling paths were summed and a sigmoid function was applied, forming a channel attention map Mc(F) of size $[B,C,1,1]$ with values in the range [0,1]. Mathematically, this combination corresponds to (1) [39]:

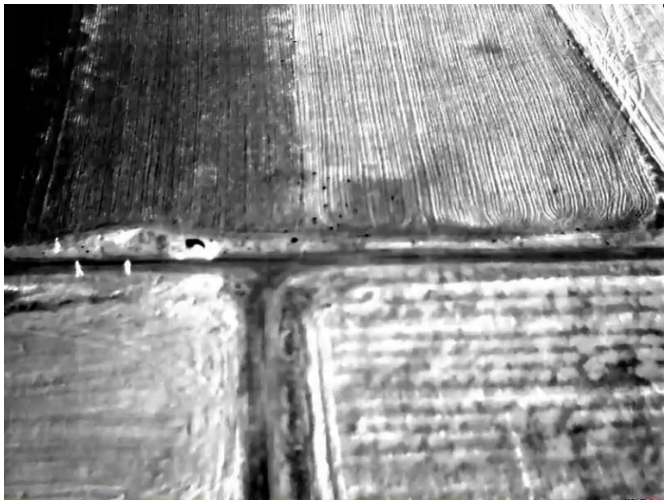

Figure 4. Original image from the dataset.

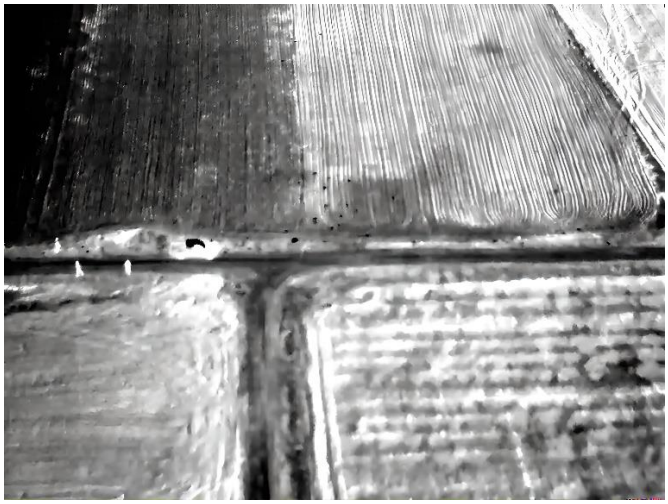

Figure 5. ESRGAN result image.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)) \quad (1)$$

The resulting channel attention map was then multiplied element-wise with the initial feature map x. Thus, feature channels considered most informative were given higher weights, while less important ones were suppressed. By performing these four steps, the channel attention module generated a per-channel attention map that strengthened relevant feature channels. The next step was to perform spatial attention, with the following steps: channel pooling, channel combination and convolution, and sigmoid and combination.

Based on the input to spatial attention, the next step was to calculate the average and maximum values along the channel dimension. This process produced two 2D maps of size $[B,1,H,W]$ that summarize the spatial information (one map of the channel average and one map of the channel max). Next, in channel combination and convolution, the two resulting 2D maps were then combined along the channel dimension into a tensor $[B,2,H,W]$. This two-channel tensor was then processed by a single 2D convolution layer of size 7 × 7, which produced a 1-channel map of size $[B,1,H,W]$. The convolution process served to identify spatial patterns in the combined channel features. In sigmoid and combination steps, the convolution output was then passed through a sigmoid function to obtain a spatial attention map Ms (F) of size $[B,1,H,W]$ with values [0,1]. The resulting map showed which pixel locations were

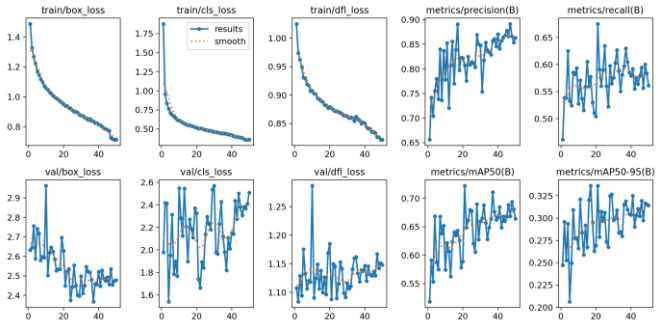

Figure 6. Results of parameter values using SR training data with YOLOv8.

important. The spatial attention map was then applied by multiplying (element-wise) the initial feature map with the mask, thereby strengthening the informative spatial areas. This process was consistent with the spatial attention formulation, namely [39].

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \quad (2)$$

The value $f^{7\times7}$ is a 7 × 7 convolution operation on the combined map. Thus, spatial attention highlighted the locations where the most important information was located in the feature map. Based on the attention channel and spatial attention, they were then combined using (3):

$$F' = M_c(F) \otimes F, F'' = M_S(F') \otimes F' \quad (3)$$

Figure 7 illustrates the results of the parameters obtained using CBAM. Based on the research results, the accuracy (precision vs recall) value obtained showed the highest precision value of 0.81, indicating the model demonstrated quite good at distinguishing valid objects from false positives. The recall yielded an average value of 0.53, meaning that the model had not captured all small/fine objects. The mAP value yielded an average of 0.58, which is quite good for a model with CBAM. The mAP50-95 value had an average of 0.25, showing a significant gap with mAP50. This indicates that the model experienced difficulty with IoU. The val box loss value remained higher than the training value, but it was stable. This indicates that there was no severe overfitting. The val class loss value fluctuated slightly but showed a decrease compared to the initial stage. Meanwhile, the val DFL value was very stable, and based on the results of the study, a high learning rate at the beginning of training proved to be effective in significantly reducing loss quickly in the early stages. After that, a gradual decrease in the learning rate allowed the model to reach peak performance, causing the precision, recall, and mAP values to reach their highest values in the middle of training. At the end of training, a very low learning rate made the model stable, with loss approaching a constant and performance metrics relatively remaining stable without major fluctuations. A comparison of the main matrix summaries for the three models can be seen in Table I.

Table I shows that the YOLOv8 + SR model achieved the best overall performance. The SR model had the highest precision value (≈86.3%) and the highest mAP@0.5 (≈66.4%). This shows that YOLOv8 + SR is slightly superior to the original model (precision 85.7%, mAP@0.5 65.2%). Meanwhile, the CBAM model had the lowest performance with a precision value of ≈77.7% and a mAP@0.5 of ~61.9%. In terms of recall, the original model was slightly higher (57.0%) than SR (56.1%), but the difference was very small. In terms of

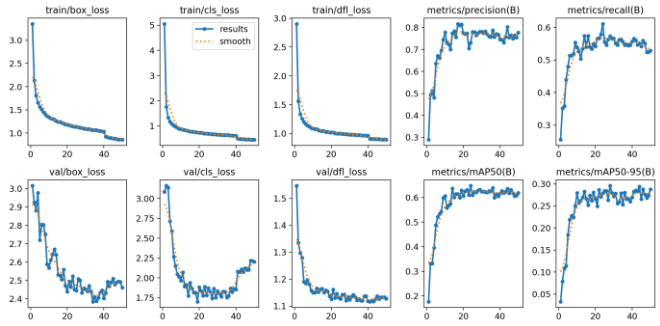

Figure 7. Results of parameter values using CBAM.

TABLE I
COMPARISON OF MAIN MATRIX SUMMARIES FOR THE THREE MODELS

No	Model	Precision	Recall	mAP @0.5	mAP@0.5:0.95
1	Original YOLOv8	85.7	57.0	65.2	30.4
2	YOLOv8 + CBAM	77.7	52.9	61.9	28.8
3	YOLOv8 + ESRGAN	86.3	56.1	66.4	31.4

mAP@0.5:0.95, which reflects the quality of detection more strictly, the ESRGAN model obtained the highest score of 31.4%, followed by the original model with 30.4%, and then CBAM with the lowest score of 28.8%. The differences in metric values from several parameters indicate that the addition of the SR module provides a slight increase in detection accuracy compared to the standard YOLOv8, while the addition of the CBAM module to YOLOv8 results in performance below the original model for thermal object detection data.

When analyzed in terms of loss convergence during training, each model exhibited different characteristics. At the beginning of training (early epochs), the CBAM model experienced a much higher training loss than the other two models, in contrast to the original and SR models, which started with a lower initial loss. The high initial loss value of CBAM is possibly due to the complexity of the CBAM attention module, which makes optimization more difficult at the beginning, or because the initial weights are not as optimal as other models.

The fastest loss reduction occurred in the CBAM model due to its very high starting point. In the first 5 epochs, the CBAM training loss dropped by more than 60%, while the original and SR models dropped by around 36%. CBAM reduced the loss the fastest at the beginning, but its starting point was much worse. After 10 epochs, the CBAM training loss was still slightly above the original and ESRGAN models. Furthermore, all models showed a trend of slower loss reduction at the end of training, indicating that the three models have reached a stable point. The training loss value of the ESRGAN model was consistently slightly lower than the original model throughout the middle to end of training. This shows that the ESRGAN model learns slightly more effectively than the original model. Meanwhile, the CBAM model has the highest loss value until the end, although the difference is not significant.

In terms of validation loss, the CBAM model was slightly lower than the original or SR models, but this does not reflect better detection performance. In fact, the mAP value for CBAM was lower. This may indicate that the loss metric is not always directly proportional to mAP, especially if the model tends to make predictions with varying degrees of confidence.

Figure 8 shows that at epoch 1, the original model and the ESRGAN model with YOLOv8 achieved high initial performance on mAP@0.5, namely 57.5% and 51.8%. This result shows that with the pretrained weights used; both models were able to detect thermal objects with good accuracy from the start of training. In contrast, the CBAM model with YOLOv8 started from a much lower performance point, at only 17.7% at epoch 1. This result shows that the CBAM model initially had difficulty performing detection, possibly because the addition of the CBAM module changed the feature distribution, requiring greater adjustment than the initial weights. Nevertheless, the CBAM model showed rapid improvement in the first few epochs, with the mAP@0.5 value jumping to 48.8% at epoch 5, which is close to the achievement of the original model or ESRGAN in the same epoch. In other words, CBAM caught up in the first 5 to 10 epochs. In fact, by epoch 10, CBAM's performance had matched and slightly surpassed that of the original model and ESRGAN. This result indicates that, in terms of learning speed, after overcoming the initial difficult phase, CBAM was able to learn quickly.

Based on the comparison of the research results, a lower training loss value does not always correspond to an increase in detection accuracy. For example, the model modified with CBAM achieved relatively stable validation loss values but produced a lower mAP (0.58) compared to the model using ESRGAN (0.72). This result indicates that the presence of an attention module helps in the feature learning process; however, it does not necessarily improve the discriminative ability of the model toward thermal image data. Conversely, the model using ESRGAN showed a lower training loss and higher precision, suggesting that improving image resolution quality directly contributes to better feature representation.

To see the speed and stability of the training process, not just the final performance, a convergence table was needed between the super-resolution and attention mechanism-based approaches. Table II shows that the model learnt the object faster than the precise location, as seen from the faster convergence of mAP50 compared to mAP50–95. Furthermore, super-resolution facilitated earlier precision learning, while the attention mechanism required an adaptation phase to effectively weigh features.

In terms of model efficiency, the use of ESRGAN increased the preprocessing stage but did not affect the inference complexity of YOLOv8 because ESRGAN only modified the input resolution. Meanwhile, integrating CBAM within the backbone increased the number of parameters and computation per layer, which may explain its slower convergence and lower overall accuracy. This result shows that although attention mechanisms are powerful for visual tasks, their direct application to thermal-based YOLOv8 can lead to excessive focus on low-texture regions, thereby reducing detection effectiveness. The precision value across all models consistently exceeded the recall value, indicating that the detectors tend to be more conservative in identifying objects, meaning they prioritize accuracy over completeness.

When the three models were analyzed side by side, YOLOv8 combined with ESRGAN demonstrated the best

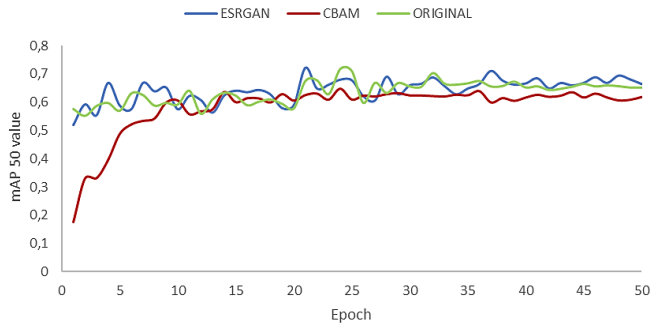

Figure 8. Comparison of mAP@0.5 (mAP 50) on Original, CBAM, and SR (ESRGAN) images.

TABLE II
CONVERGENCE TABLE

No	Model	Matrix	Final mAP	90% Final	Epoch at 90%
1	ESRGAN +YOLOv 8	mAP50	0.66394	0.314 32	4
2	ESRGAN +YOLOv 8	mAP50 - 95	0.59754 6	0.282 888	2
3	CBAM+ YOLOv8	mAP50	0.61902	0.287 53	9
4	CBAM+ YOLOv8	mAP50 - 95	0.55711 8	0.258 777	10

balance among accuracy metrics. The relatively small improvement compared to the original YOLOv8 indicates that super-resolution contributes most effectively when handling small or low-texture objects. The research results also revealed that the modification using CBAM appeared to be less suitable for the characteristics of thermal images, which typically exhibit weak spatial gradients. Based on these findings, it can be concluded that image resolution enhancement is more beneficial than network architecture modification in the context of thermal object detection.

D. ERROR ANALYSIS

The research results show that the YOLOv8 model combined with ESRGAN achieved the highest overall accuracy, followed by the standard YOLOv8 and the CBAM modification, which is consistent with previous studies [14], [16], [17], [40]. Based on earlier research, CBAM focused on specific regions, thereby potentially improving object detection performance [18], [19]. However, the findings of this study indicate that the CBAM modification did not significantly improve object detection parameters compared to the original model and the model combined with ESRGAN. This discrepancy may be attributed to differences in the type of objects studied. In previous research, CBAM was not applied to UAV imagery, whereas this study utilized thermal images captured by UAVs [18]–[20]. This result suggests that attention mechanisms cannot fully replace the need for improving the quality of the initial input image.

Although the use of ESRGAN, which enlarges objects, has the potential to introduce synthetic texture artifacts, in the context of human detection using thermal imagery from UAV cameras with small target objects, applying ESRGAN prior to the detection process proves to be a more effective strategy than modifying the detection network architecture. In the context of real-world implications, the findings of this study have direct relevance to the implementation of drone-based SAR systems. Based on the obtained results, the ESRGAN-based YOLOv8 model can detect human presence more reliably in low-light environments or mountainous areas, thereby significantly accelerating the victim identification process.

In the model that enhances image resolution using ESRGAN, although an additional preprocessing stage is introduced, the training process remains stable with smooth convergence at each epoch. This stability indicates that improving the input image resolution helps the model learn feature representations more efficiently while also reducing fluctuations during the backpropagation process. In contrast, the model modified with CBAM requires longer training time and greater computational resources due to the addition of convolutional layers in the backbone. The increased number of parameters introduced by the attention mechanism causes the learning process to become slower, particularly during the early stages of training.

In real-world applications, maintaining a balance between accuracy and computational load is crucial. Although ESRGAN improves feature clarity and detection reliability, it adds computational overhead at the preprocessing stage. However, since preprocessing can be performed prior to the inference stage, ESRGAN remains a practical choice for drone-based thermal detection, especially when using edge hardware that supports parallel processing. This makes the ESRGAN–YOLOv8 model suitable for onboard implementation using embedded GPUs such as NVIDIA Jetson or other similar edge computing devices.

V. CONCLUSION

Based on the research results, it can be concluded that the ESRGAN model is the most superior model in terms of detection accuracy compared to the original and CBAM models. The ESRGAN model can detect thermal objects with slightly greater accuracy than the original version with the value of precision and mAP@0.5, and mAP@0.5:0.95 values, where the model using ESRGAN achieved the highest scores of 86.3, 66.4, and 31.4. The original YOLOv8 model itself has proven to be very strong (close to ESRGAN), where the original YOLO model achieved the highest recall value, which was 57.0, so it can be an option if training time efficiency is considered. Further research can be carried out by modifying the SR module, and the testing section can be enriched by using the ablation method or by using 5-fold cross-validation.

CONFLICTS OF INTEREST

The authors declare that in the process of conducting and preparing this research, there were no conflicts of interest, whether financial, personal, or professional, that could affect the objectivity and integrity of the research results. The entire research process was conducted independently without any influence from any party with an interest in the results of this research. The results of this research are presented objectively.

AUTHORS' CONTRIBUTIONS

Conceptualization, Helfy Susilawati and Akhmad Fauzi Ikhsan; methodology, Helfy Susilawati; validation, Arief Suryadi Satyawan; analysis, Helfy Susilawati and Firman; investigation, Chandra; resources, Helfy Susilawati; data curation, Helfy Susilawati; writing—original draft preparation, Helfy Susilawati; writing—reviewing and editing, Helfy Susilawati; visualization, Akhmad Fauzi Ikhsan.

REFERENCES

[1] Krisdianto, "Pencarian pendaki yang hilang di Gunung Binaya dilanjutkan." Kementerian Kehutanan RI. Access date: 29-Sep-2025. [Online]. Available: https://www.kehutanan.go.id/news/pencarian-pendaki-yang-hilang-di-gunung-binaya-dilanjutkan

[2] M.B. Ibrahim, "Pemuda Sidoarjo hilang saat mendaki Gunung Buthak." Detikjatim. Access date: 29-Sep-2025. [Online]. Available: https://www.detik.com/jatim/berita/d-8068473/pemuda-sidoarjo-hilang-saat-mendaki-gunung-buthak

[3] M. Erdelj, E. Natalizio, K.R. Chowdhury, and I.F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervasive Comput.*, vol. 16, no. 1, pp. 24–32, Jan.-Mar. 2017, doi: 10.1109/MPRV.2017.11.

[4] Meenakshi *et al.*, "AI-driven autonomous robots for search and rescue operations in disaster zones," in *2025 Int. Conf. Data Sci. Agents Artif. Intell. (ICDSAAI)*, 2025, pp. 1–6, doi: 10.1109/ICDSAAI65575.2025.11011570.

[5] F. Mercogliano *et al.*, "ITINERIS geophysical technologies @CNR-IREA: Drone-based tests at Altopiano di Verteglia, Avellino (Southern Italy)," *Comput. Sci. Appl. – ICCSA 2025 Workshops*, 2025, pp. 373–383, doi: 10.1007/978-3-031-97663-6_33.

[6] H. Shi *et al.*, "Improving surface soil moisture content estimation of winter oilseed rape by integrating multimodal remote sensing information and structural characteristics," *Comput. Electron. Agric.*, vol. 237, pp. 1–18, Oct. 2025, doi: 10.1016/j.compag.2025.110779.

[7] T. Sam-Odusina *et al.*, "Detection and geolocation of peat fires using thermal infrared cameras on drones," *Drones*, vol. 9, no. 7, pp. 1–19, Jul. 2025, doi: 10.3390/drones9070459.

[8] B. Chang *et al.*, "Enhancing wildlife detection using thermal imaging drones: Designing the flight path," *Drones*, vol. 9, no. 1, pp. 1–14, Jan. 2025, doi: 10.3390/drones9010052.

[9] S. Yeom, "Thermal image tracking for search and rescue missions with a drone," *Drones*, vol. 8, no. 2, pp. 1–15, Feb. 2024, doi: 10.3390/drones8020053.

[10] S. Yeom, "Multi-rotor drone-based thermal target tracking with track segment association for search and rescue missions," *Drones*, vol. 8, no. 11, pp. 1–18, Nov. 2024, doi: 10.3390/drones8110689.

[11] H. Farman *et al.*, "Enhancing UAV-based human detection in thermal imaging with custom dataset," in *Proc. ICSDI 2024 Vol. 3*, 2024, pp. 437–444, doi: 10.1007/978-981-97-8345-8_53.

[12] J. McGee, S.J. Mathew, and F. Gonzalez, "Unmanned aerial vehicle and artificial intelligence for thermal target detection in search and rescue applications," in *2020 Int. Conf. Unmanned Aircr. Syst. (ICUAS),* 2020, pp. 883–891, doi: 10.1109/ICUAS48674.2020.9213849.

[13] V. Teju and D. Bhavana, "An efficient object tracking in thermal imaging using optimal Kalman filter," *Int. J. Eng. Trends Technol.*, vol. 69, no. 12, pp. 197–202, Dec. 2021, doi: 10.14445/22315381/IJETT-V69I12P223.

[14] L. Courtrai, M.T. Pham, and S. Lefèvre, "Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks," *Remote Sens.*, vol. 12, no. 19, pp. 1–19, Sep. 2020, doi: 10.3390/rs12193152.

[15] I. García-Aguilar, J. García-González, R.M. Luque-Baena, and E. López-Rubio, "Object detection in traffic videos: An optimized approach using super-resolution and maximal clique algorithm," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 18999–19013, Sep. 2023, doi: 10.1007/s00521-023-08741-4.

[16] Y.R. Musunuri, O.-S. Kwon, and S.-Y. Kung, "SRODNet: Object detection network based on super resolution for autonomous vehicles," *Remote Sens.*, vol. 14, no. 24, pp. 1–19, Dec. 2022, doi: 10.3390/rs14246270.

[17] S. Ren *et al.*, "Towards efficient video detection object super-resolution with deep fusion network for public safety," *Secur. Commun. Networks*, vol. 2021, pp. 1–14, May 2021, doi: 10.1155/2021/9999398.

[18] Y. Zhang *et al.*, "Research of maritime object detection method in foggy environment based on improved model SRC-YOLO," *Sensors*, vol. 22, no. 20, pp. 1–14, Oct. 2022, doi: 10.3390/s22207786.

[19] W. Wang *et al.*, "Improved detection and reidentification algorithm for natural gas pipeline excavator dynamic tracking based on deep learning," *Measurement*, vol. 257, pp. 1–17, Jan. 2026, doi: 10.1016/j.measurement.2025.118567.

[20] Z. Lv, S. Zhu, D. Wang, and Z. Liang, "Infrared-visible person re-identification via dual-channel attention mechanism," *Multimed. Tools Appl.*, vol. 82, no. 15, pp. 22631–22649, Jun. 2023, doi: 10.1007/s11042-023-14486-y.

[21] C. Song *et al.*, "Smoke video detection based on double spectrum," *Int. J. Robot. Autom.*, vol. 39, no. 2, pp. 161–169, Jan. 2024, doi: 10.2316/J.2024.206-1020.

[22] M. Rizk and I. Bayad, "Bringing intelligence to SAR missions: A comprehensive dataset and evaluation of YOLO for human detection in TIR images," *IEEE Access*, vol. 13, pp. 17208–17235, Jan. 2025, doi: 10.1109/ACCESS.2025.3529484.

[23] J. Dong, K. Ota, and M. Dong, "UAV-based real-time survivor detection system in post-disaster search and rescue operations," *IEEE J. Miniaturization Air Space Syst.*, vol. 2, no. 4, pp. 209–219, Dec. 2021, doi: 10.1109/JMASS.2021.3083659.

[24] Z. Song *et al.*, "An infrared dataset for partially occluded person detection in complex environment for search and rescue," *Sci. Data*, vol. 12, pp. 1–14, Feb. 2025, doi: 10.1038/s41597-025-04600-0.

[25] G. Batchuluun *et al.*, "Deep learning-based thermal image reconstruction and object detection," *IEEE Access*, vol. 9, pp. 5951–5971, Dec. 2020, doi: 10.1109/ACCESS.2020.3048437.

[26] F. Ponzini, D. van Hamme, and M. Martelli, "Human detection in marine disaster search and rescue scenario: A multi-modal early fusion approach," *Ocean Eng.*, vol. 340, pp. 1–15, Nov. 2025, doi: 10.1016/j.oceaneng.2025.122341.

[27] A. Vaswani *et al.*, "Attention is all you need," 2023, *arXiv:1706.03762*.

[28] P. Shetty, D.K.R, and D.S. Padre, "Deepfake detection using hybrid deep learning: Enhancing accuracy with ResNet, GANs, and attention mechanisms," in *2025 3rd Int. Conf. Augment. Intell. Sustain. Syst. (ICAISS)*, 2025, pp. 95–99, doi: 10.1109/ICAISS61471.2025.11041976.

[29] F. Ying *et al.*, "Improvements in automatic diagnosis methods for knee osteoarthritis based on deep learning," *Chinese J. Tissue Eng. Res.*, vol. 29, no. 35, pp. 7511–7518, 2025, doi: 10.12307/2026.533.

[30] J. Wang, J. Gao, and B. Zhang, "A small object detection model in aerial images based on CPDD-YOLOv8," *Sci. Rep.*, vol. 15, pp. 1–16, Jan. 2025, doi: 10.1038/s41598-024-84938-4.

[31] Y. Gao and B. Lin, "Research and design of a small object detection and tracking system based on YOLOv8 and ByteTrack algorithms," in *2024 IEEE 4th Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*, 2024, pp. 109–113, doi: 10.1109/ICDSCA63855.2024.10859594.

[32] H. Zhang, M. Liang, and Y. Wang, "YOLO-BS: A traffic sign detection algorithm based on YOLOv8," *Sci. Rep.*, vol. 15, pp. 1–11, Mar. 2025, doi: 10.1038/s41598-025-88184-0.

[33] R. Ma *et al.*, "InSAR-YOLOv8 for wide-area landslide detection in InSAR measurements," *Sci. Rep.*, vol. 15, pp. 1–22, Jan. 2025, doi: 10.1038/s41598-024-84626-3.

[34] J. Li *et al.*, "A small-scale target enhancement framework for aerial pineapple images on accurate agricultural information," *Comput. Electron. Agric.*, vol. 239, pp. 1–14, Dec. 2025, doi: 10.1016/j.compag.2025.110874.

[35] Y. Chen, J. Li, Y. Niu, and J. He, "Small object detection networks based on classification-oriented super-resolution GAN for UAV aerial imagery," in *2019 Chin. Control Decis. Conf. (CCDC)*, 2019, pp. 4610–4615, doi: 10.1109/CCDC.2019.8832735.

[36] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," 2016, *arXiv:1609.04802*.

[37] L. Riz *et al.*, "The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios," 2023, *arXiv:2304.05417*.

[38] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," 2019, *arXiv:1809.00219*.

[39] S. Woo, J. Park, J. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

[40] H.G. Correa *et al.*, "Thermal image super-resolution using real-ESRGAN for human detection," in *Proc. 20th Int. Jt. Conf. Comput. Vis. Imaging Comput. Graph. Theory Appl.*, 2025, pp. 247–254, doi: 10.5220/0013078800003912.

Jurnal Nasional Teknik Elektro dan Teknologi Informasi
Volume 15 Number 2 May 2026

This page is intentionally left blank